Unicode

Die Vereinheitlichung der weltweiten Schriftsysteme

Heinz Gnehm, 16. Februar 2005

Geschichte [i]

- Baudot-Code und andere Telegraphencodes wie ITA2 von der CCITT (1870-1930)
- ANSI X3.4-1963 (ASCII) basierend auf dem älteren FIELDATA
- EBCDIC bei IBM, basierend auf älteren Codes für Lochkarten (12 Löcher in 80 Zeilen)
- ISO 646 mit der Möglichkeit für nationale Varianten von ASCII ab 1972

Geschichte [ii]

- Nach 1985 entstanden die 16 ISO 8859-Standards, die alle europäischen Sprachen sowie Hebräisch und Arabisch umfassen
- ISO 8859-1 ist das bekannte Latin-1, also eigentlich 8-Bit ASCII
- Daneben existieren noch zahlreiche länder- und sprachspezifische Codierungen, hauptsächlich im asiatischen Raum

Geschichte [iii]

- 1984 begannen ISO/IEC JTC1/SC2/WG2 und das amerikanische Unicode-Konsortium mit der Arbeit an einem umfassenden System für alle Schriftsysteme der Welt
- 1993 wurde Version 1.1 des Unicode-Standards (ISO 10646-1:1993) fertiggestellt
- 2000 wurde Unicode 3.0 vorgestellt und seit 2003 ist Unicode 4.0 der offiziell gültige Standard

Schriftsysteme

- Es gibt weltweit vier grosse Schriftsysteme
  - Alphabetschriften in Europa
  - Konsonantenschriften im nahen und mittleren Osten
  - Silbenschriften auf dem indischen Subkontinent
  - Zeichenschriften in Ostasien (China, Taiwan, Korea und Japan)
- Die Komplexität der Schriftsysteme nimmt interessanterweise von West nach Ost zu (USA -> Japan)

Unicode - Allgemeines

- Unicode umfasst heute über 90’000 Zeichen
- Es werden auch ausgestorbene Schriftsysteme unterstützt, etwa Runen, Ogham und Keilschriften (die ägyptischen Hieroglyphen sind auf der Warteliste)
- Die korrekte Zeichendarstellung muss von den Programmen gewährleistet werden

Unicode - Grundsätze

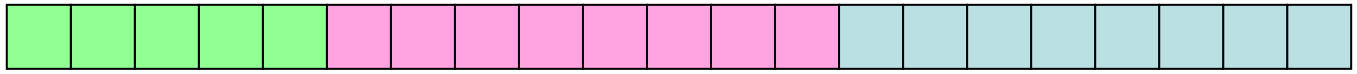
- Unicode codiert nur einzelne Zeichen, nicht aber die visuelle Darstellung dieser Zeichen
- Zeichen werden immer in der logischen (Lese-)Richtung codiert
- Gleiche Zeichen sollen auch gleiche Codes erhalten, ungeachtet ihrer sprachlichen Verwendung
- Bestehende Codierungen sollen so weit wie möglich berücksichtigt werden

Terminologe

1. **Zeichenumfang**
   Abstract character repertoire
2. **Zeichensatz**
   Coded character set, character encoding space
3. **Zeichencode**
   Character encoding form
4. **Zeichencodierung**
   Character encoding scheme
5. **Zeichenübertragungscodierung**
   Transfer encoding syntax

Unicode - Codierung [i]

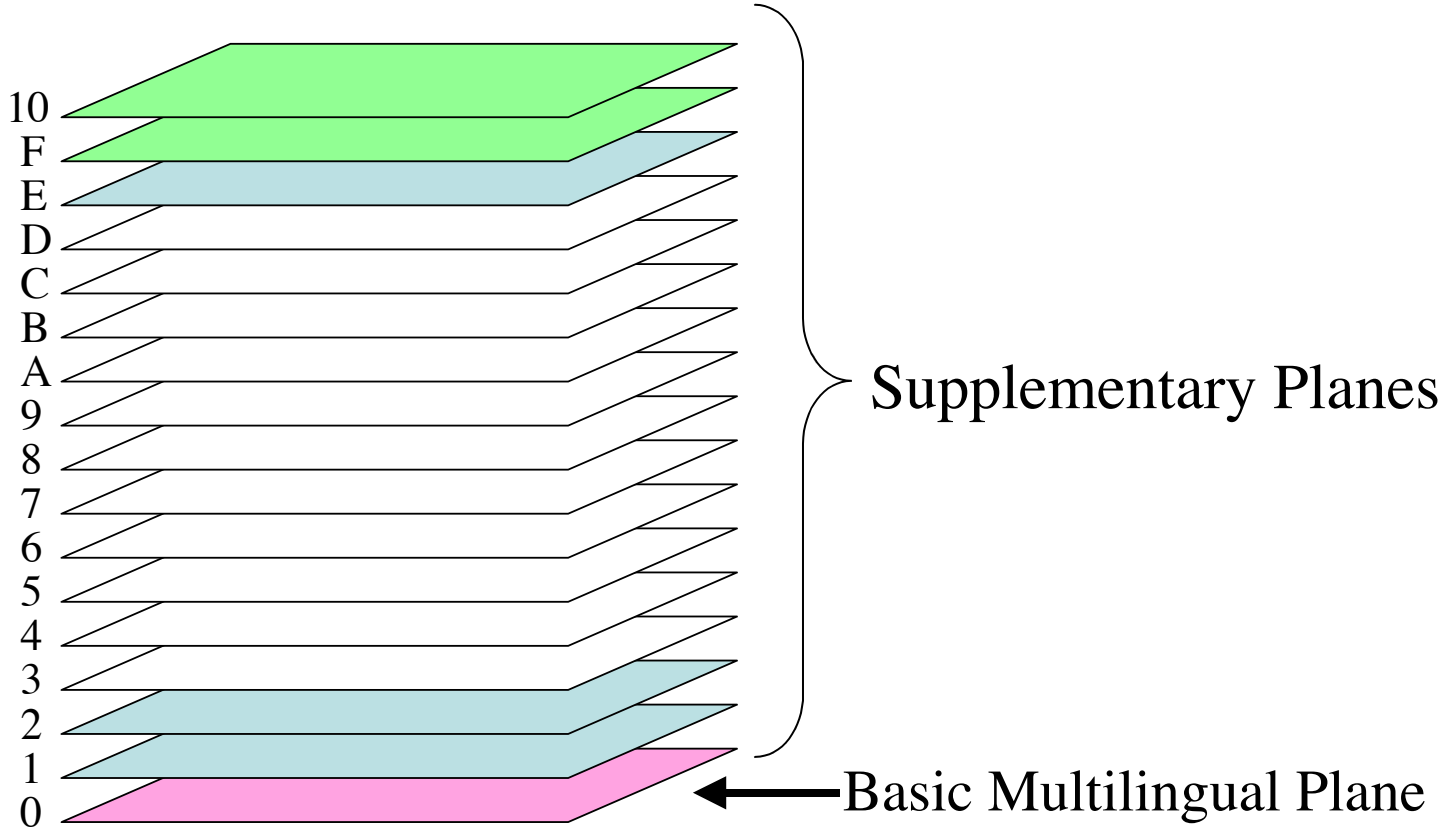
1. **Zeichenumfang**
   Alle Schriftsysteme der Welt
2. **Zeichensatz**
   21 Bits (17 Ebenen x 256 Zeilen x 256 Zellen)



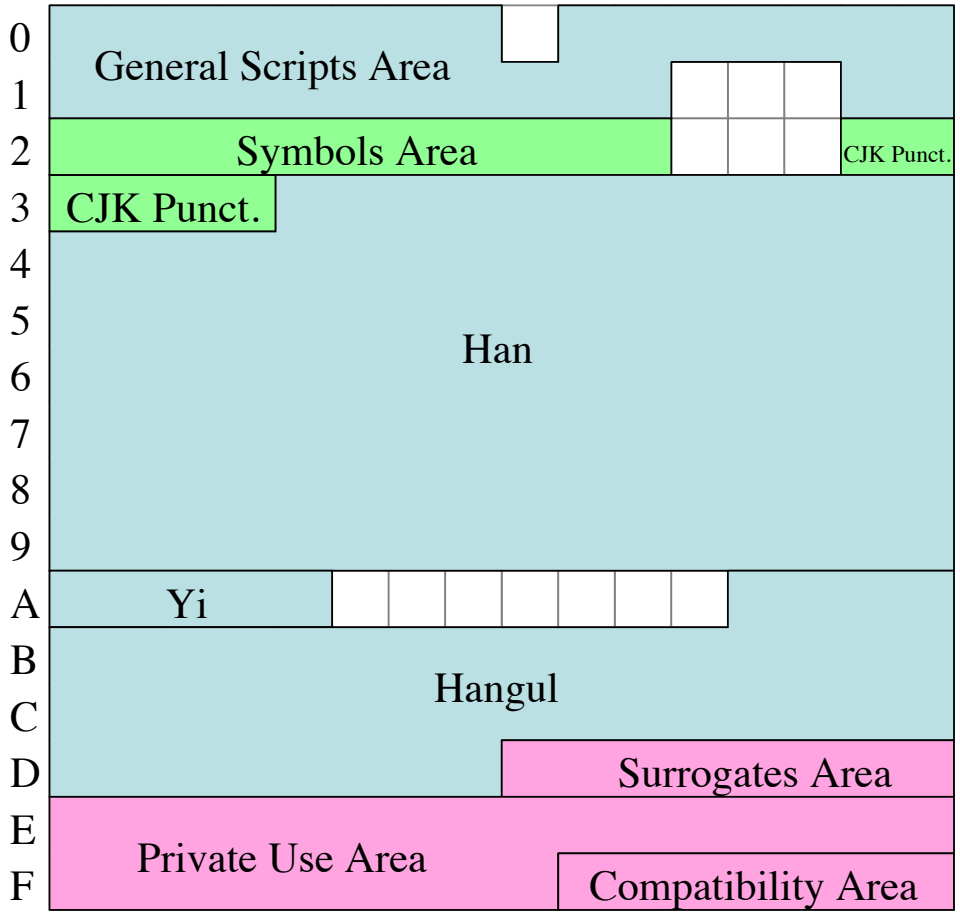
3. **Zeichencodierung**
   «Unicode Transformation Format» (UTF-32, UTF-16, UTF-8)
4. **Zeichensatzcodierung**
   «big-endian» und «little-endian»

Unicode - Codierung [ii]

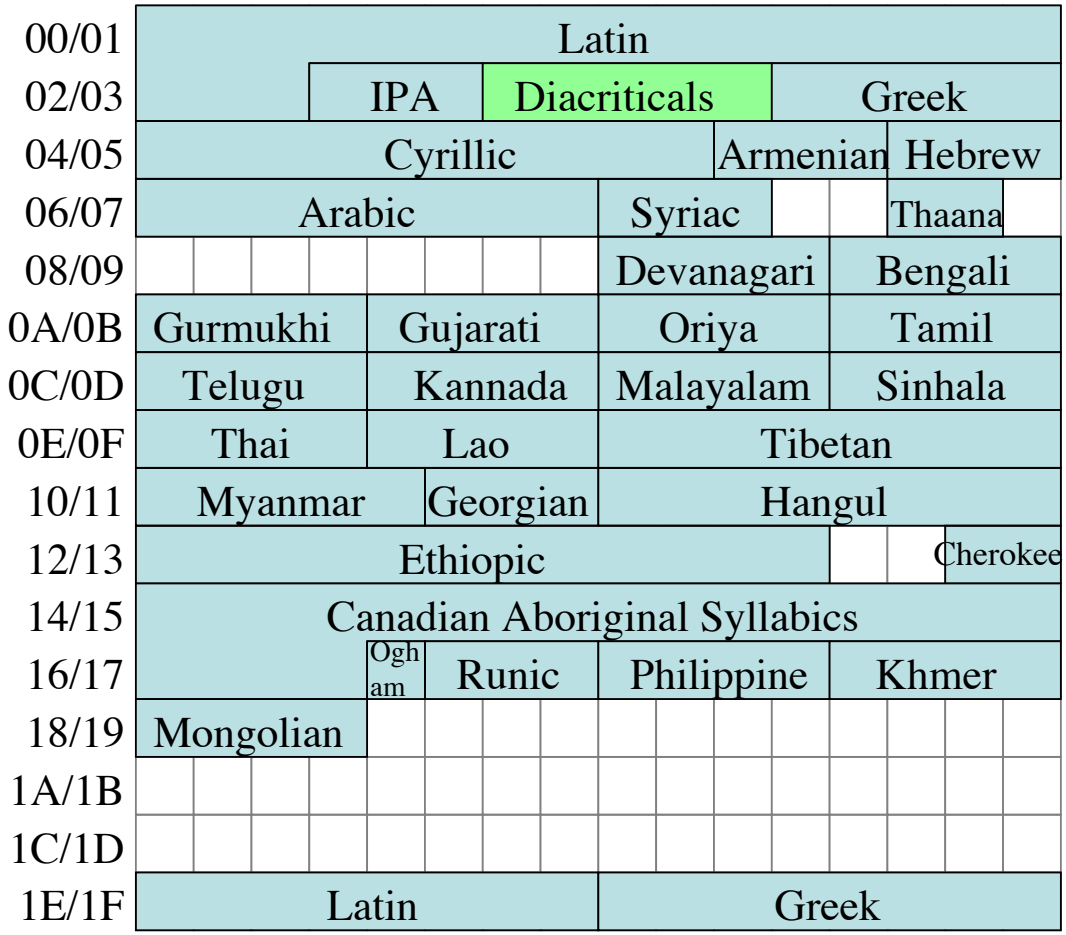
Unicode - Planes (17 Ebenen)



Unicode - Basic Multilingual Plane



Unicode - General Scripts Area



Herausforderungen

- In Japan wird Unicode stark kritisiert und Alternativen verwendet
- Die Vereinheitlichung der chinesischen Schriftzeichen stösst auf Widerstand
- Die technische Umsetzung von Unicode ist kompliziert (liegt aber nicht an Unicode)
- Der Speicherbedarf wird nahezu verdoppelt