



Praktikum zur Veranstaltung XML-Technologie: **Vorübungen**

Wiederholung einiger
wichtiger Unix-Kommandos,
Unicode u.a. Zeichensätze,
UTF-8 Codierung



Unix-Vorübungen



- Dateisystem
 - inodes, Verzeichniseinträge, Links (hard & soft)
- Kernel, Shell
 - Speziell: Die bash
 - Eingestellt als default-Shell?!
 - Befehlszeilenpuffer, -vervollständigung
- Prozesse
 - pid, Scheduler, Priorisierung; Kommandos dazu
 - Vorder- und Hintergrundprozesse
- I/O
 - stdin, stdout, stderr; /dev, „mounten“



- Das Prinzip „Unix-Werkzeugkasten“
 - Pipes bilden:
stdout von Prozess 1 wird stdin von Prozess 2
 - Output umlenken
 - Beispiele:
 - 1) `cat file1 | grep pattern-a | wc > resultfile`
 - 2) `find ~ -name *.zip 2&> /dev/null`
- Patterns, reguläre Ausdrücke
 - `rm *` # ☹
 - `ls a*.???` `b[1236-9]` `cd`



- Grundlage zum Wiederholen:
 - Übungen aus dem Praktikum zur Einführung in die Informatik 1.
- Wichtige Kommandos:
 - `man, info`
 - `ls, cd, pwd`
 - `mkdir, rmdir`
 - `cp, mv, rm, ln`
 - `cat, more, head, tail`
 - `grep, find`
 - `ps, pstree, kill, nice, time, fg, bg`
 - `mount, umount`



- Weitere Kommandos
 - `chmod, chgrp; touch`
 - `gzip, gunzip, gzcat; zip, unzip; tar`
- Anwendungen, Editoren
 - `vi, emacs`
 - `acroread`
 - `mozilla`
 - Speziell für XML:
 - `nsgmls` (später mehr)



Unicode

... und andere Zeichensätze



Unicode



- Informationen:
 - <http://czyborra.com/>
zu Zeichensätzen allgemein
 - <http://www.unicode.org/>
Speziell zu Unicode
- Beispiel: Buchstabe „ü“
 - Codepage 437 (DOS): 0x81
 - ISO-8859-1: 0xFC
 - Unicode (composite): U+00FC
 - Unicode (combining): U+0075, U+0308
 - Unicode, UTF-8 (s.u.): U+00FC = 0xC3, 0xBC



Unicode-Zeichenarten



- Basiszeichen
 - Unser normales Verständnis eines Zeichens
- Ideographische Zeichen
 - z.B. fernöstliche wie Kanji-Zeichen
- *combining characters*
 - „Pünktchen“, Akzentzeichen u.a.
 - Sie ergeben zusammen mit ihrem jeweiligen Vorläuferzeichen in einem String das endgültige Symbol
 - Beispiel: à = a`
 - Diese Zeichenkombinationen ergänzen die bereits vorhandenen Spezialzeichen
 - Die Kombinationsmethode schafft mit relativ wenigen Unicode-Einträgen eine große Vielfalt an möglichen Symbolen.
- *extenders*
 - (Unicode-Spezialthema, hier nicht behandelt)



Unicode-Codierungen



- UCS-4:
 - Die allgemeine 4-Byte-Angabe: U+xxxxxxxx
- UTF-8, UTF-16, UTF-32
- Unterscheidung im Fall UTF-16:
 - *high-endian* vs. *low-endian* mittels Sonderzeichen xFEFF
- UTF-8 Codierung:

U+00000000	–	U+0000007F	0xxxxxxxx	
U+00000080	–	U+000007FF	110xxxxx	10xxxxxx
U+00000800	–	U+0000FFFF	1110xxxx	10xxxxxx 10xxxxxx
U+00010000	–	U+001FFFFF	11110xxx	(10xxxxxx) ₃
U+00200000	–	U+03FFFFFF	111110xx	(10xxxxxx) ₄
U+04000000	–	U+7FFFFFFF	1111110x	(10xxxxxx) ₅

 - 1 bis 6 Oktets pro Unicode-Zeichen (31 bits), niemals xFE oder xFF.
 - Stets klar, ob Folgebyte vorliegt und wieviele Folgebytes notwendig!



Suchaufgaben



- Aufgabe:
 - Ermitteln Sie die Codes der umseitig folgenden Zeichen
- Hinweise:
 - Dokumentieren Sie ihre Ergebnisse tabellarisch in Datei **xmltech-v1.txt** – diese werden noch benötigt!
 - Verwenden Sie die angegebenen Internetquellen!
 - Geben Sie stets den Unicode an!
 - Geben Sie den Code aus einer der ISO-8859-Tabellen an, incl. der Tabellenummer selbst, sofern ein ISO-Code für das Zeichen existiert.



Suchaufgaben



- A) Westliche Sonderzeichen
 - Ä, ä, Ö, ö, Ü, ü; ß
- B) Währungszeichen
 - Britisch Pounds, Euro: £, €
- C) Mathematische Sonderzeichen
 - Quantoren: „Für alle“, „es existiert“, „es existiert nicht“: $\forall, \exists, \nexists$
 - „daraus folgt“, „ist äquivalent“, „ist gleich“, „ist ungleich“: $\Rightarrow, \Leftrightarrow, =, \neq$
 - Sonstiges: „ist Element von“; „alpha“, „beta“, „gamma“, das Gradzeichen (37°C): $\in, \alpha, \beta, \gamma, ^\circ$
- D*) Kanji
 - Schreiben Sie „Japan“ - auf japanisch (Ni-Hon): 日本



Codierungsaufgabe



- UFT-8 Codes berechnen:
Ermitteln Sie die UTF-8 Codes (Oktet-Sequenzen) der folgenden drei Unicode-Zeichen:
 - A) ß
 - B) €
 - C*) 日 („Ni“, Kanji-Zeichen für "Sonne")
- Abgabe
 - Tragen Sie Ihre Ergebnisse ein in Datei **xmltech-v2.txt**



Abgaben



- Geben Sie folgende 2 Dateien ab:
 - **xmltech-v1.txt**
 - **xmltech-v2.txt**
- Abgaberegeln:
 - Gemäß der allgemeinen Abgaberegeln (beschrieben in eigener PDF-Datei).
- **Hinweise:**
 - Immer Name, Vorname, MatNr, etc. in der Datei angeben
 - Keine Teamlösungen - "Kopien" erhalten keine Punkte!
 - Verwenden Sie "cp", keinen Filemanager / GUI-Tools!