



Mediale Informatik

Vorlesung im 3. Semester Medieninformatik
Wintersemester 2002/2003

Prof. Dr. Michael Weber



Kapitel B2

Text

Schrift und Alphabet



- Schrift:
System zur Lesbarmachung der gesprochenen Sprache
 - d.h. Transfer von kontinuierlicher Sprache in diskrete 2-dimensionale räumliche Darstellung
- Alphabet:
Festlegung der Kodierungselemente, der Symbole



Schrift und Alphabet



■ Historie

■ Entwicklung der Schrift

- 4000 v. Chr. sumerische Bilderschrift
- 3000 v. Chr. Hieroglyphen, Keilschrift
- 2000 v. Chr. Buchstabenschrift (semitisches Uralphabet)

■ Kodierungsvorgang

- Handschrift (mit Feder, Pinsel, Meißel)
- Druckverfahren (J. Gutenberg, 1455)
- Schreibmaschine (P. Mitterhofer, 1864)
- Computertastatur als Eingabegerät zur digitalen Speicherung der Texte
- Spracherkennungs-Software



Schrift und Alphabet



- Phonetisches Alphabet
 - z.B. die Buchstabenalphabete der westlichen Welt
- Silbenalphabet
 - z.B. Koreanisches Hanguk
- Ideographisches Alphabet
 - Symbole repräsentieren ganze Worte oder Konzepte
 - z.B. Kinji
- Ziffern, Punktierungszeichen und ähnliches gehören ebenfalls zum Alphabet



Schrift und Alphabet



- Digitale Repräsentation
 - Abbildung der (abstrakten) Symbole des Alphabets auf digitale Werte
 - Das Bild heißt Zeichensatz – Character Set
 - Der Wertebereich (das Alphabet) heißt Symbolrepertoire – Character Repertoire
 - Für jedes genutzte Symbol im Repertoire definiert der Zeichensatz einen Kodewert bzw. Kodepunkt



Kodierungsstandards



- ASCII
 - American Standard Code for Information Interchange
 - Seit den 70ern der dominante Zeichensatz
 - Ein anderer früher genauso vorkommender: EBCDIC (Extended Binary Coded Decimal Interchange Code) von IBM
 - 7 Bit Kode → 128 Codepunkte
 - 0 .. 31 Steuerzeichen,
32 .. 127 druckbare Zeichen
 - ASCII ist vorwiegend geeignet für US-English



Kodierungsstandards



- ASCII
 - 1972 Standardisierung als ISO 646
 - Unterschiedliche länderspezifische Varianten mit unterschiedlichen spezielleren Symbolen
 - z.B. ISO 646-US, ISO 646-UK
 - Das 8. Bit diente ursprünglich als Paritätsbit zur Fehlerkorrektur
 - Mit verbesserten Übertragungsmöglichkeiten wurde dies unnötig
 - → dieses Bit kann genutzt werden, um den Kode zu erweitern
 - Herstellerspezifische Belegungen der Codepunkte 128 .. 255 entstanden



Kodierungsstandards



- ISO 8859-X

- Standardisierung in den 80ern als sog. Multipart-Standard
- z.Zt. 10 standardisierte 8-Bit-Erweiterungen von 7-Bit ASCII um länderspezifische Zeichen
 - ISO 8859-1 (ISO Latin1) westliche Alphabete
 - ISO 8859-2 (ISO Latin2) osteuropäische (tschech., slowakisch, kroatisch)
 - ISO 8859-5 kyrillisch
 - ISO 8859-6 arabisch
 - ISO 8859-7 griechisch
 - ISO 8859-8 hebräisch
 - ISO 8859-0 (ISO Latin0) Latin 1 + Eurosymbol



Kodierungsstandards



- ISO 8859-X
 - Nachteile
 - Es gibt immer noch herstellerspezifische Abweichungen
 - Multilinguale Anwendungen sind nicht möglich
 - Ideographische Alphabete werden nicht unterstützt
- Fazit:
ein 8 Bit Code hat einfach zu wenige Codepunkte



Kodierungsstandards



- ISO 10646
 - 32-Bit Code (1991)
 - Gruppierung von Zeichensätzen als Hyperkubus
 - g = Gruppe
 - p = Ebene (Plane)
 - r = Zeile (Row)
 - c = Spalte (Column)
 - Mit Wertebereich jeweils 0..255
 - Symbole werden als Tupel (g,p,r,c) angegeben
 - $(0,0,0,*)$ ist ISO Latin1
 - $(0,0,*,*)$ ist die Basic Multilingual Plane
 - identisch mit Unicode



Kodierungsstandards



- ISO 10646
 - Für Chinesisch, Japanisch und Koreanisch werden die Zeichensätze unifiziert (CJK consolidation)
 - 39000 Codepunkte sind definiert
 - 6400 Codepunkte sind für private Nutzung



Kodierungsstandards



- Unicode
 - 16-Bit Code
 - Verwendet u.a. in HTML, XML, Java
 - Beinhaltet die gängigsten Alphabete für Sprache
 - Lateinisch, Griechisch, Kyrillisch, Armenisch, Hebräisch, Arabisch, Devanagarisch, Bengalisch, Gurmukhisch, Gujaratisch, Oriya, Tamilisch, Telugisch, Kannada, Malayalam, Thai, Lao, Georgisch, Tibetanisch
 - Chinesische, japanische und koreanische Ideogramme
 - japanische und koreanische phonetische und Silbenskripte
 - Zeichensetzungen, Akzente und Tilden, mathematische Symbole, Dingbats-Symbole



Kodierungsstandards



- Zeichenkodierung – Character Encoding
 - Abbildung der Codepunkte auf eine Bytesequenz zur Speicherung bzw. Übertragung
 - Quoted-Printable (QP)
 - 8-Bit Code in 7-Bit darstellen
 - Zeichen zwischen 128 und 255 werden als 3 Byte dargestellt
 - z.B. é ist kodiert als 233 ($E9_{16}$) in ISO Latin 1, als quoted-printable ist é kodiert als ´ = E9 ´
 - = ist kodiert als ´ = 3D ´



Kodierungsstandards



- Zeichenkodierung – Character Encoding
 - ISO 10646 Kodierung
 - UCS-4 1:1 Abbildung auf 4 Byte
 - UCS-2 nur die niedrigen 2 Byte werden kodiert (entspricht Unicode)
 - Unicode Transformation Formats, UTF
 - UTF-7 alles kodiert in ASCII, ähnliches Prinzip wie Q-P
 - UTF-8 7-Bit ASCII werden als 1 Byte kodiert, alles andere in bis zu 6 Bytes mit 8.Bit = 1
 - UTF-16 Paare von 16-Bit Werten können als einzelner 32-Bit Wert zusammengefasst werden;
dies erschließt 15 weitere Ebenen aus ISO 10646 (ca. 1 Mio. Symbole)



Schriftart



- Glyphen dienen als visuelle Repräsentation von Symbolen
- Schriftart - Font
 - Zusammenstellung von zueinander passenden Glyphen zu einem Alphabet
 - früher Bleiletern
 - heute Beschreibung der grafischen Darstellung
- Schriftart-Familie
 - Passende Glyphen mit unterschiedlichen Größen und Schnitt



Schriftart



- Schriftgröße
 - Ursprung im Bleisatz
 - Die Größe ist festgelegt durch die Kegelhöhe des Bleiletters
 - d.h. etwas größer als das Druckbild
 - beinhaltet das Fleisch um das Druckbild
 - Einheit Didot-Punkt (0,376 mm)
 - [Didot, 1784]
 - Computerschriften verwenden Pica-Punkte (0,351mm)
 - d.h. 12 pt Pica entspricht ca. 11 pt Didot

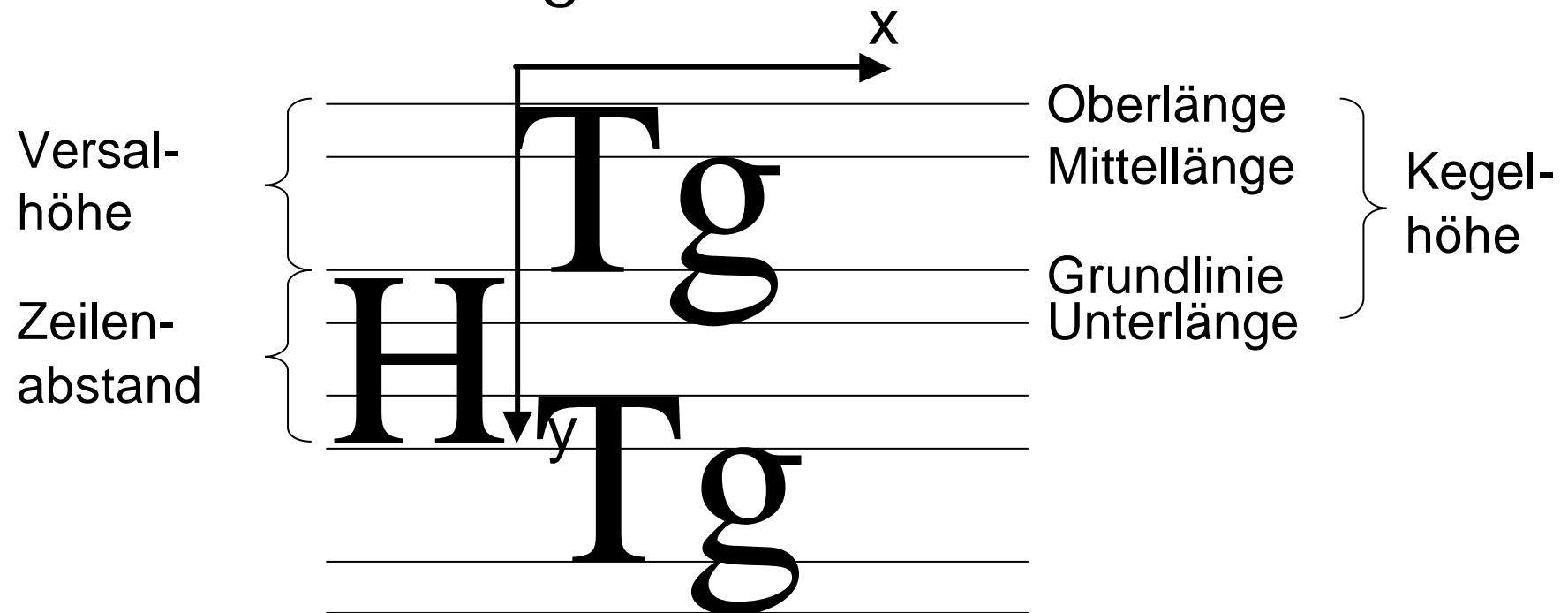


Schriftart



■ Schriftmaße

- Die Referenzkoordinate liegt oben links
- Der Zeilenabstand wird entsprechend berücksichtigt



Schriftart



- Alternative Schriftmaße
 - ex Höhe des kleinen x
 - dies entspricht der Länge zwischen Grund- und Mittellänge
 - em Breite des großen M und damit des Kegels
 - Dies ist oft identisch mit der Fontgröße in pt
 - en Breite des großen N = $\frac{1}{2}$ em



Schriftart



- Wegen unterschiedlichen Fleisches sind Schriftarten gleicher Größe unterschiedlich groß im Druckbild

Dies ist 48 Punkt



Schriftart



- Schriftattribute verändern die Präsentation der Zeichen, z.B.
 - **Fett**
 - *Kursiv*
 - Schattiert
 - Relief



Schriftart



- es gibt tausende Schriftarten, z.B.
 - Times New Roman
 - Arial
 - Σψμβολ Symbol
 - Zapf Chancery
 - ϕ)(■ υο ρ)(■ υο ◆ Dingbats
 - ...
- Die Typografie beschäftigt sich mit den Frage- und Problemstellungen zu Schriften



Schriftart



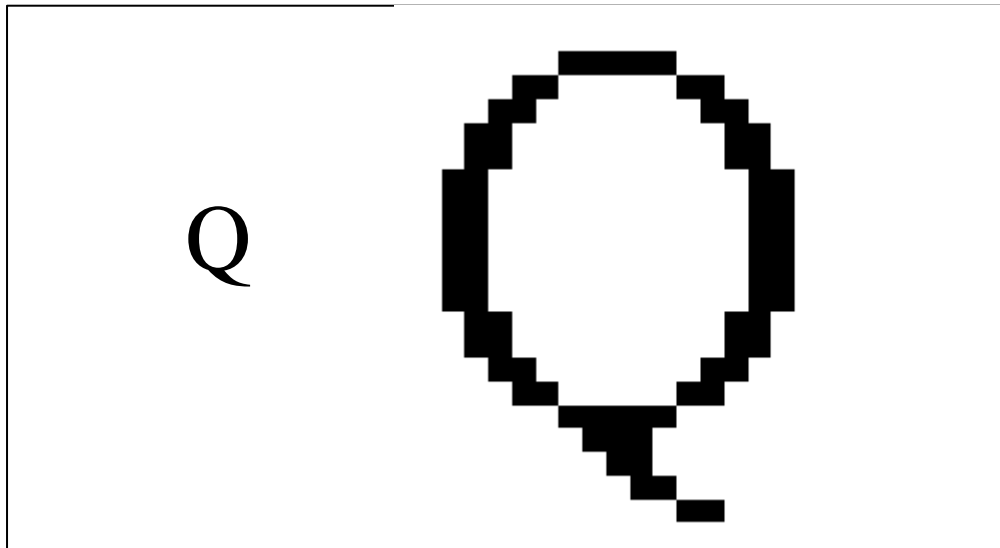
- Grafische Repräsentation der Glyphen
 - Bitmap-Font
 - Repräsentation als Bitmap
 - Outline-Font
 - Repräsentation als Vektorgrafik



Schriftart



- Bitmap Font
 - Glyphen werden in Rasterform gespeichert
 - bei Bedarf werden sie in den Speicher geladen
 - Beispiel mit 8-facher Vergrößerung (Font = Times New Roman)



Schriftart



- Eigenschaften von Bitmap Fonts
 - Auflösungsabhängig
 - schlecht skalierbar
 - Bitmap Fonts brauchen bei zunehmender Größe viel Speicher
 - Zeichen mit Schriftattributen müssen zusätzlich gespeichert werden
 - Betriebssysteme haben meist ihre speziellen zugeschnittenen Bitmap-Fonts



Schriftart



- Outline-Font
 - Plattformübergreifende Formate
 - Eine standardisierte Fontbeschreibung garantiert gleiches Aussehen auf unterschiedlichen Plattformen
- Beispiele
 - Adobe Type 1 (Postscript Fonts)
 - Vorwiegend auf Macintosh
 - Apple TrueType
 - Vorwiegend auf Windows, aber auch Mac



Schriftart



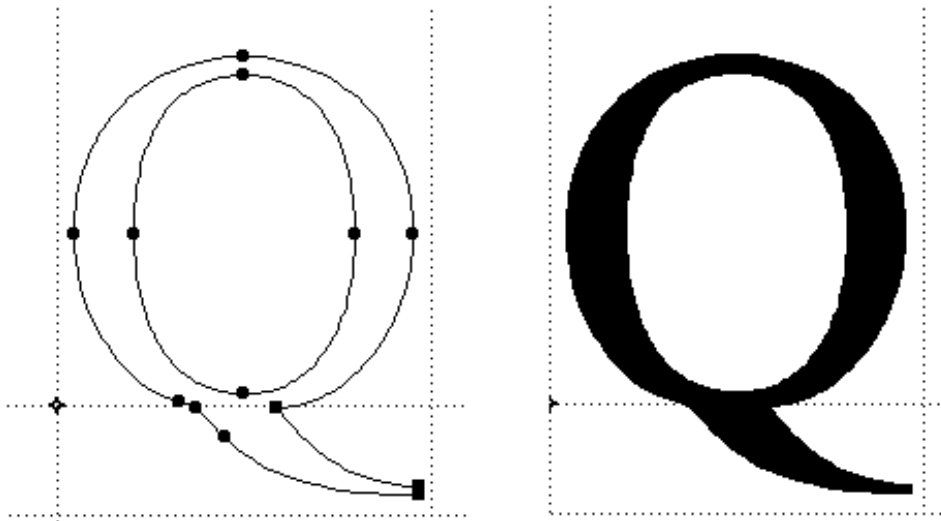
- Adobe Type 1
 - Glyphen werden in einer Untermenge von Postscript beschrieben
 - Grundlage der Beschreibung sind kubische Bézierkurven
 - Der Adobe Type Manager (ATM) führt quasi als Treiber das Rendering durch
- TrueType
 - Grundlage der Beschreibung sind quadratische Kurven
 - Das Rendering ist ins Betriebssystem eingebaut



Schriftart



- TrueType
 - Umrisse der Zeichen werden als quadratischer Kurvenzug angegeben
 - zur Darstellung wird der Kurvenzug ausgefüllt
 - Procedural Instructions dienen der Verbesserung bei niedriger Auflösung



Schriftart



- Interpolation & Approximation mit Splines

- stückweise linear:

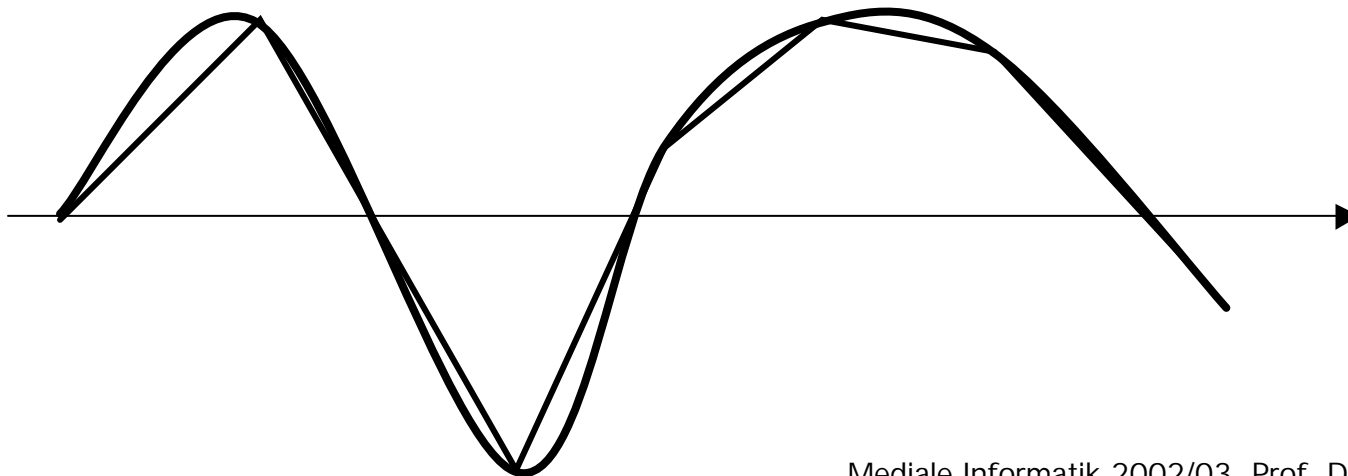
$$f_i(x) = a_i x + b_i$$

- an den Stützpunkten stetig:

$$f_i(x) = f_{i+1}(x)$$

- stückweise kubisch:

$$f_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i$$



Schriftart



- Interpolation & Approximation mit Splines
 - Berechnung:
 - stetig:
$$f_i(x_k) = s_k,$$
$$f_i(x_{k+1}) = s_{k+1}$$

→ 2n Gleichungen
 - 'glatt', d.h. Ableitungen gleich:
$$f'_i(x) = f'_{i+1}(x)$$

→ 2(n-1) Gleichungen
 - Gleichungssystem mit 4n Unbekannten:

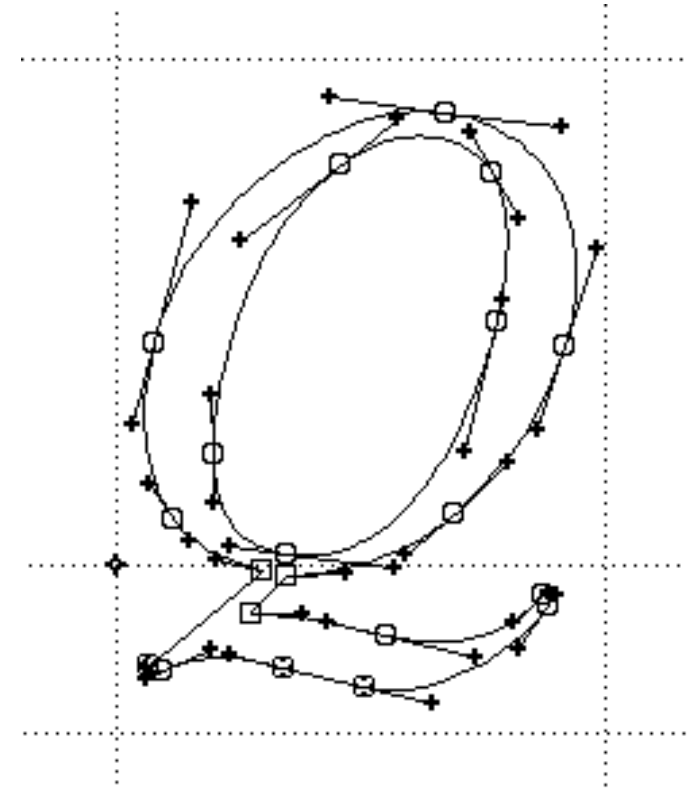
→ 2n + 2n - 2 Gleichungen
 - je nach Randbedingungen verschiedene Approximationseigenschaften



Schriftart



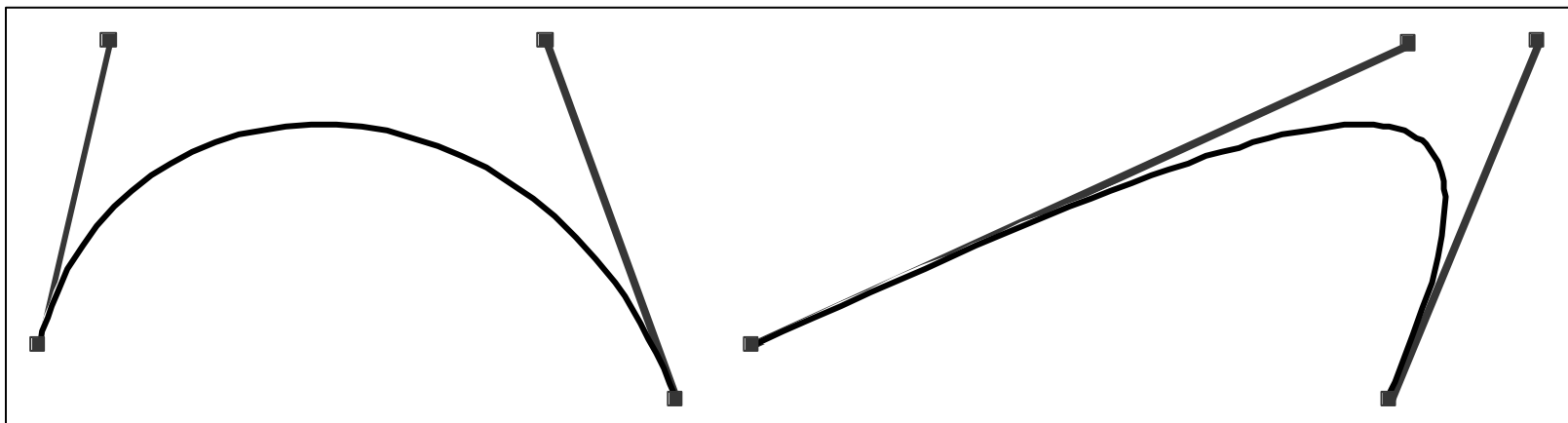
- Adobe Type 1
 - Fontparameter
 - Zeichenparameter
 - Bézier-Kurven zur Beschreibung des Umrisses
 - Deklarative Hints zur Verbesserung bei niedriger Auflösung



Schriftart



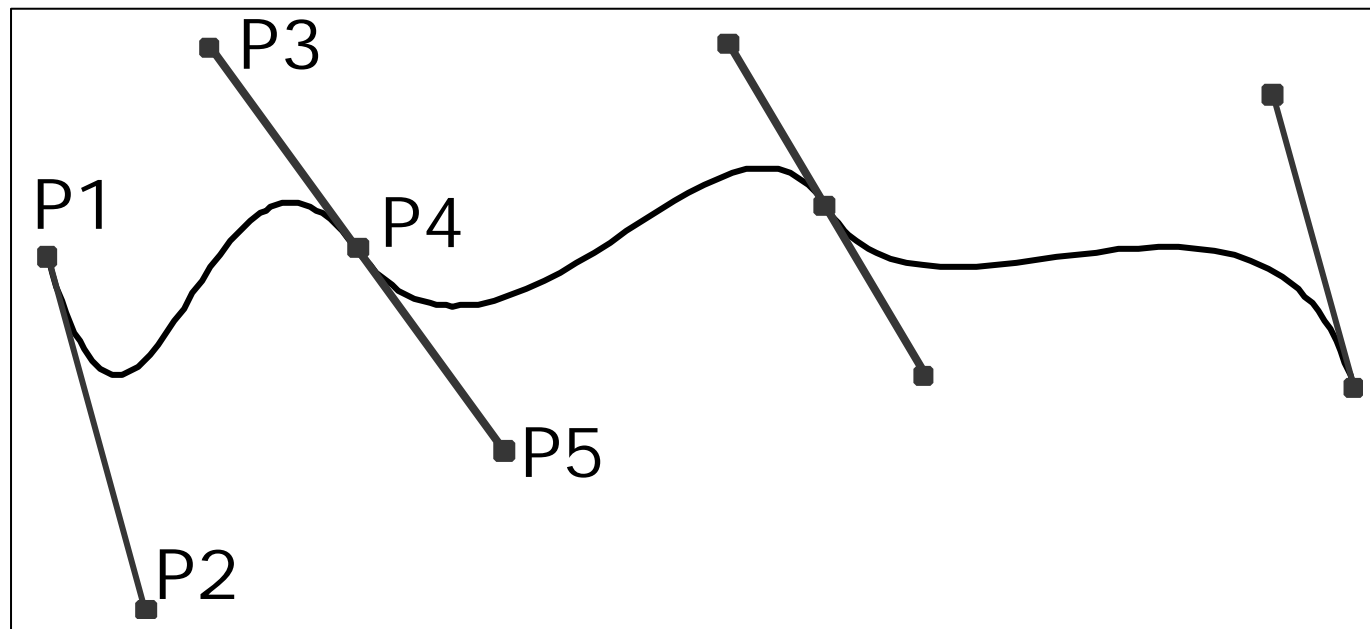
- Bézier-Kurven
 - Beschreibung der Kurve durch vier Punkte
 - Anfangspunkt und Endpunkt
 - 2 Kontrollpunkte legen Tangentenrichtung fest und „ziehen“ je nach Vektorlänge die Kurve an



Schriftart



- zusammengesetzte Bézier-Kurven
 - Kontrollpunkte so legen, dass die Kurve glatt wird
 - 3 fortlaufende Kontrollpunkte P_3 , P_4 , P_5 auf einer Geraden (1. Ableitung gleich)
 - gleicher Abstand zwischen P_3 , P_4 sowie P_4 , P_5



Schriftart



- Berechnung von Beziér-Kurven
 - Für 4 Punkte $P_1 \dots P_4$:
$$Q(t) = (1-t)^3 P_1 + 3t (1-t)^2 P_2 + 3t^2 (1-t) P_3 + t^3 P_4$$
 - Durch frei gewählte Unterteilung von $[0,1]$ für t erhält man beliebig viele Kurvenpunkte, die durch Strecken verbunden werden



Schriftart



- OpenType
 - Dateiformat, um Type 1 und TrueType Fonts zu speichern
 - Entwicklung von Adobe und Microsoft, 1997/98



Schriftart



- Multiple Master Fonts
 - Basiert auf Adobe Type 1
 - Eine Schriftart-Familie wird durch Parameter beschrieben
 - Designachsen
 - Gewicht ultra-light – ultra-bold
 - Breite condensed – extended
 - Optische Größe Proportionen der Glyphen, Kerning, Ligaturen, ...
 - Serifenstil sans serif – serif
 - Font-Instanzen können aus einem Parametersatz generiert werden



Schriftart



- Multiple Master Fonts
 - Nachteil
 - Multiple Master Font benötigt mehr Speicher als einzelner Font
 - Vorteile
 - Lokal nicht vorhandener Font kann leichter substituiert werden
 - Copyrighted Fonts müssen nicht im Dokument mitgespeichert werden
 - Beispiel
 - Adobe Portable Document Format, PDF
 - Der Acrobat Reader hat je einen generischen Serif und Sansserif Master Font



Fortgeschrittene Textrepräsentationen



- Marked-up Text

- Text enthält Inhalt und Form

- z.B. troff `.ce`

dies ist zentrierter Text

- T_EX, HTML, SGML

- Strukturierter Text

- Text enthält Strukturinformationen

- z.B. über Kapitelstruktur

- XML, RTF



Fortgeschrittene Textrepräsentationen



- Hypertext
 - Text ist nicht-linear
 - er enthält Knoten (den Text selbst) und Verbindungen (Links) zwischen Knoten
- Mischformen sind häufig anzutreffen

