Introduction to Media Informatics: Text and Hypertext

Dr. Mathias Lux
Associate Professor
Alpen-Adria Universität Klagenfurt

Agenda

- Text & Markup
- Hypertext
- WWW
  - URL, DNS, HTTP
- HTML

Text Formats

- American Standard Code for Information Interchange (ASCII)
- Defines 128 character, 95 of them printable

USASCII code chart

Bits b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row ↓ Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

Src. Wikipedia, http://en.wikipedia.org/wiki/ASCII

ASCII

- Letter has a 7-bit code
  - A -> $1000001_2 = 65_{10}$
- Control Codes also have 7-bit codes

000 0111	007	7	07	BEL	BEL	^G	\a	Bell
000 1000	010	8	08	BS	BS	^H	\b	Backspace[d][e]
000 1001	011	9	09	HT	HT	^I	\t	Horizontal Tab[f]
000 1010	012	10	0A	LF	LF	^J	\n	Line feed
000 1011	013	11	0B	VT	VT	^K	\v	Vertical Tab
000 1100	014	12	0C	FF	FF	^L	\f	Form feed
000 1101	015	13	0D	CR	CR	^M	\r	Carriage return[g]

Src. Wikipedia, http://en.wikipedia.org/wiki/ASCII

Which Characters are Missing?

USASCII code chart

b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
Bits b_4 ↓	b_3	b_2	b_1	Row ↓ Column →	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

Src. Wikipedia, http://en.wikipedia.org/wiki/ASCII

Text Formats: ISO/IEC 8859-1

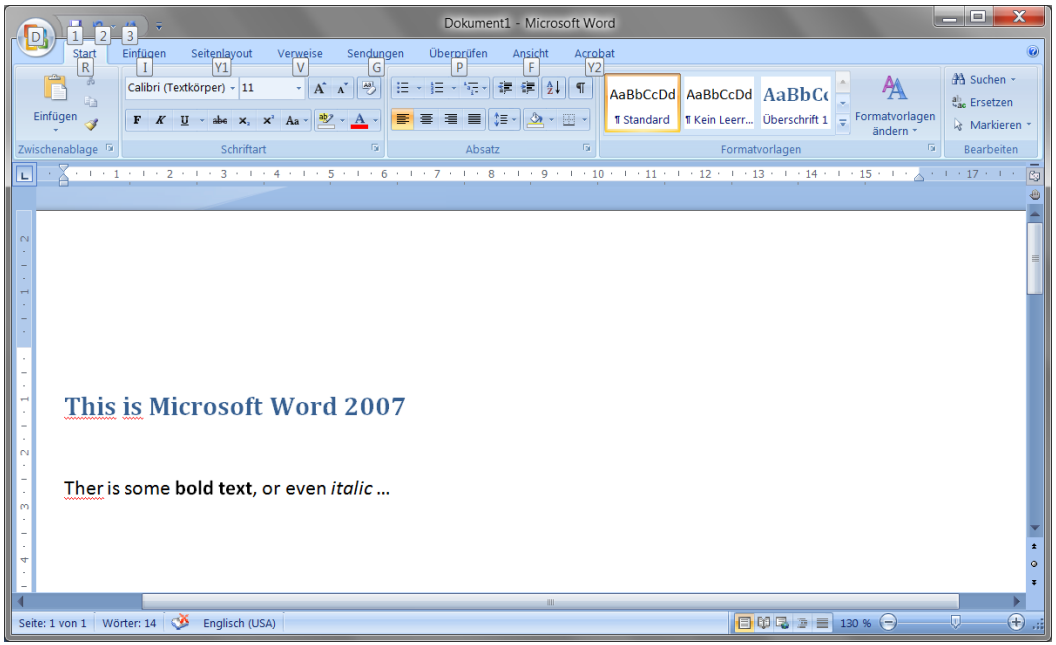
- Commonly referred to as "Latin-1"
- Each character has an 8-bit code
- 191 characters supported
- Sufficient for many Western European languages
  - But still some characters missing

Text Formats: Unicode Standard

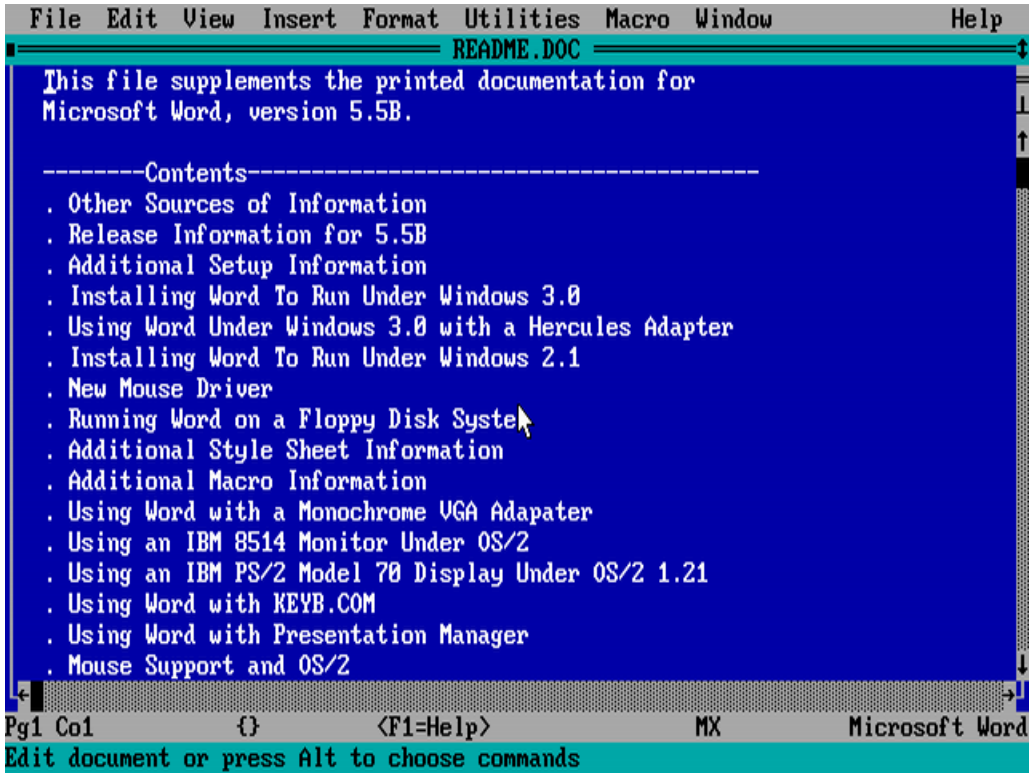
- More than 100,000 characters
- Supporting left-to-right and right-to-left scripts
- Defines character properties, rendering, ...
- Proposes different encodings like UTF-8, ..

WYSIWYG Text Editing

- Short for "What you see is what you get"
  - Means that you can actually see what it will look like when printing it …



WYSIWYG Text Editing



WYSIWYG: Related Concepts

- Type faces
  - were "bought with the printer"
  - expensive printers had nice postscript fonts
- Programs would send to a printer
  - a font set command and
  - the text to print
- Printers would even split documents in pages

WYSIWYG: True Type Fonts

- Developed 1980s by Apple and Microsoft
- Competitors of Adobe Type 1 (Postscript) fonts
  - Licensing, etc.
- Provide a more flexible way to
  - define outlines of fonts
  - control display at different font sizes
- First common font families
  - Times Roman, Courier, Helvetica

WYSIWYG: True Type Fonts

- Not the same look on every system
  - different outline -> pixel renderers
- Can now be loaded into printers as "soft fonts"
- The open source implementation of True Type is called "Free Type"
- OpenType is the successor created by Microsoft and Adobe

Some Font Facts ...

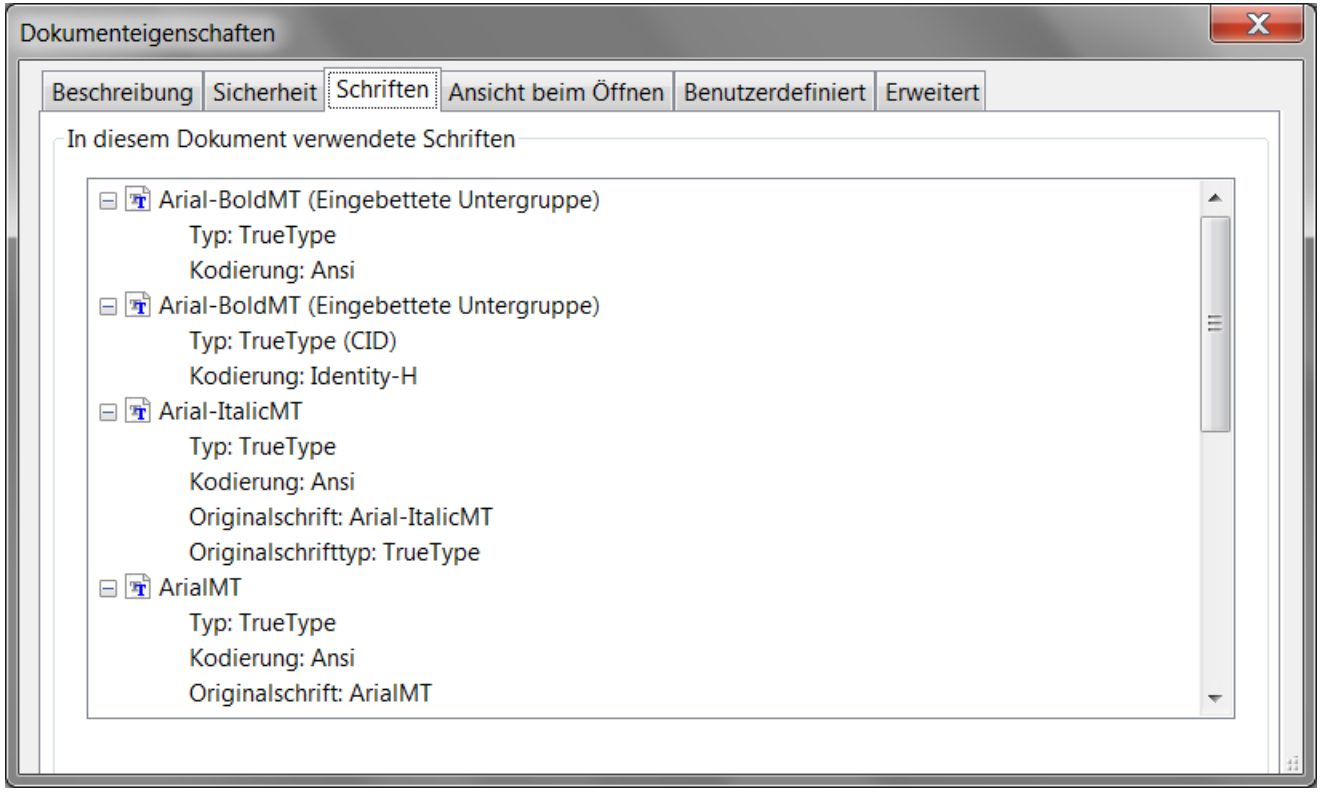
- A font can be
  - Serif, Sans Serif, Slab Serif
  - Monospaced

Times New Roman: The quick brown fox jumps over the lazy dog.

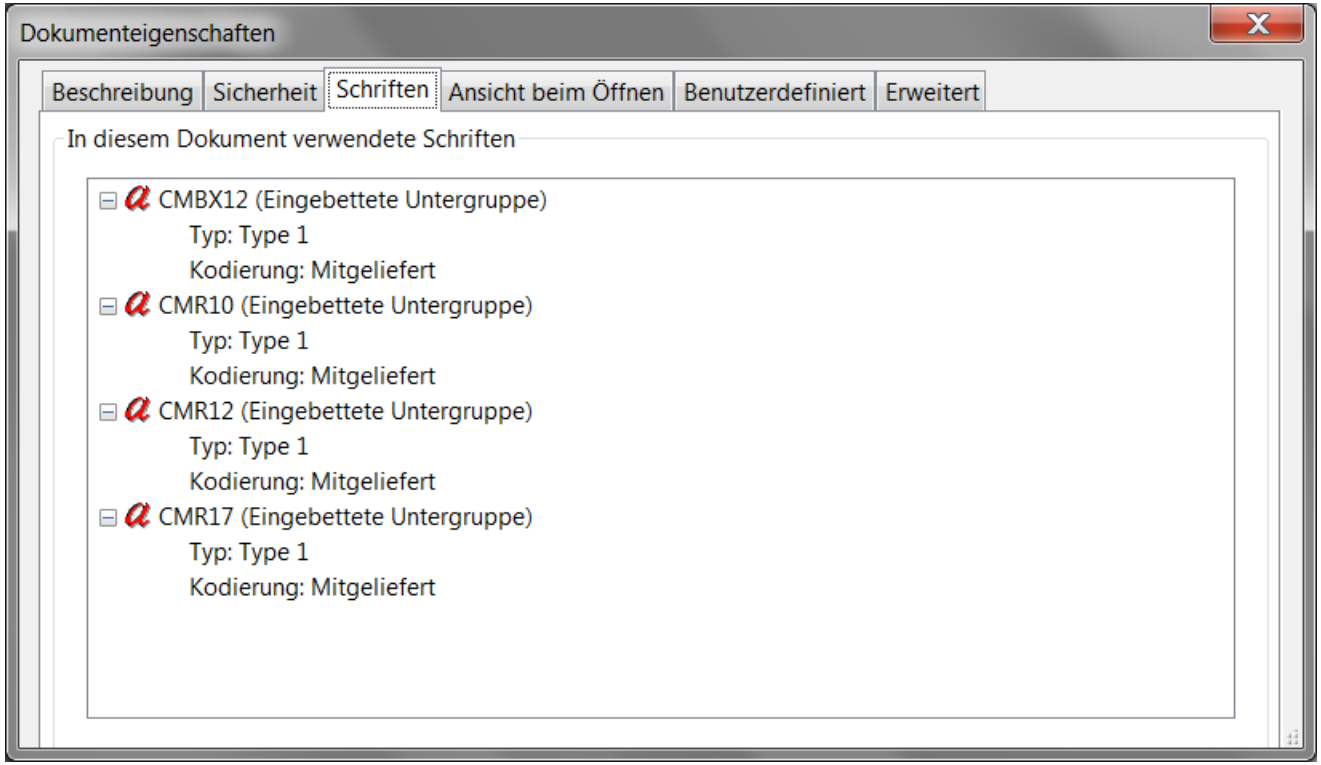
Droid Sans: The quick brown fox jumps over the lazy dog.

Courier New: The quick brown fox jumps over the lazy dog.

Font Embedding vs. Reference

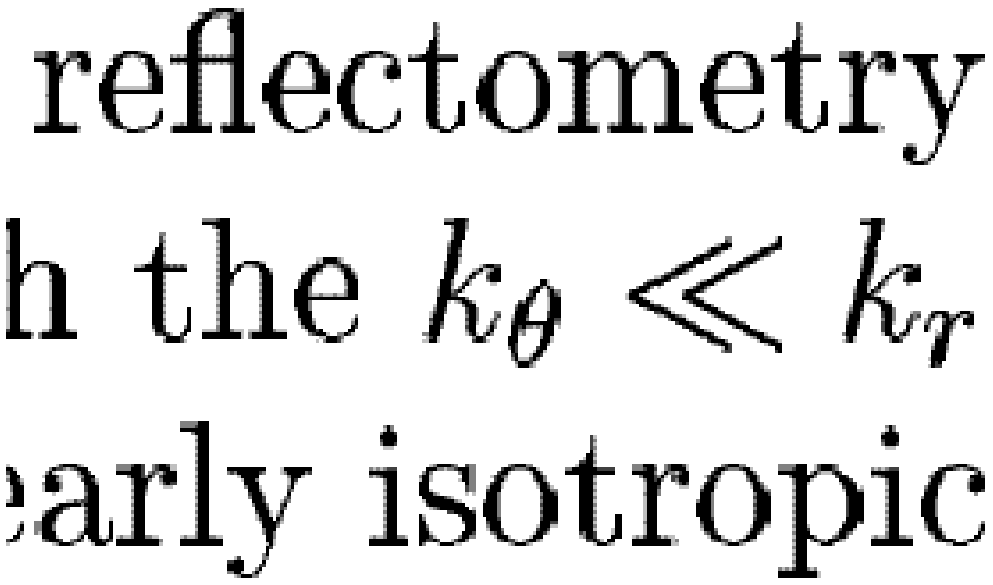


Font Embedding vs. Reference



Bitmap Fonts

- Pre-rendered fonts embedded in a document
- Lead to aliasing & artifacts when zoomed in



Typography

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec pellentesque odio in risus sodales aliquam. Aliquam orci tellus, elementum ultrices felis vitae, hendrerit gravida dui. Aenean sapien ligula, rhoncus et elementum quis, elementum eget augue. Aenean bibendum, elit sed interdum malesuada, diam turpis facilisis sem, vitae condimentum ipsum velit eget est. Sed in mauris tincidunt, tincidunt tellus sit amet, commodo ligula. Pellentesque porttitor, mauris sit amet malesuada fringilla, tortor justo dictum nibh, ac commodo lectus nisl at ante. Cras vitae neque fringilla, porttitor augue quis, euismod risus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec pellentesque odio in risus sodales aliquam. Aliquam orci tellus, elementum ultrices felis vitae, hendrerit gravida dui. Aenean sapien ligula, rhoncus et elementum quis, elementum eget augue. Aenean bibendum, elit sed interdum malesuada, diam turpis facilisis sem, vitae condimentum ipsum velit eget est. Sed in mauris tincidunt, tincidunt tellus sit amet, commodo ligula. Pellentesque porttitor, mauris sit amet malesuada fringilla, tortor justo dictum nibh, ac commodo lectus nisl at ante. Cras vitae neque fringilla, porttitor augue quis, euismod risus.

Reading Type:Rider

- Play, experience and read Type:Rider, the video game
  - Up to level 5 - Clarendon

Markup Languages

- Text (content) is annotated by markup,
  - cp. "marking up" text with a red pencil
- Markups are syntactically distinguishable from the actual text (content)
- Different Types of markup languages
  - presentational ... ie. WYSIWYG editors
  - procedural ... processing instructions
  - descriptive ... semantics of the text

SGML

- Direct ancestor of Scribe (1980, Brian Reid)
  - Scribe was based on a grammar
- Structural description of a document
  - instead of a presentation, ie. style is defined separately
- Long-term valid, machine-readable documents
- SGML ...
  - is a rooted acyclic directed graph.
  - supports document type declarations

TeX

- Typesetting system widely spread in academia
  - Released in 1978 by Donald Knuth
- TeX is free software
  - available on Linux distributions, Mac and Win
  - LaTeX is a popular packaging for TeX
  - MikTeX is a popular collection of tools for Windows (http://miktex.org/)

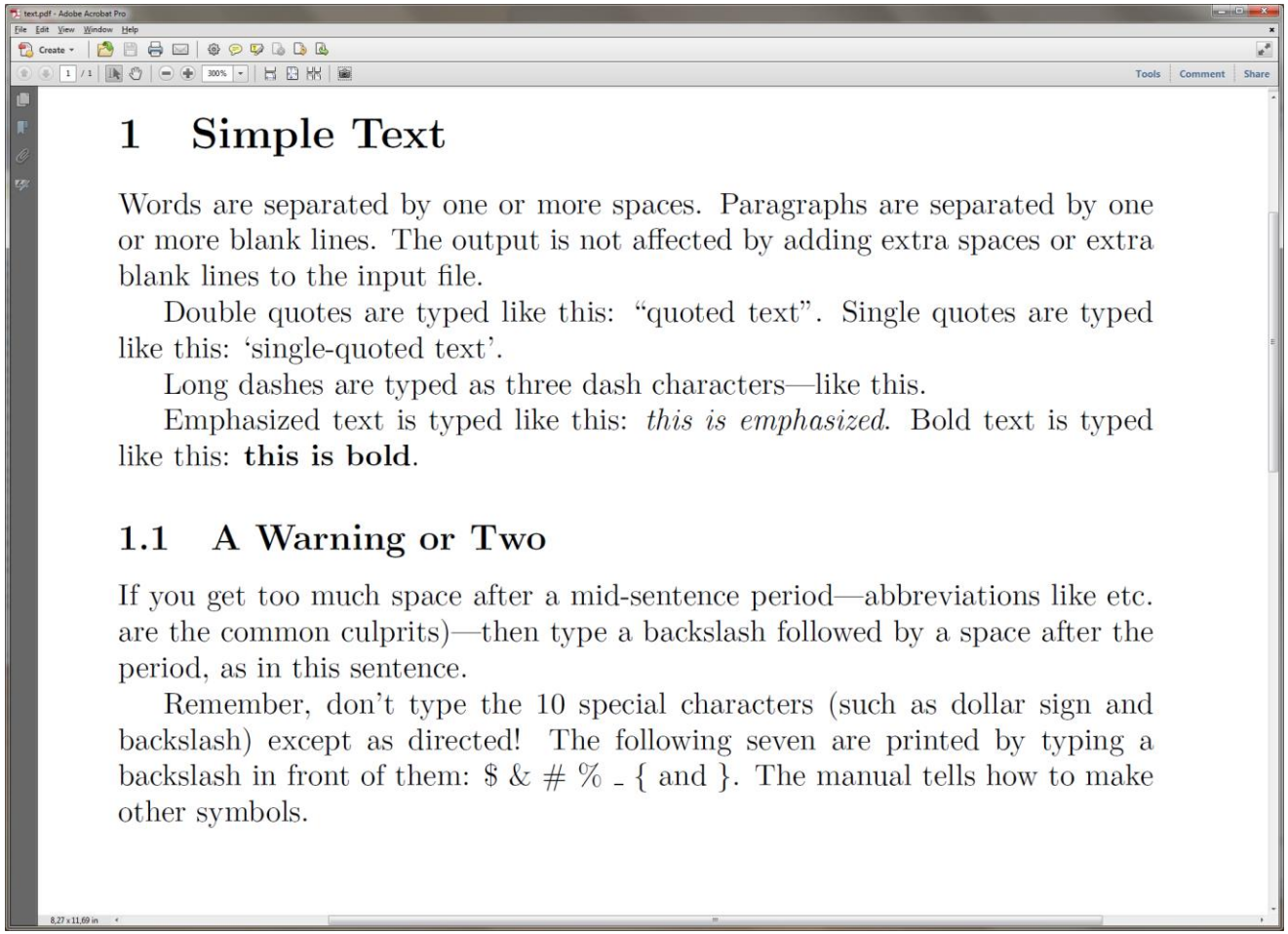
Sample LaTeX Document

```
\documentclass{article}
\begin{document}
\section{Simple Text}

Words are separated by one or more spaces.  Paragraphs are separated by one or more blank lines.
The output is not affected by adding extra spaces or extra blank lines to the input file.

Double quotes are typed like this: ``quoted text''.
Single quotes are typed like this: `single-quoted text'.

Long dashes are typed as three dash characters---like this.

Emphasized text is typed like this: \emph{this is emphasized}.
Bold       text is typed like this: \textbf{this is bold}.

\subsection{A Warning or Two}

If you get too much space after a mid-sentence period---abbreviations like etc.\ are the common
culprits)---then type a backslash followed by a space after the period, as in this sentence.

Remember, don't type the 10 special characters (such as dollar sign and backslash) except as directed!
The following seven are printed by typing a backslash in front of them:  \$  \&  \#  \%  \_  \{  and
\}. The manual tells how to make other symbols.

\end{document}
```

LaTeX

- Demo ...

Sample LaTeX Document



XML

- Successor of SGML
  - less complex, easier to read (for humans), easier to parse (for machines)
- Standardized by W3C
  - Builds a basis for many standards
  - SVG, XHTML, SMIL, MPEG-7, ...

XML

- Based on a tree model & Unicode
  - Supports document type declarations, etc.
  - Documents can be strictly defined by XML Schema
- Main components
  - root element, child elements, both with attributes
  - text in between elements

XML

- Two promising parsing models
- Document Object Model – DOM
  - the XML document is treated as a tree data structure
- Simple API for XML - SAX
  - event based model for parsing
  - sample events: document begins, element begins, etc.
  - no need to store document in-memory

XML Benefits & Drawbacks

- XML is a tree structure
  - simple to maintain, but complicated to use for graph data
- XML cannot be streamed
  - only the complete document can be validated#
- XML brings serious overhead
  - markups, DOM data structure, etc.

XML Examples

```
<?xml version="1.0" encoding="UTF-8" ?>
<俄语>данные</俄语>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Lightweight Markup Languages

- Simple (non complex) markup language
- Easy to read and write for humans
  - even with a simple text editor
- Used for
  - immediate editing, short texts
  - simple styled documents with clear guidelines

Bulletin Board Code - BBCode

- Used for many message boards
- Examples:
  - [b]bold text[/b]
  - [i]italicized text[/i]
  - [u]underlined text[/u]
  - [s]strikethrough text[/s]
  - [url]http://example.org[/url]
  - [url=http://example.com]Example[/url]
  - [img]http://www.cnn.com/test.png[/img]
  - [quote]quoted text[/quote]
  - [code]monospaced text[/code]

Wikitext

- Simple markup with bi-directional links
- Wiki Link: [[Another Site]]
  - link is resolved by wiki engine
  - target site is notified or created (implicitely)
- Wiki formatting
  - Different for different wiki engines, ie. Mediawiki, Dokuwiki, ...
  - eg. http://en.wikipedia.org/wiki/Help:Cheatsheet

Creole

- Light weight markup language

```
//emphasized// (e.g., italics)

**strongly emphasized** (e.g., bold)

* Bullet list
* Second item
** Sub item

# Numbered list
# Second item
## Sub item
```

See http://en.wikipedia.org/wiki/Creole_(markup)

Creole

```
Link to [[wikipage]], [[link_address|link text]]

= Extra-large heading (closing optional)
== Large heading
=== Medium heading
==== Small heading

Force\\linebreak

---- (horizontal line)

{{Image.jpg|title}}

|= |= table |= header |
| a | table | row |
| b | table | row

{{{
This text will //not// be **formatted**.
}}} |
```

Markdown

- Very simple text format
- Can be rendered to HTML, etc.
  - text markup itself is much like ASCII formatting
- Renderers are available in several languages

Sample Markdown Document

```
LIRE Solr Integration Project
=============================
Includes a RequestHandler and some utility
classes for a fast start.

The request handler supports four different
types of queries

1. Get random images ...
2. Get images that are looking like the one with id ...
3. Get images looking like the one found at url ...
4. Get images with a feature vector like ...

Preliminaries
-------------

Supported values for feature field parameters,
e.g. lireq?field=cl_ha:

-  **cl_ha** .. ColorLayout
-  **ph_ha** .. PHOG
-  **oh_ha** .. OpponentHistogram
-  **eh_ha** .. EdgeHistogram
-  **jc_ha** .. JCD
```

LIRE Solr Integration Project

Includes a RequestHandler and some utility classes for a fast start.

The request handler supports four different types of queries

1. Get random images ...
2. Get images that are looking like the one with id ...
3. Get images looking like the one found at url ...
4. Get images with a feature vector like ...

Preliminaries

Supported values for feature field parameters, e.g. lireq?field=cl_ha:

- **cl_ha** .. ColorLayout
- **ph_ha** .. PHOG
- **oh_ha** .. OpponentHistogram
- **eh_ha** .. EdgeHistogram
- **jc_ha** .. JCD

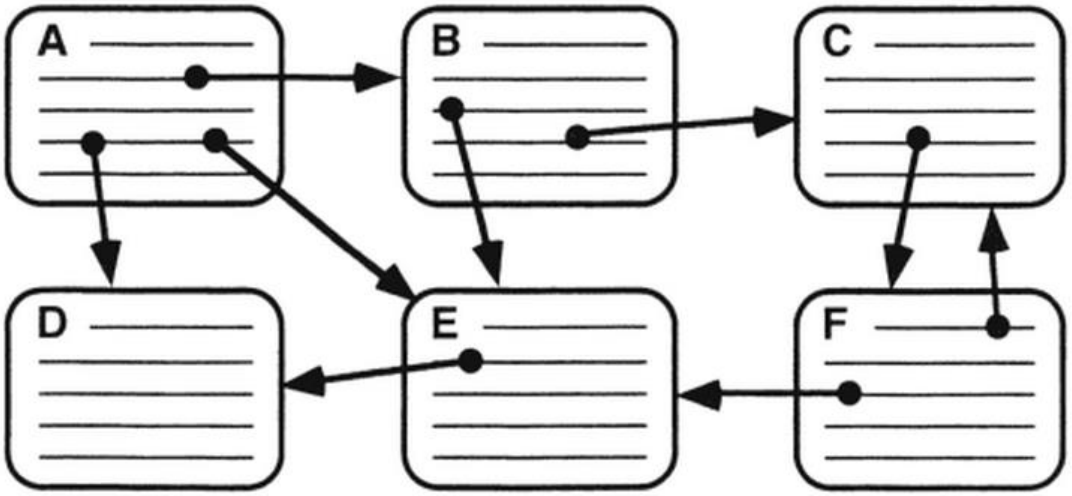
HTML & XHTML

- Prominent markup languages for the web
- HTML (<5) is based on SGML
- XHTML is based on XML
- HTML 5 is based on XML

- Both are used for creating hypertext systems

Hypertext – A Definition

- Text is printed & consumed sequentially
- Hypertext is non-sequential



Src. Nielsen J (1995) Multimedia and Hypertext: The Internet and Beyond, Morgan & Kaufmann

Hypertext: A Definition

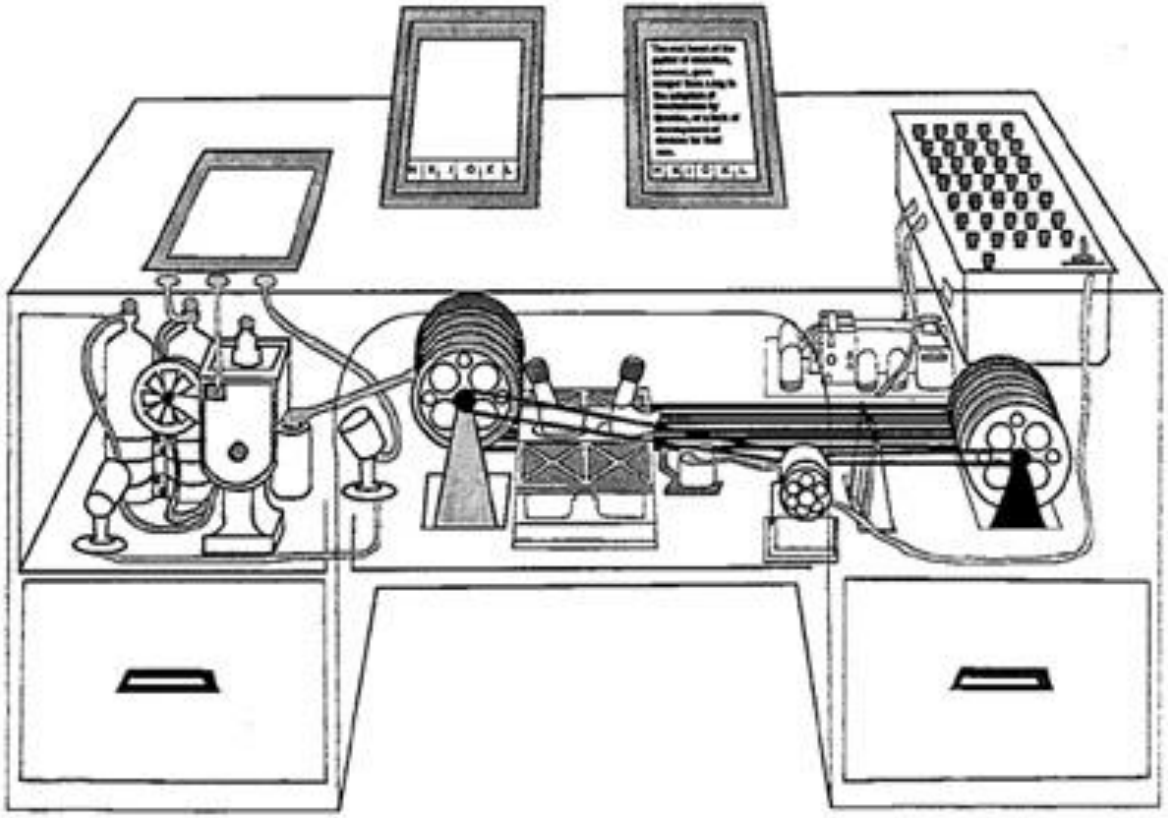
- Text has some "hypertext" elements
  - footnotes, indexes, (cross-)references, glossaries
- Hypertext is more general
  - hypertext nodes are connected by links
  - links can be bi- or unidirectional
  - links can be "typed"
    - footnote, related, include, ...
- Readers of hypertext traverse links

MEMEX

Memory **Ex**tender – Vannevar Bush

- Published in 1945 (Atlantic Monthly)
- An electromechanical device for
  - viewing books and films
  - adding information and comments
  - interlinking information
  - browsing links
- MEMEX is an early hypertext system

Memex



Uniform Resource Locators (URLs)

A URL consists of

- the scheme name (protocol)
- a colon and two slashes
- a host (domain name or IP address)
- a port number (optional)
- full path of the resource

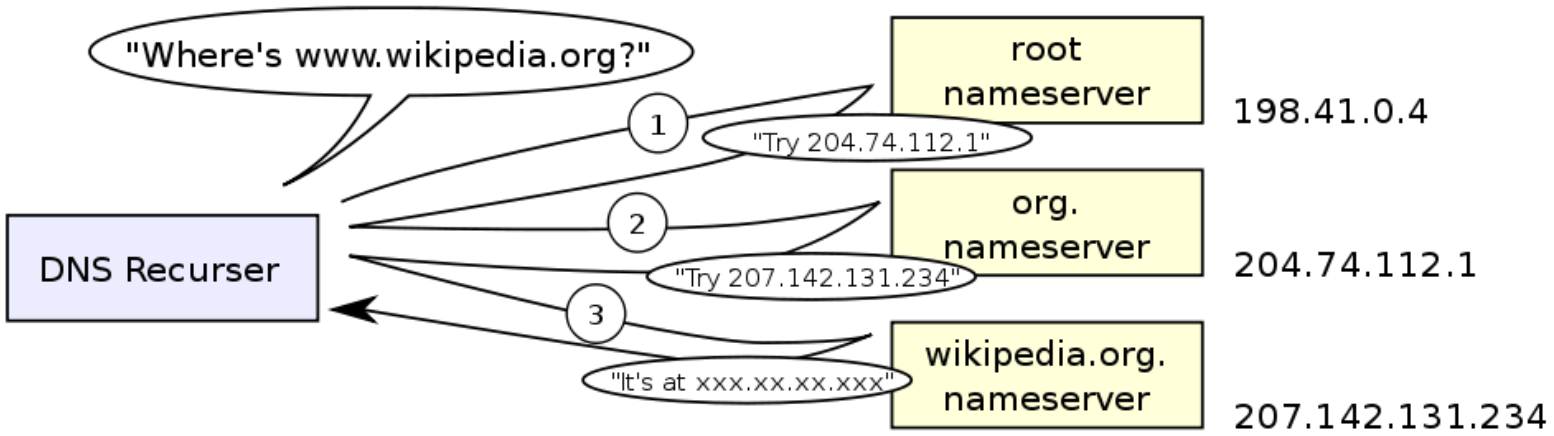
`https://code.google.com/p/lire/`

Domain Names

- A domain name is a string that defines a realm of authority
  - cnn.com, öamtc.at, uni-klu.ac.at, ...
- A top level domain defines country and/or intent of the domain
  - .com, .ac.at, ...
- A sub domain points to a specific IP address
  - www.aau.at, ftp.uni-klu.ac.at

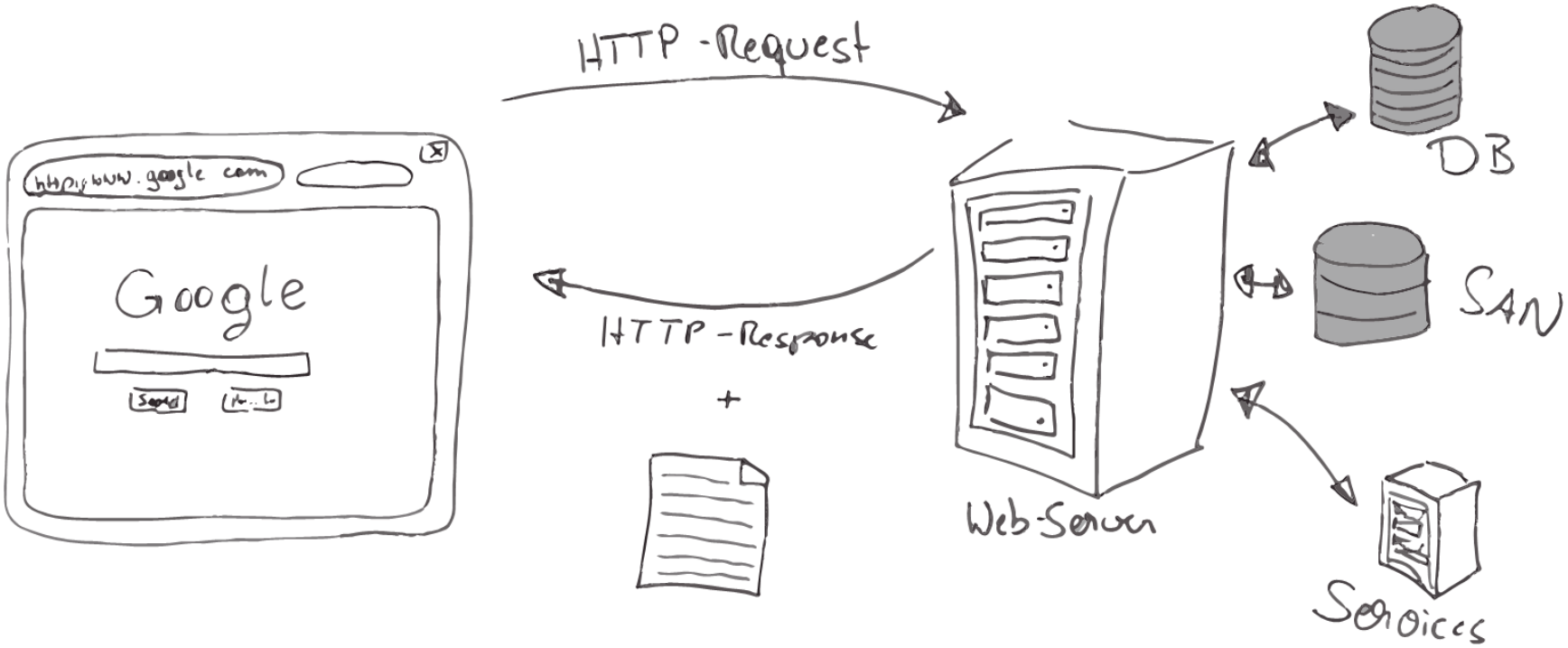
Domain Name System

- Hierarchical system where domain names are "registered"



Src. http://en.wikipedia.org/wiki/File:An_example_of_theoretical_DNS_recursion.svg

WWW


HTTP - Request
http://www.google.com
Google
HTTP - Response
+
DB
SAN
Web-Server
Services

WWW – Die Maus (WDR)

- http://www.youtube.com/watch?v=QKLz4ufCuKk

HTTP

Retrieve http://www.somehost.com/path/file.html

- Open socket to www.somehost.com:80
- Send something like this:

```
GET /path/file.html
HTTP/1.0
From: someuser@jmarshall.com
User-Agent: HTTPTool/1.0
[blank line here]
```

Src. http://www.jmarshall.com/easy/http/

HTTP Response

- 200 OK
  - The request succeeded, resource is returned in the message body.
- 404 Not Found
  - The resource doesn't exist.
- 302 Moved Temporarily
  - Used to redirect.
- 500 Server Error
  - An unexpected server error.

Sample Response

```
HTTP/1.0 200 OK
Date: Fri, 31 Dec 1999 23:59:59 GMT
Content-Type: text/html
Content-Length: 1354

<html>
  <body>
    <h1>Happy New Millennium!</h1>
    (more file contents) . . .
  </body>
</html>
```

Src. http://www.jmarshall.com/easy/http/

HTTP – Some more facts

- A web page typically is more than one file
- Browsers try to re-use HTTP connections
  - Ask for multiple files in one single connection
  - „keep alive“
- HTTP is based on TCP & defaults to port 80
  - Works well behind firewalls
- Other services try to use HTTP too
  - HTTP video streaming, etc.

HTML Document Structure

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML
  4.01 Transitional//EN"
  "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
  <title>
    Beschreibung der Seite
  </title>
</head>
<body>
</body>
</html>
```

HTML Document Structure