Web Engineering & Design 1

HTML

Markup languages

Tobias Blaser

HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

FHO Fachhochschule Ostschweiz

Content

- What is markup?
- Markup languages
- SGML
- Encoding

Learning goals

You are able to...

- ...explain markup languages, compare them to programming languages and set examples of both.
- ...evaluate which parts of a document markup and which parts data are.
- ...illustrate what effect the encoding of a document has and set an example.
- ...draw how UTF-8 is structured.
- ...fix a document encoded the wrong way.

Simple text documents

Do you remember text only emails?

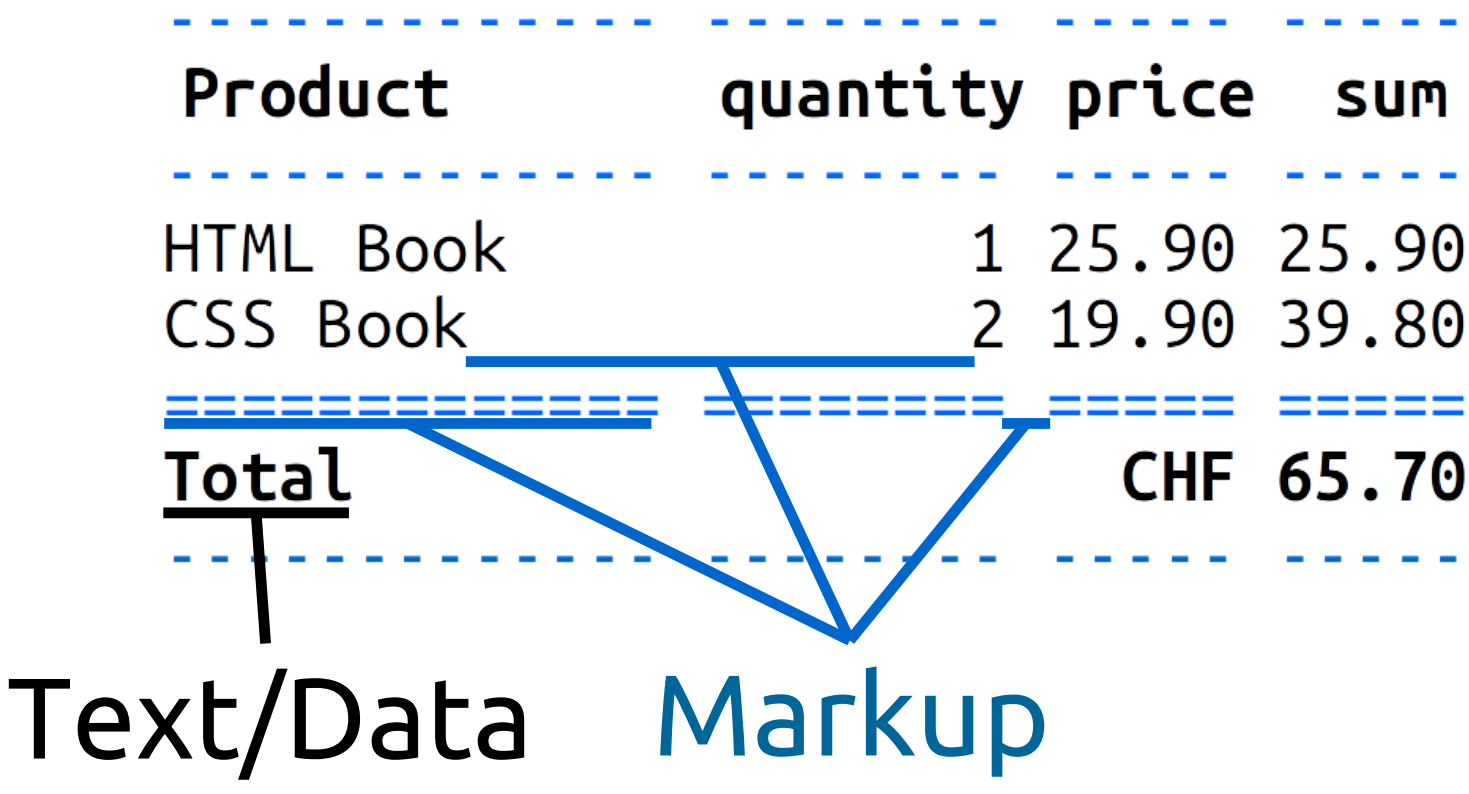
```
Dear Tim Sanders

Thanks for your order. Your order number is 1234-5678.

------------ -------- ----- -----
Product       quantity price  sum
------------ -------- ----- -----
HTML Book            1 25.90 25.90
CSS Book             2 19.90 39.80
============ ======== ===== =====
Total                  CHF 65.70
------------ -------- ----- -----

Best regards

Your Skyshop team
```



«A markup language is a system for annotating a document in a way that is syntactically distinguishable the text.»

Wikipedia

Markup

- Data + **Markup** → Information
- Annotate the data → labels or white spaces
- Clarifies the intention:

```
«The **boss** is out of office.»
```

«The **boss** is out of office.»

→ focus on *who*

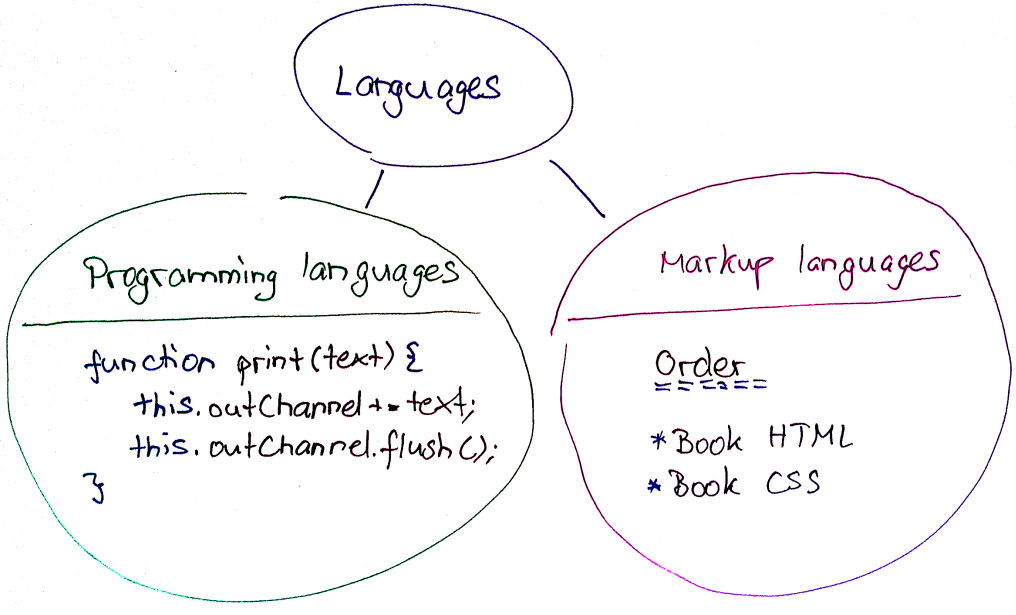
```
«The boss is **out of office**.»
```

«The boss is **out of office**.»

→ focus on *what*

➡ Adds semantics

Markup Languages



Markup Language Types

- Semantic markup
- Presentation markup

Markup languages vs. programming languages

Programming language

- Defines **instructions** for a machine → flow
- Mostly imperative → executed sequentially

Markup language

- Declarative
- Defines a meaning → semantic **annotations**
- May be translated into **different outputs**

  Example *LaTeX*:
  - text document (article/book package)
  - presentation (beamer package)

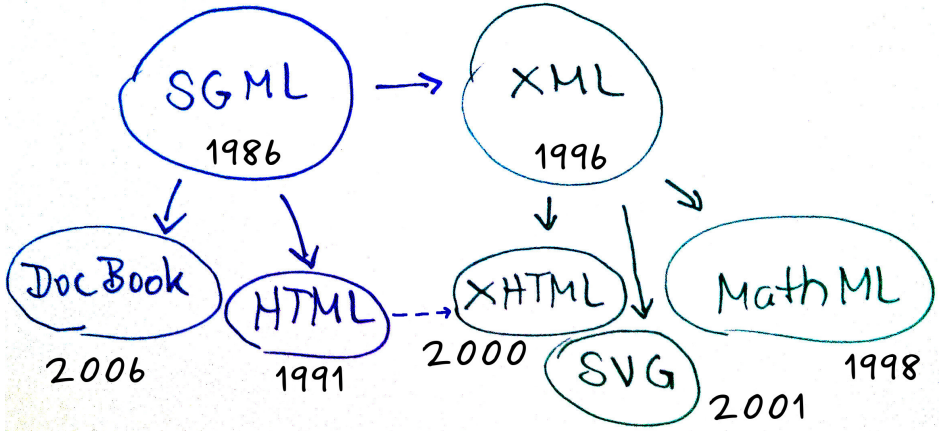
Examples

	Markdown	Latex	HTML
Title	`Title` `=====`	`\section{Title}`	`<h1>Title</h1>`

Item 1
Item 2
	`* Item 1`
	`* Item 2`
		`\begin{itemize}`
		`  \item Item 1`
		`  \item Item 2`
		`\end{itemize}`
			`<ul>`
			`  <li>Item 1</li>`
			`  <li>Item 2</li>`
			`</ul>`

Loud	`**Loud**`	`\important{Loud}`	`<strong>Loud</strong>`

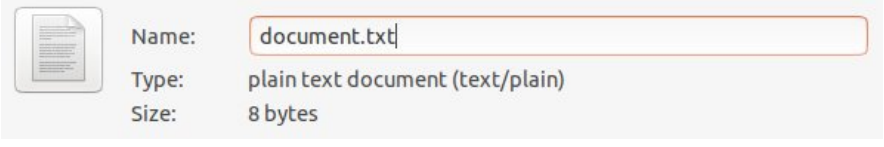
SGML



- Markup language definition language
- Brackets: `< >`
- Tags: `<item></item>`
- Entities: `&lt;   &amp;`

Encoding

What contains the following document ?




```
hexdump -x "document.txt"
```

```
0000000 c35a 72bc 6369 0068 0000007
```

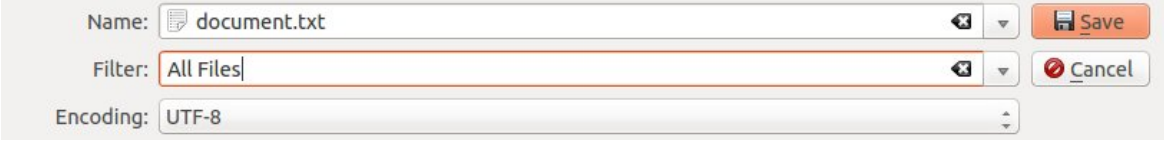
→ We don't know even we know **how to interprete** it!

Encoding = How to interprete the bytes of a document

Content	Encoding	Bytes (Hex)
"Zuerich"	ISO 8859-1	`0000000 755a 7265 6369 0a68 0000008`
"Zuerich"	UTF-8	`0000000 755a 7265 6369 0068 0000007`
"Zuerich"	UTF-16	`0000000 005a 0075 0065 0072 0069 0063  0068 000a 0000010`
"Zürich"	ISO 8859-1	`0000000 fc5a 6972 6863 0000006`
"Zürich"	UTF-8	`0000000 c35a 72bc 6863 0000006`
"Zürich"	UTF-16	`0000000 005a 00fc 0072 0069 0063 0068 000000c`

- **UTF-8**: most commong encoding

➡ Use whenever possible **UTF-8**



UTF-8

Universal Coded Character Set + Transformation Format

- variable-length

Character		Binary UTF-8	Hexadecimal UTF-8
$	U+0024	00100100	24
¢	U+00A2	11000010 10100010	C2 A2
€	U+20AC	11100010 10000010 10101100	E2 82 AC

Source: en.wikipedia.org/wiki/UTF-8

- 8-bit code units
- ASCII-backward compatible → first 128 characters

Wrong encoding

What happens, if the characters are decoded the wrong way?

```
0000000 c35a 72bc 6369 0068 0000007
UTF-8 Big endian:
|Z |ü   |r |i |c |h |
|5a|c3bc|72|69|63|68|

ISO 8859-1:
|5a|c3|bc|72|69|63|68|
|Z |Ã |¼ |r |i |c |h |
```

- Example UTF-8/ISO 8859-1
- 2-byte UTF-8 characters will be interpreted as 2 different characters
- ü → c3bc → c3 bc → Ã¼

Web Engineering & Design 1

HTML

Markup languages

Tobias Blaser

HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

FHO Fachhochschule Ostschweiz

Self Check

1. What's the goal of markup?
2. What's the difference between "data" and "information"?
3. What's the difference between a markup language and a programming language?
4. What is SGML?
5. Why should you use UTF-8 instead of ASCII to encode your documents?
6. UFT-8 is "variable-length". How does this affect on byte level?

Solutions

1. Add a meaning to data
2. Data: mud of characters, Information: data + meaning
3. markup language: declarative, content meaning
programming language: iterative, programm flow
4. Markup definition language
5. ASCII contains 256 characters but no european language characters (→ ISO 8859-1). UTF-8 contains most of the characters of the world.
6. Early defined chars are short encoded (one or two bytes), later added characters are encoded by multiple bytes.

Web Engineering & Design 1

HTML

Markup languages

Tobias Blaser

HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

FHO Fachhochschule Ostschweiz