# Document and Content Analysis

Summer 2009

Lecture 5
Layout and Markup

Thomas Breuel
Faisal Shafait

# Outline

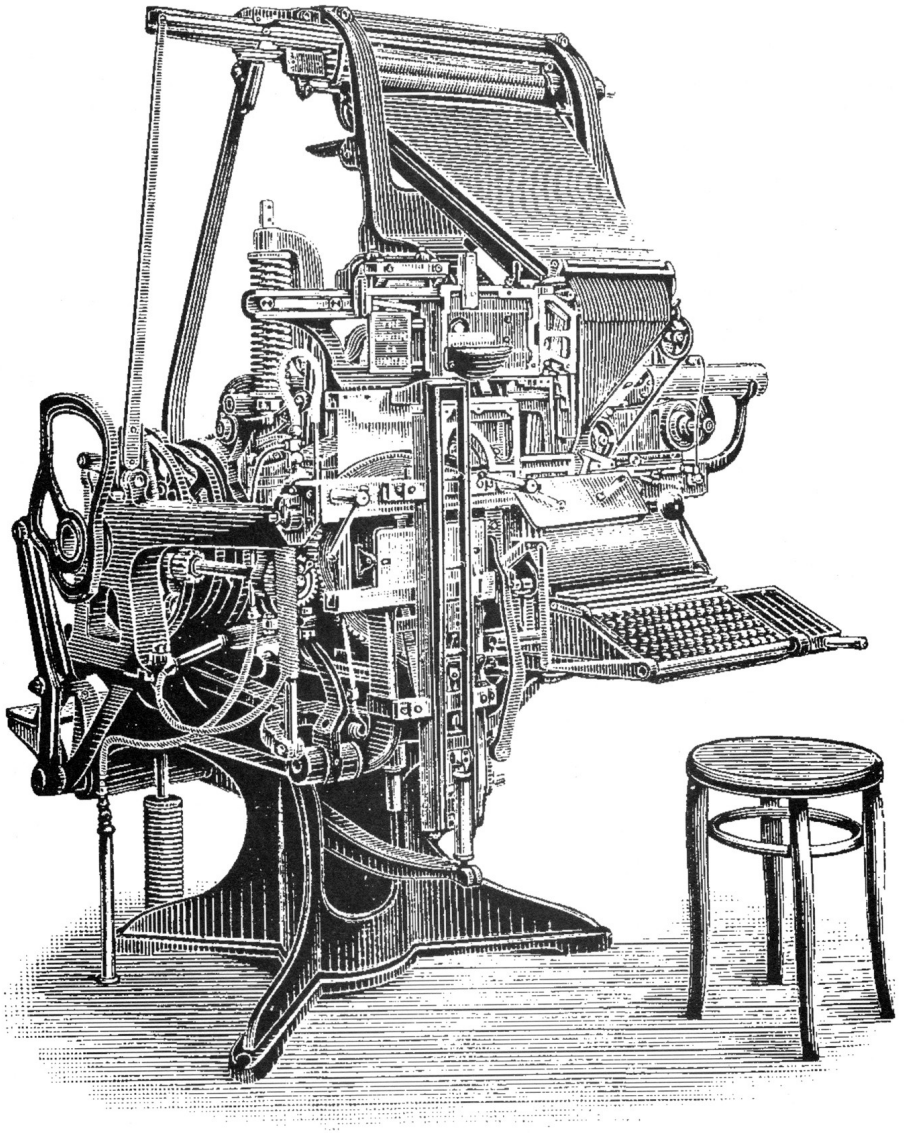
- markup
- logical vs physical
- TeX, LaTeX, SGML, DocBook, HTML, CSS
- wiki formats
- microformats
- typesetting, page layout, reflow

pre-computer output

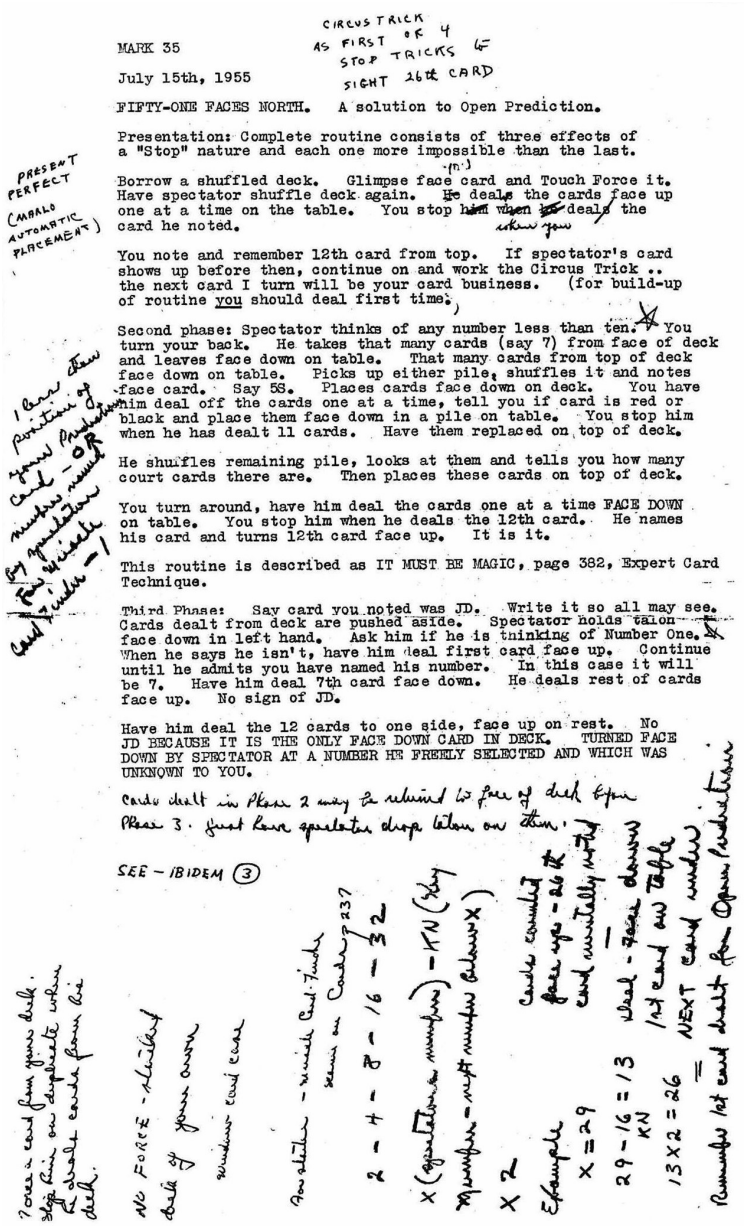
# stylus and pen

# typesetter

# typesetter



# typewriter



MARK 35

July 15th, 1955

FIFTY-ONE FACES NORTH. A solution to Open Prediction.

Presentation: Complete routine consists of three effects of a "Stop" nature and each one more impossible than the last.

Borrow a shuffled deck. Glimpse face card and Touch Force it. Have spectator shuffle deck again. He deals the cards face up one at a time on the table. You stop him when you deals the card he noted.

You note and remember 12th card from top. If spectator's card shows up before then, continue on and work the Circus Trick .. the next card I turn will be your card business. (for build-up of routine you should deal first time.)

Second phase: Spectator thinks of any number less than ten. You turn your back. He takes that many cards (say 7) from face of deck and leaves face down on table. That many cards from top of deck face down on table. Picks up either pile, shuffles it and notes face card. Say 5S. Places cards face down on deck. You have him deal off the cards one at a time, tell you if card is red or black and place them face down in a pile on table. You stop him when he has dealt 11 cards. Have them replaced on top of deck.

He shuffles remaining pile, looks at them and tells you how many court cards there are. Then places these cards on top of deck.

You turn around, have him deal the cards one at a time FACE DOWN on table. You stop him when he deals the 12th card. He names his card and turns 12th card face up. It is it.

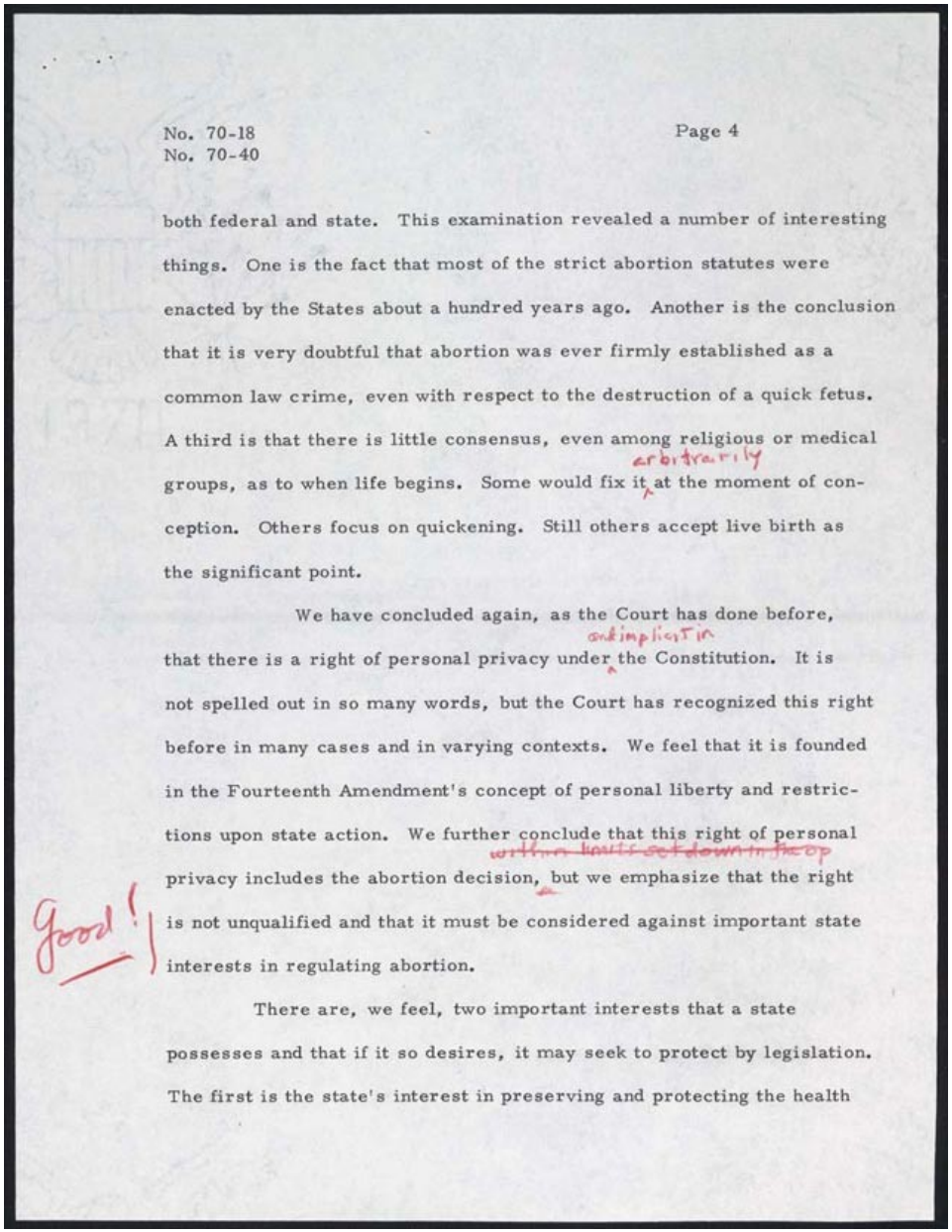
This routine is described as IT MUST BE MAGIC, page 382, Expert Card Technique.

Third Phase: Say card you noted was JD. Write it so all may see. Cards dealt from deck are pushed aside. Spectator holds talon face down in left hand. Ask him if he is thinking of Number One. When he says he isn't, have him deal first card face up. Continue until he admits you have named his number. In this case it will be 7. Have him deal 7th card face down. He deals rest of cards face up. No sign of JD.

Have him deal the 12 cards to one side, face up on rest. No JD BECAUSE IT IS THE ONLY FACE DOWN CARD IN DECK. TURNED FACE DOWN BY SPECTATOR AT A NUMBER HE FREELY SELECTED AND WHICH WAS UNKNOWN TO YOU.

SEE - IBIDEM (3)





No. 70-18
No. 70-40



both federal and state. This examination revealed a number of interesting things. One is the fact that most of the strict abortion statutes were enacted by the States about a hundred years ago. Another is the conclusion that it is very doubtful that abortion was ever firmly established as a common law crime, even with respect to the destruction of a quick fetus. A third is that there is little consensus, even among religious or medical groups, as to when life begins. Some would fix it arbitrarily at the moment of conception. Others focus on quickening. Still others accept live birth as the significant point.

We have concluded again, as the Court has done before, that there is a right of personal privacy under the Constitution. It is not spelled out in so many words, but the Court has recognized this right before in many cases and in varying contexts. We feel that it is founded in the Fourteenth Amendment's concept of personal liberty and restrictions upon state action. We further conclude that this right of personal privacy includes the abortion decision, but we emphasize that the right is not unqualified and that it must be considered against important state interests in regulating abortion.

Good!

There are, we feel, two important interests that a state possesses and that if it so desires, it may seek to protect by legislation. The first is the state's interest in preserving and protecting the health

# pre-computer output

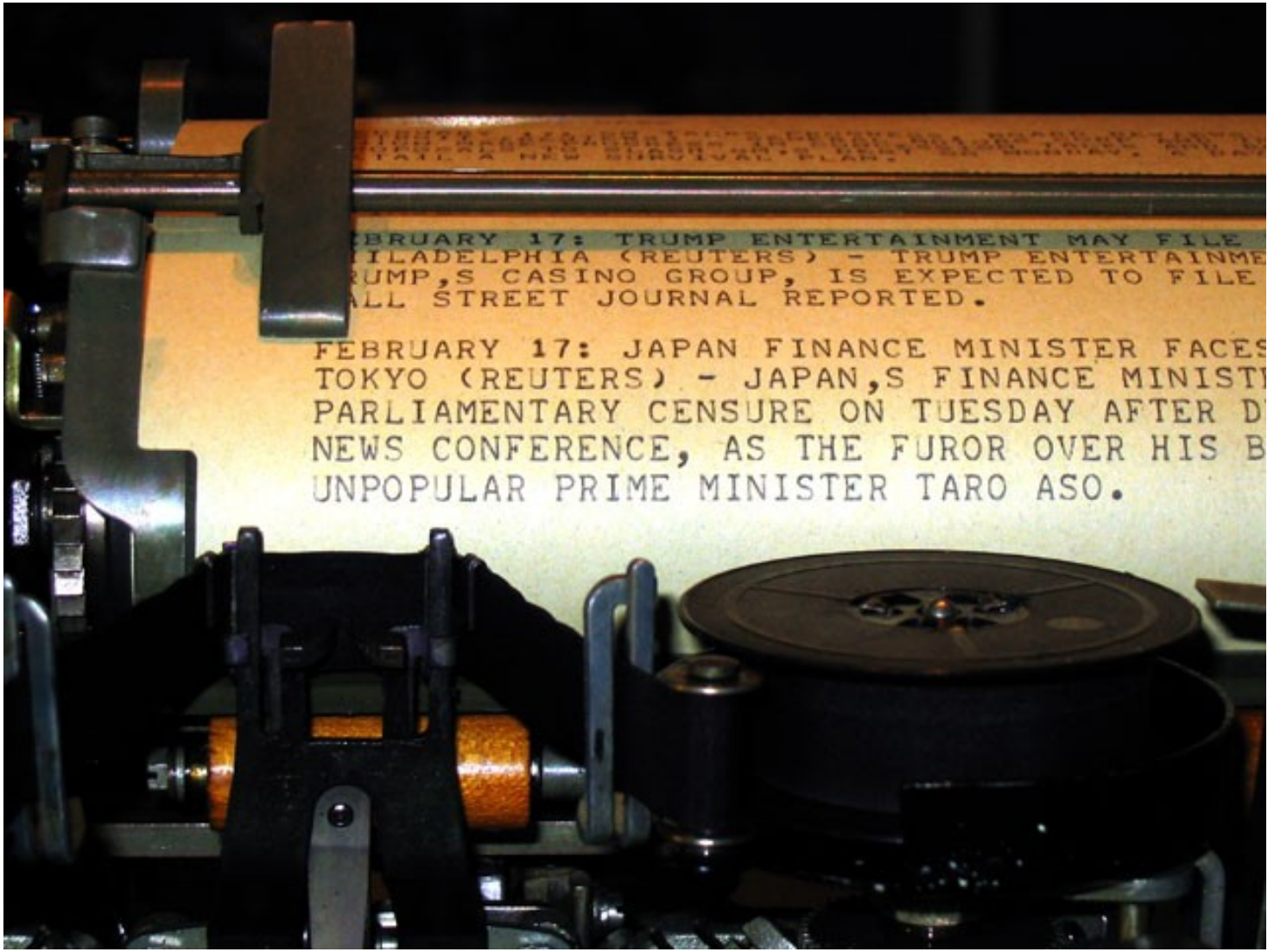
- **do anything that the medium permits**
  - arbitrary positioning
  - manual annotations
  - drawing lines, scribbles, ...

computer printers

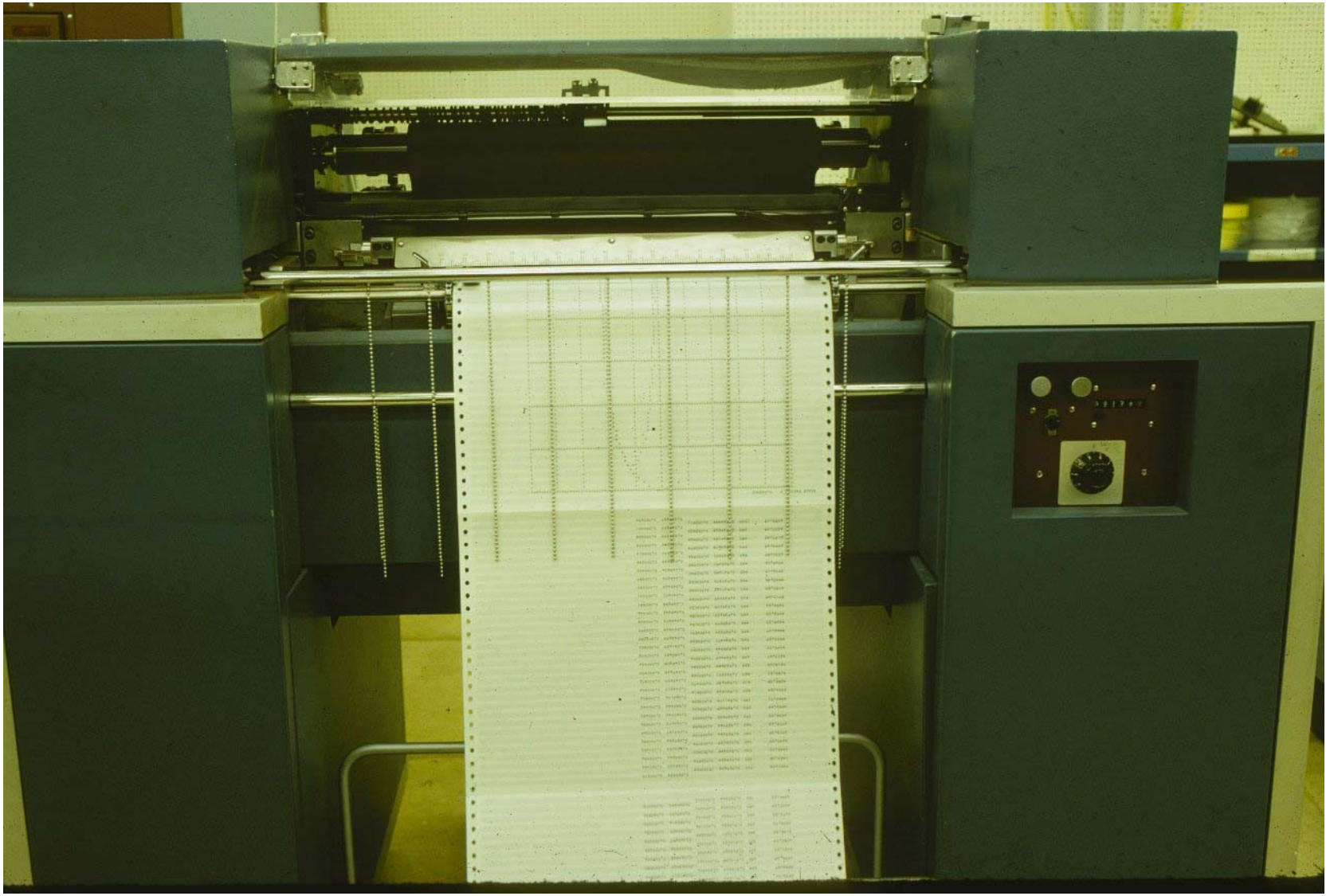
# printing technologies

- **teletype**
  - upper case, monospaced
- **electronic typewriters, line printers**
  - full character set, monospaced
- **photographic typesetting machines**
  - high quality, proportional spacing
- **dot matrix, laser printers**
  - pixel-addressable, render any b/w image

# teletype


BRUARY 17: TRUMP ENTERTAINMENT MAY FILE
ILADELPHIA (REUTERS) - TRUMP ENTERTAINME
UMP,S CASINO GROUP, IS EXPECTED TO FILE
LL STREET JOURNAL REPORTED.
FEBRUARY 17: JAPAN FINANCE MINISTER FACES
TOKYO (REUTERS) - JAPAN,S FINANCE MINISTE
PARLIAMENTARY CENSURE ON TUESDAY AFTER D
NEWS CONFERENCE, AS THE FUROR OVER HIS B
UNPOPULAR PRIME MINISTER TARO ASO.

# line printer



# IBM Selectric

# editing and markup

- **typewriter**
  - everything is manual: character placement, line breaks
- **early editor**
  - grid of character model
  - each character just gets put where you write it
  - pagination done by printer
  - long lines fall off the edge
- **problem**
  - you have to manually redo layout after any change
  - (still beats re-typing)

# early markup tasks

- **output**
  - line printer
- **input**
  - monospaced text from text editor (no WYSIWYG)
- **automate**
  - line breaks (no truncation)
  - page breaks (other than physical breaks)
  - headers / footers / line numbers

# RUNOFF

- .line length
- .indent
- .single space
- .double space
- .begin page
- .header
- .break
- .adjust
- .nojust
- .fill
- .nofill
- .space
- .center
- .literal

Saltzer, 1964

http://mit.edu/Saltzer/www/publications/CC-244.html

# photographic type setting

*ABCDEFGHIJKLMNOP*
*QRSTUVWXYZabcdefghi*
*jklmnopqrstuvwxyz*
*1234567890(.;!?&$£)*

*Qa*

# troff typesetting

- **troff is an evolution of RUNOFF / roff**
  - proportional spacing
  - multiple fonts
  - arbitrary character placement
  - macro packages
  - mathematics and tables
- **output**
  - photographic typesetter (now emulated on modern devices)
  - can also output to monospaced device (nroff)
  - still used for manual pages

# troff input / output

```
.\"  Questo è un esempio di documento scritto utilizz
.\"  di composizione Troff.
.\"
.\"  Viene definita la dimensione del testo: il margi
.\"  4 cm, e l'ampiezza del testo di 8 cm.
.po 4c
.ll 8c
.\"  Inizia il documento.
.ft B
1. Introduzione a Troff

.ft P
Questo \(`e un esempio di documento scritto in modo
tale da poter essere elaborato con Troff.
In questo caso, si presume che verr\(`a utilizzato
lo stile ``\fBs\fP'' (con l'opzione \fB\-ms\fP).

.ft B
1.1 Paragrafi
```

**1. Introduzione a Troff**

Questo è un esempio di documento scritto in modo tale da poter essere elaborato con Troff. In questo caso, si presume che verrà utilizzata lo stile **"s"** (con l'opzione **–ms**).

**1.1 Paragrafi**

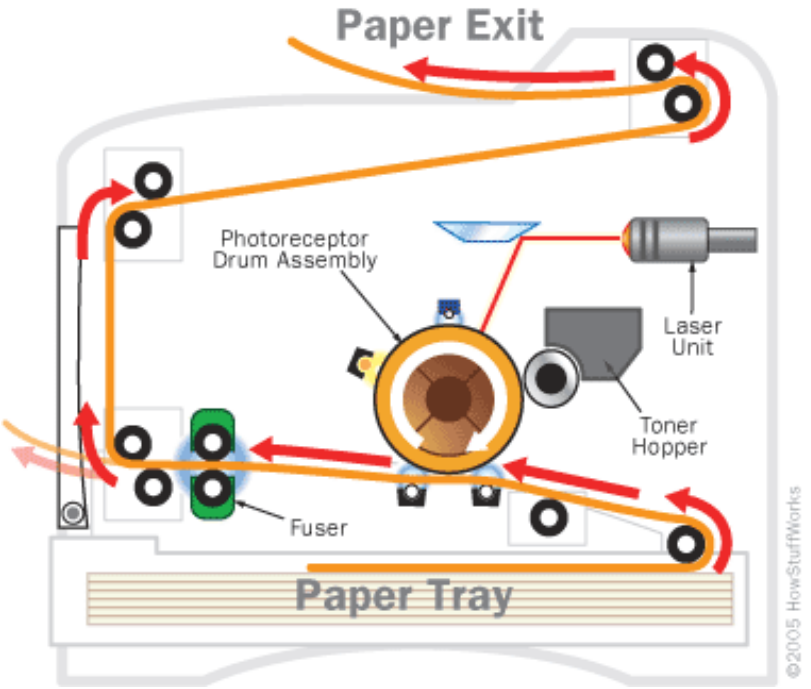
Il testo di un paragrafo termina quando nel sorgente viene incontrata una riga vuota.

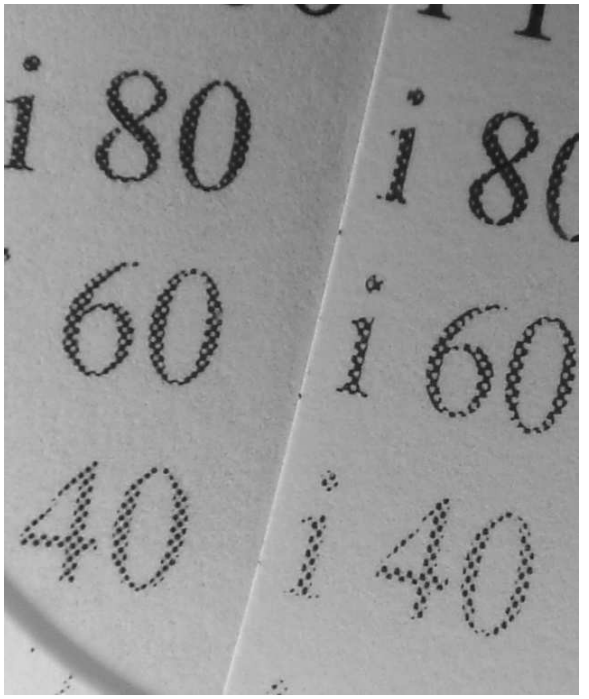
Per la precisione, gli spazi verticali vengono rispettati, per cui le righe vuote si traducono in spazi tra i paragrafi, anche quando queste sono più di una.

# dot matrix printers

ystem where a ...i
ld allow us t... t
mercial supplier.

# laser printers



# TeX / Metafont / LaTeX

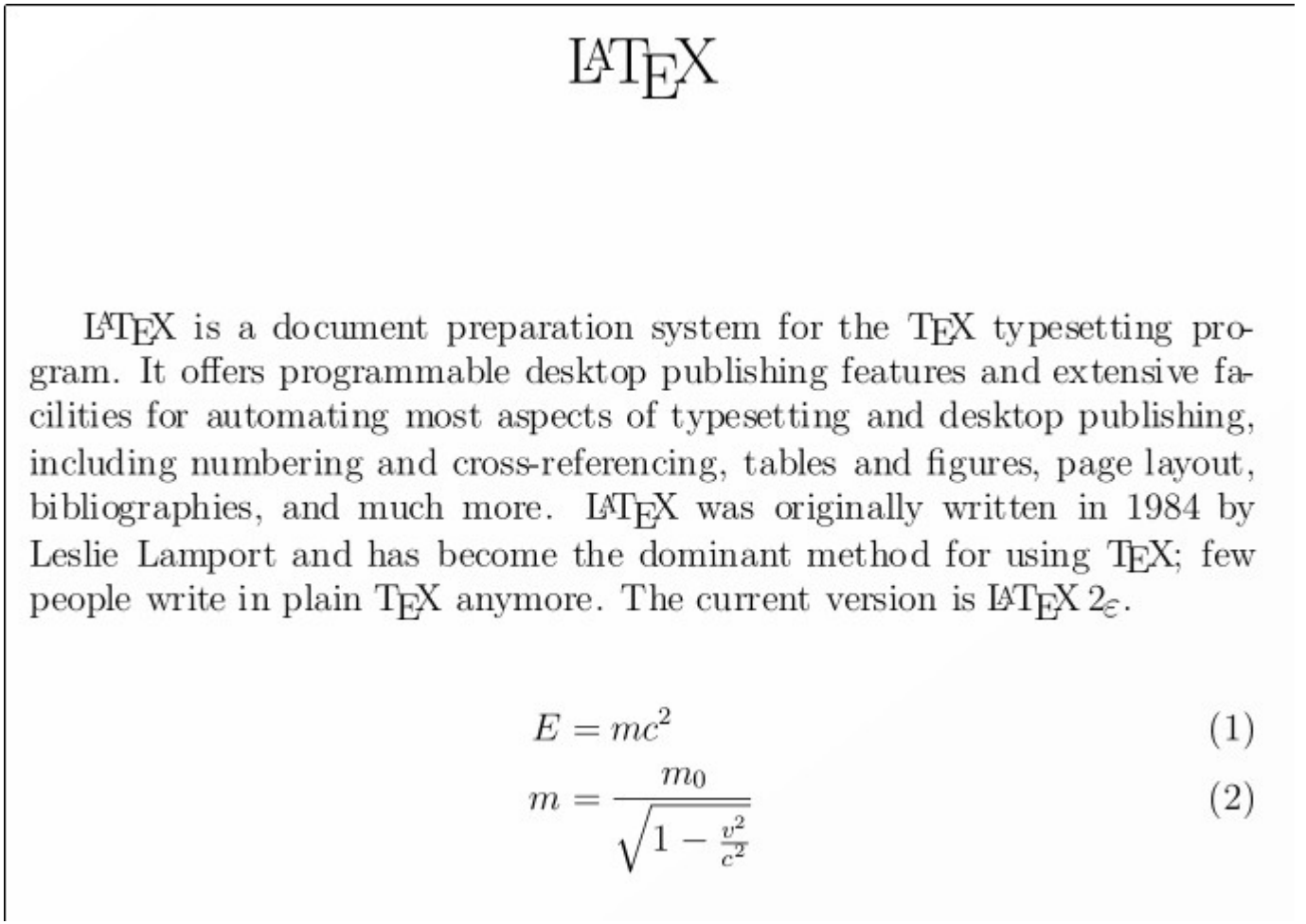
- **new features**
  - output to bitmapped devices
  - scalable fonts in METAFONT
  - elaborate built-in macro language
  - new hyphenation algorithm
  - elaborate optimization algorithms (line/page breaking, ...)
  - built-in mathematics support
  - syntax partially fixed by LaTeX macro package

  - written by Donald Knuth for writing computer science papers

# TeX / LaTeX

```
\documentclass[12pt]{article}
\usepackage{amsmath}
\title{\LaTeX}
\date{}
\begin{document}
  \maketitle
  \LaTeX{} is a document preparation system for the \TeX{}
  typesetting program. It offers programmable desktop publishing
  features and extensive facilities for automating most aspects of
  typesetting and desktop publishing, including numbering and
  cross-referencing, tables and figures, page layout, bibliographies,
  and much more. \LaTeX{} was originally written in 1984 by Leslie
  Lamport and has become the dominant method for using \TeX; few
  people write in plain \TeX{} anymore. The current version is
  \LaTeXe.

  % This is a comment, it is not shown in the final output.
  % The following shows a little of the typesetting power of LaTeX
  \begin{align}
    E &= mc^2                                   \\
    m &= \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}
  \end{align}
\end{document}
```

# TeX / LaTeX



LaTeX

LaTeX is a document preparation system for the TeX typesetting program. It offers programmable desktop publishing features and extensive facilities for automating most aspects of typesetting and desktop publishing, including numbering and cross-referencing, tables and figures, page layout, bibliographies, and much more. LaTeX was originally written in 1984 by Leslie Lamport and has become the dominant method for using TeX; few people write in plain TeX anymore. The current version is LaTeX $2_\epsilon$.

$$E = mc^2 \tag{1}$$

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2}$$

# factors driving markup

- **input**
  - monospaced text editors
- **output**
  - line printer → typesetter → laser printer
- **document change**
  - reflow
  - pagination
  - style sheets...

# style sheets

## *Forest Products Journal* Publication Style Sheet

The *Forest Products Journal's* primary purposes include communicating research findings at the applied or practical level, communicating news and items of current interest to the membership, and describing Society programs and activities. Technical manuscripts submitted for publication are reviewed in a double-blind review process by referees selected by the Editor. Final evaluation of the material rests with the Editor.

### Form of Manuscript

The original submission of a manuscript for review must be sent in both electronic form and on a double-spaced printed hard copy. Mail one hard copy plus one floppy disk or CD that contains the manuscript in an MS Word file. After the author page, the sequence of material in the manuscript as submitted should be: title page, abstract, text, literature cited, captions for figures, tables, and figures (i.e., drawings and photographs). Do not incorporate tables into the text. Manuscripts should be submitted in correct English form. Authors speaking English as a second language should have their manuscripts edited by a native English speaker prior to original submission.

A. Authors' names, titles, affiliations, complete addresses of the affiliation, and e-mail addresses should be included on a separate page. If acknowledgments are necessary, they should be written as a footnote on the author page. The title page should only include the title of the manuscript, which should be as concise as possible.

B. The abstract should contain, in very condensed form (250 words for an article, 75 for a note), the essence of the whole work. It should summarize why the work was done, what was done and how, and results and conclusions, perhaps with a mention of the significance.

C. In the text, make certain all figures, tables, and references are mentioned. If fewer than six references are mentioned, they should be typed as footnotes at the bottom of the page. If six or more references are mentioned, they should be cited in parentheses at the appropriate location in the text using the author-date style. For example: . . . (Brown and Banks 1985, Adams 1989, Evans et al. 1999) — in chronological order.

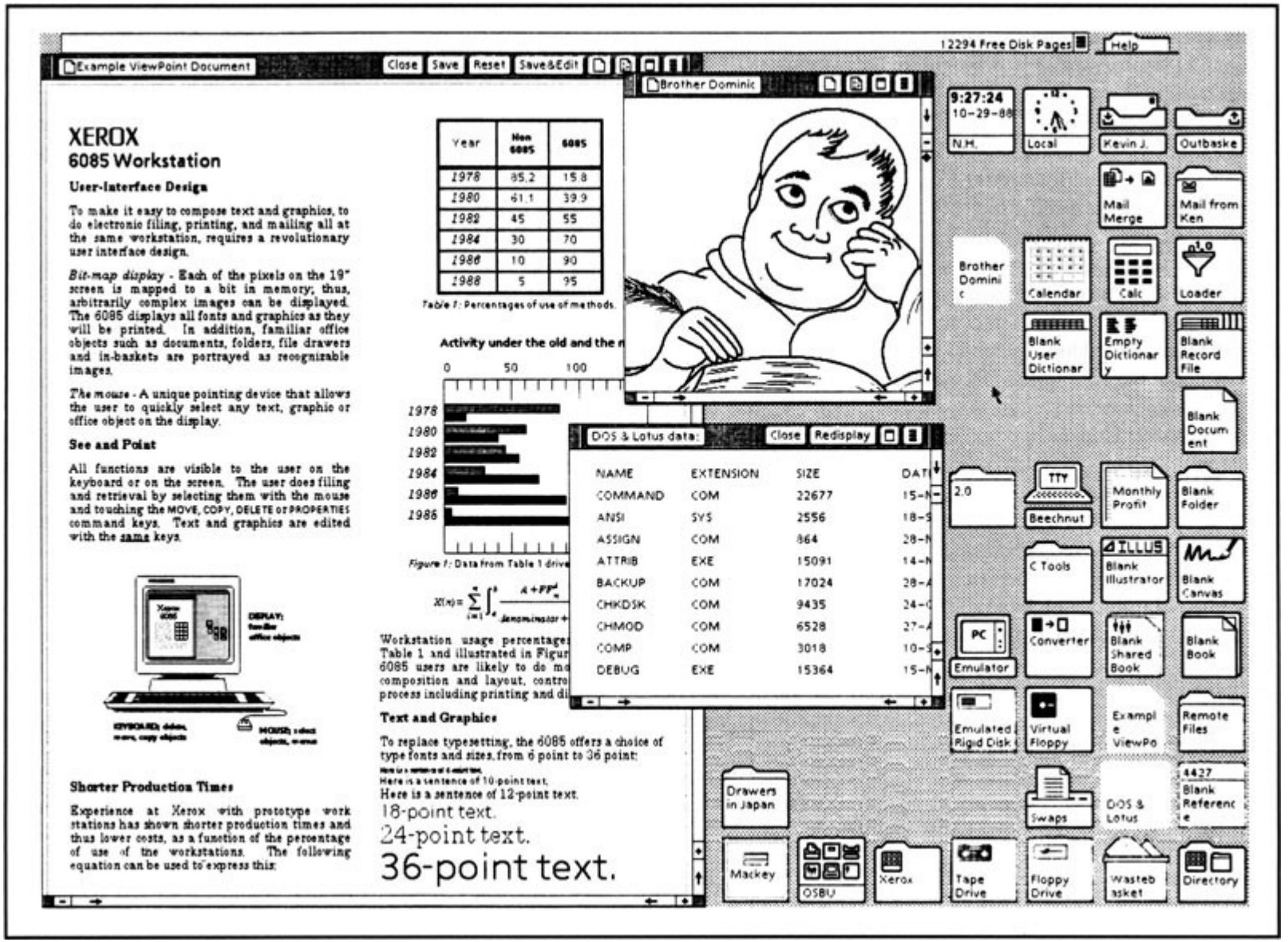
# physical vs logical markup

- **\centerline{\bf John Smith}**
  - typeset “John Smith” in bold face, centered
- **\author{John Smith}**
  - “John Smith” is the author
  - the style says that the author is typeset in bold face, centered

# physical vs logical markup

- **change of style**
  - physical markup: edit all instances of markup
  - logical markup: just change the electronic style sheet
- **the evolution of logical markup was driven by editing and printing**
- **using logical markup as semantic information came later**

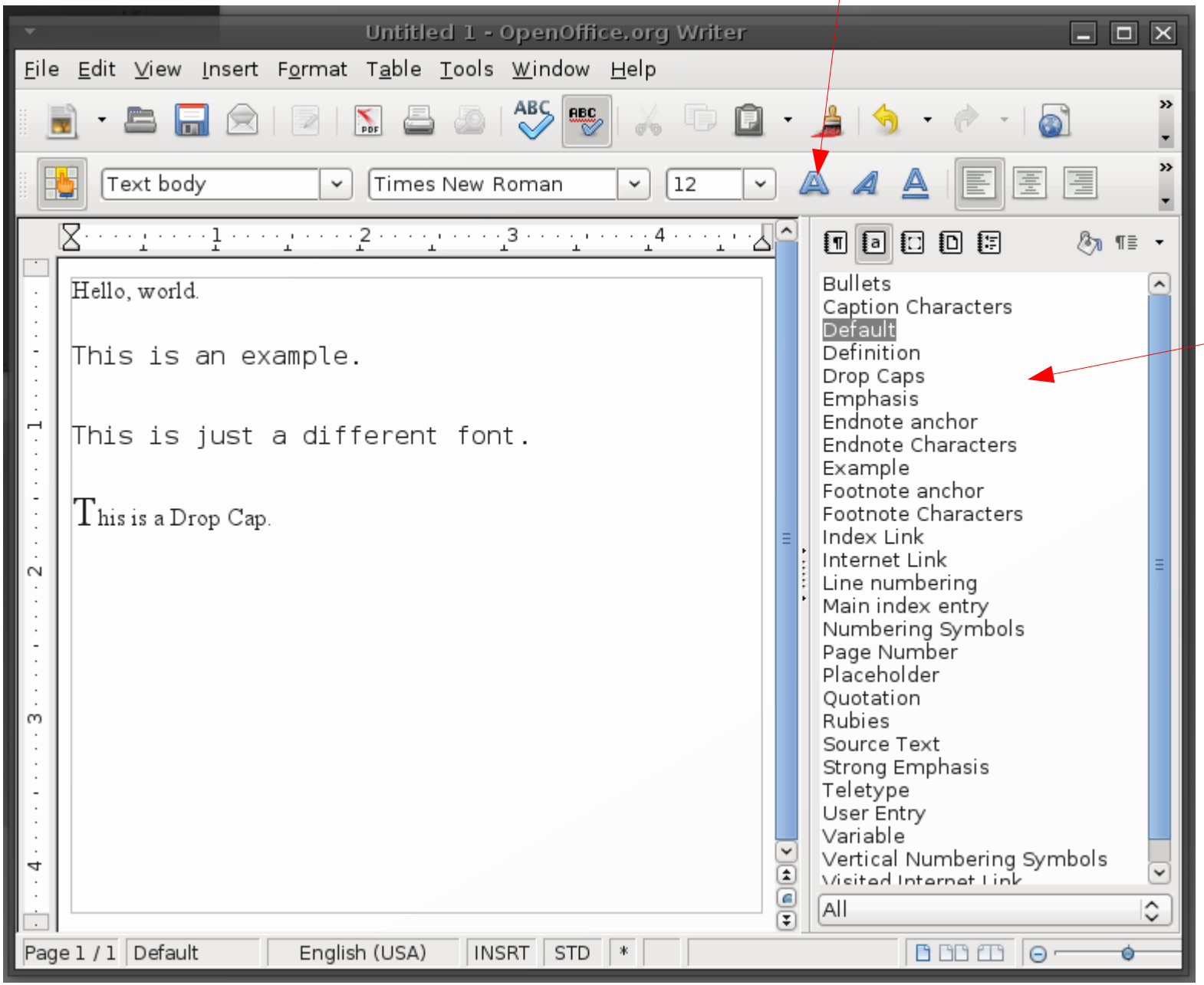
# WYSIWYG editing



# WYSIWYG editing

- **WYSIWYG**
  - what you see is what you get
  - enablers: bitmapped displays, fast processors
  - display typeset document directly on screen
  - line breaks, pagination updated while typing
  - algorithms similar to TeX etc.
  - easy to change formatting and see effect
- **big issue...**
  - how does logical markup work for WYSIWYG

# physical vs logical markup



# markup

- **commands embedded in text**
  - determine formatting after processing
- **physical markup**
  - changes appearance directly
- **logical markup**
  - changes appearance based on a style sheet
  - may convey semantic information

# SGML, HTML, XML

# SGML

- **standard generalized markup language**
- **developed in the 1960's for documentation**
- **specific doc syntax defined by DTD**

# SGML

- **abbreviations to make it easier to type**
  - <QUOTE TYPE=example> instead of “example”
  - <ITALICS/word/
  - <QUOTE//
  - </>
  - <QUOTE><ITALICS>word</QUOTE>
- **(putting lipstick on a dog)**

# DocBook

- **SGML, now XML**
- **semantic markup—computer documentation**
- **widely used for open source now**
  - Linux Documentation, Gtk+, KDE, ...
- **non-WYSIWYG**

# XML

- successor to SGML
- cleaned up syntax and semantics
- better defined
- document schemas and tools
- transformation languages (XSLT, ...)

# XML vision (?)

- **web sites send semantically marked up data to web browser**
- **web sites send XML → HTML transformation rules to browser**
- **browser puts the two together and renders a page**
- **client-side XSLT**

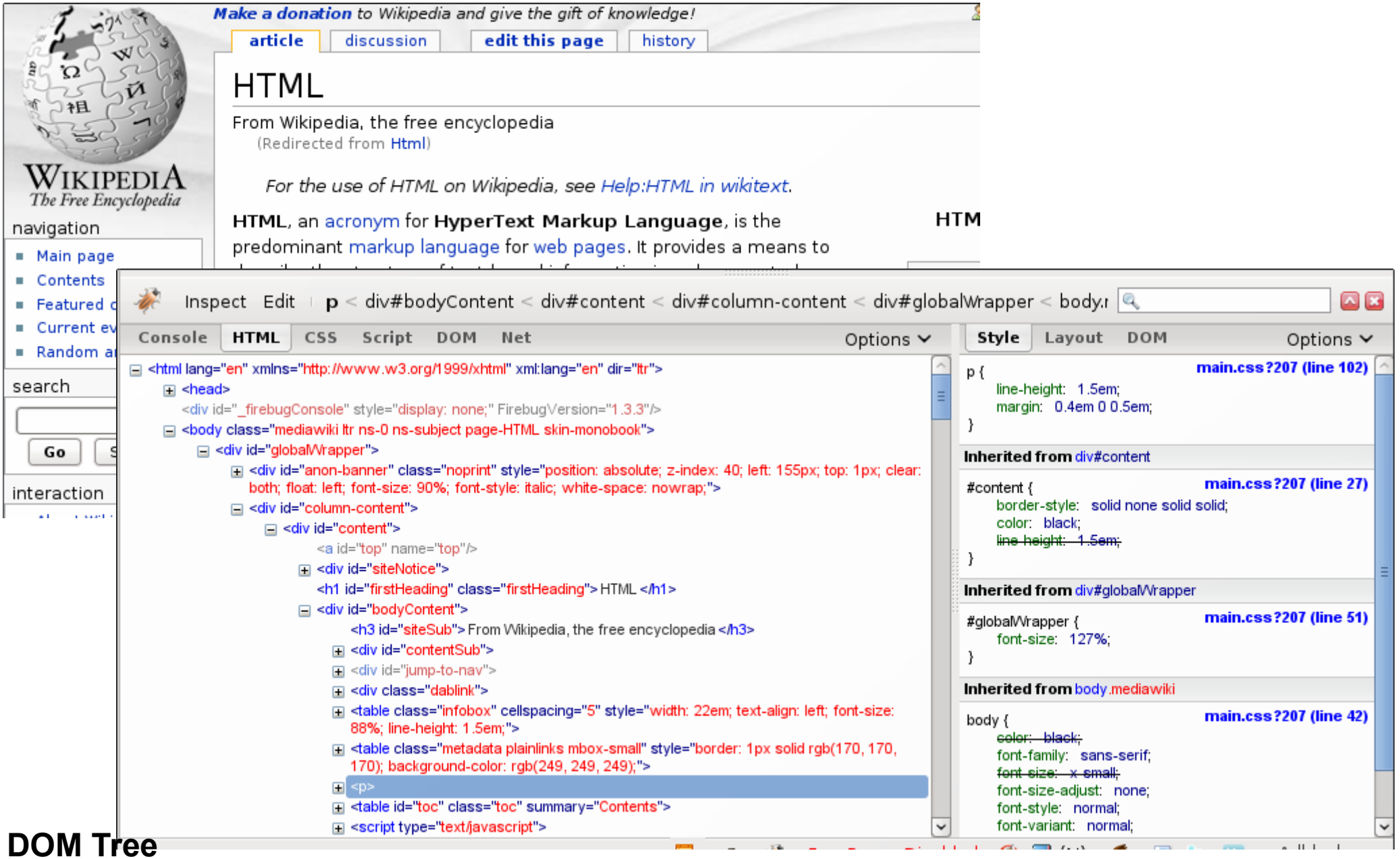
# HTML

- **initially**
  - simplified SGML syntax
  - specific set of tags (similar to troff)
  - mostly physical markup
  - some hypertext facilities
- **now**
  - styles, style sheets, semantic markup
  - XML-based variant

# HTML

```
<!DOCTYPE html>
<html>
  <head>
    <title>Hello HTML</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```

# HTML DOM




**DOM Tree**

# HTML + CSS

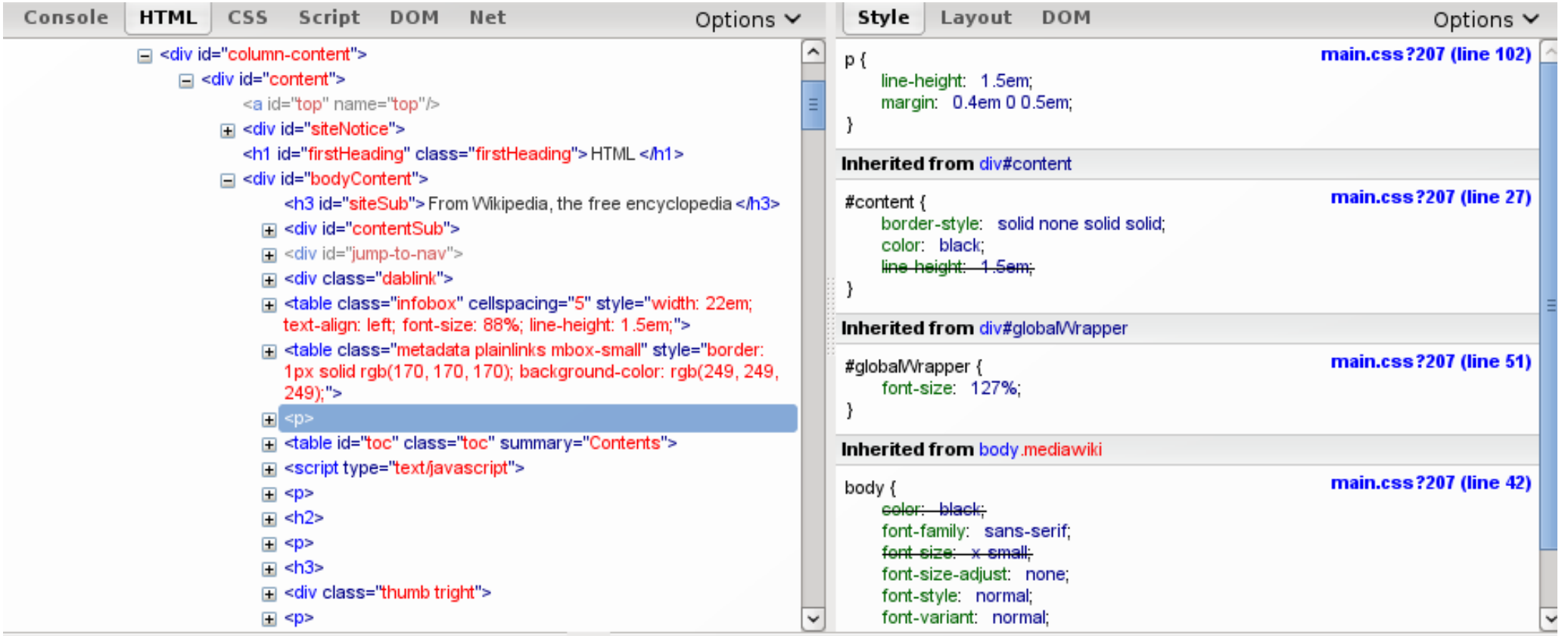
<h1>This is a Section</h1>

<h2>This is a Subsection</h2>

Hello, world.

```
h1 { color: white; background: orange !important; }
h2 { color: white; background: green !important; }
```

# Cascading Style Sheet



# HTML + CSS

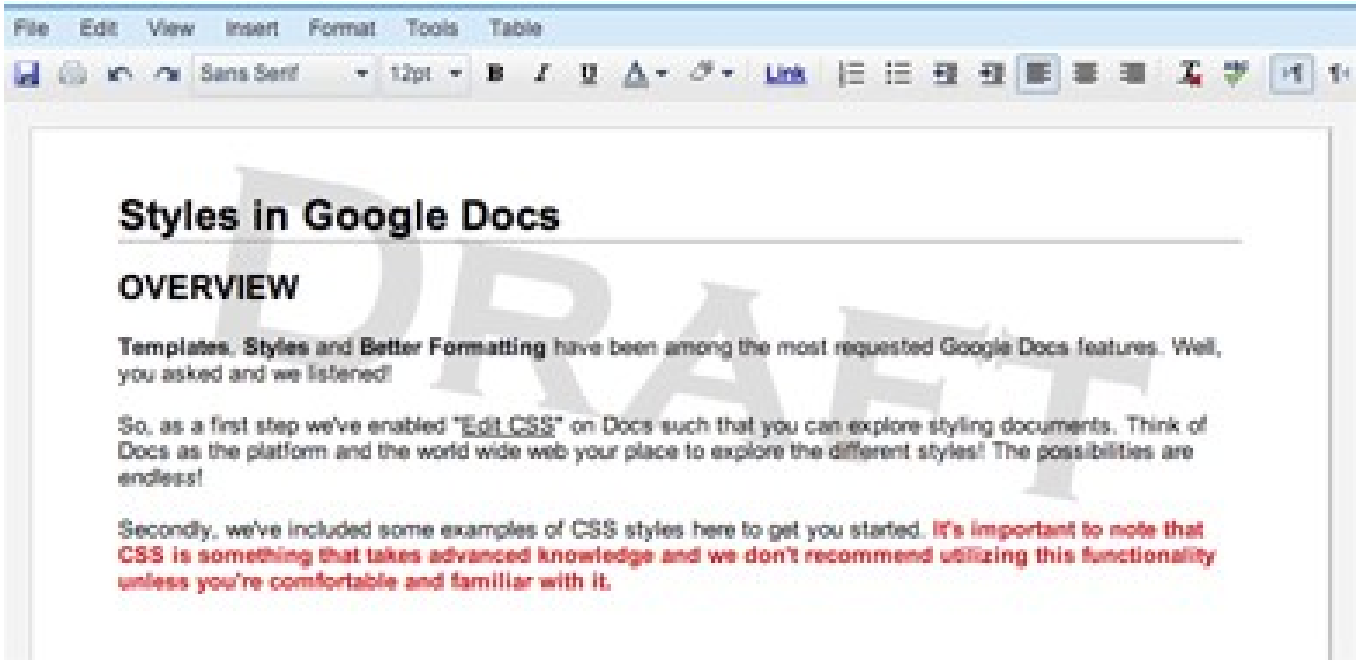
- **HTML/styles associated by patterns**
  - tag names, DOM path, style attribute
- **styles are “cascading”**
  - many sources of style information (document, browser, server, ...)
  - styles are combined by overriding lower priority styles with higher priority ones
  - priorities are determined by order and explicit declaration
- **CSS syntax itself is not XML**

# CSS + word processing

- **Google Docs**
  - on-line word processor (use with Google account)
  - formatting determined by (optional) CSS style sheets
  - Google Docs > Edit > Edit CSS...

# Google Docs + CSS

```
body {
 background-image: url('File?id=ad8wdwbvms_890m8v5pjdm_b');
 background-repeat: no-repeat;
 background-position: 50% 20px;
}
```



# HTML + CSS

- only addresses style sheets for rendering
- what about semantic information?

wiki markup

# problems with HTML

- **non-WYSIWYG editing**
  - hard to type
  - tricky syntax, hard to learn
  - even simple text requires lots of “noise”
- **WYSIWYG editing**
  - didn't used to be supported in browsers
  - hard to add semantic information

# wiki markup etc.

```
=== Mouth ===

Cats have highly specialized [[tooth|teeth]] for the killing of prey and the tearing of meat. The [[premolar]] and [[Molar (tooth)|first molar]] together compose the [[carnassial]] pair on each side of the mouth, which efficiently functions to shear meat like a pair of [[scissors]]. While this is present in [[Canidae|canids]], it is highly developed in felines. The cat's [[tongue]] has sharp spines, or [[Filiform papilla|papillae]], useful for retaining and ripping flesh from a carcass. These papillae are small backward-facing hooks that contain [[keratin]] which also assist in their [[Personal grooming|groom]]ing.

As facilitated by their oral structure, cats use a variety of vocalizations for [[cat communication|communication]], including meowing, purring, hissing, growling, squeaking, chirping, clicking, and grunting.<ref name=Channel3000Meows/>  Their types of [[Cat body language|body language]]: position of ears and tail, relaxation of whole body, kneading of paws, all are indicators of mood.
```

# wiki markup

- **input**
  - plain simple text comes out OK
  - section headings are simulated (“===”)
  - a few special characters are used for links
- **output**
  - output is reflowable HTML with different fonts, styles
  - line breaks in the input are not heeded (unless marked)

# reStructuredText

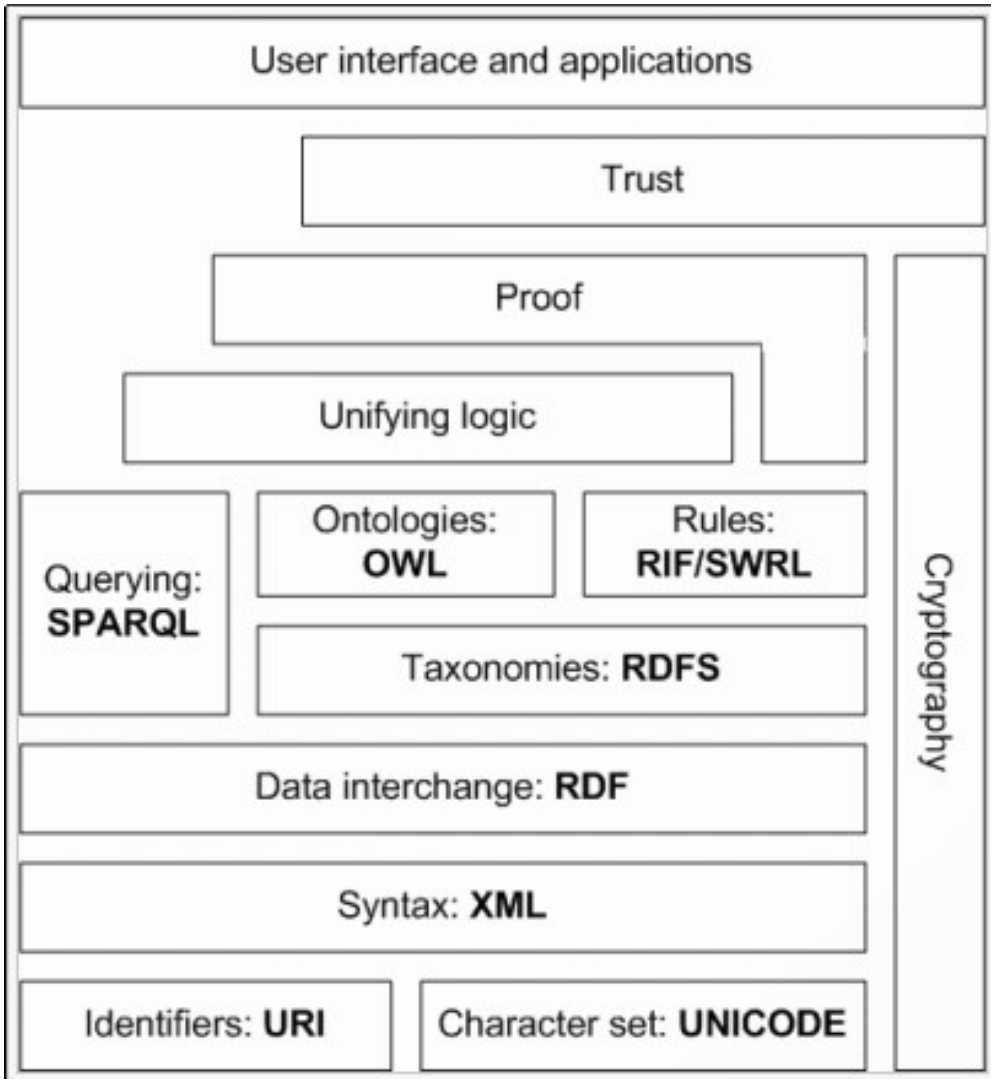
reStructuredText is an easy-to-read, what-you-see-is-what-you-get plaintext markup syntax and parser system. It is useful for in-line program documentation (such as Python docstrings), for quickly creating simple web pages, and for standalone documents.

# reStructuredText

```
This is a paragraph.  It's quite
short.

   This paragraph will result in an indented block of
   text, typically used for quoting other text.

*This* is **another** one.
```

# Microformats

# semantic web




- **XML, XML Schema, RDF, RDF Schema, OWL, SparQL, RIF, ...**
- **prescription (?)**
  - change to XML
  - express your information according to some schema
  - ...
  - profit ???

# adding semantic information

- **use case:**
  - publish business vCard, visitor incorporates with one click
  - publish calendar/event information, add to calendar from web
  - publish bibliography, incorporate from browser
- **generally**
  - add semantic information to web pages
  - allow users to pick up and use information easily

# solution: microformats

- **microformats**
  - are valid HTML that renders correctly in common browsers
  - incorporate semantic information
  - are equivalent to XML formats

# common microformats

- XHTML Friends Network (XFN)
  - social relationships
- hCard
  - represents people, companies, organizations, places
- hCalendar
  - iCalendar embedding for events, appointments
- hReview
  - reviews of products, services, etc.
- hAtom
  - embed Atom feeds directly in your page
- hResume
  - resume semantics (LinkedIn uses it)

# microformats services

- **Technorati**
- **Flickr (geo microformat)**
- **Yahoo (reviews, search, etc.; in and out)**
- **LinkedIn**
- **Technorati**
- **Magnolia**

# microformat ingest

- **directly by user**
  - Operator, Tails extensions for Firefox
  - recognize microformats in-line and add to address book
- **indirectly**
  - search engines, web crawlers recognize microformats
  - semantic information is used to enhance services, search

# microformat styling

```
<div class="vcard">
    <img src="img/thedude.jpg" ... class= "photo" />
    <h2>
        <a class= "fn n url" href="http://www.....com">
            <span class="given-name">Jeffrey</span>
            <span class="family-name">Lebowski</span>
        </a>
    </h2>
    <h3>address</h3>
    <div class="adr">
        <div class="street-address">123 Palm Drive</div>
        <span class="locality">Los Angeles</span>,
        <span class="region">CA</span>,
        <span class="postal-code">123456</span>
    </div>

    <h3>phone</h3>

    <div class="tel">+1 (123) 456-7899</div>

    <h3>email</h3>

    <a class="email" href="mailto...">...</a>
</div>
```

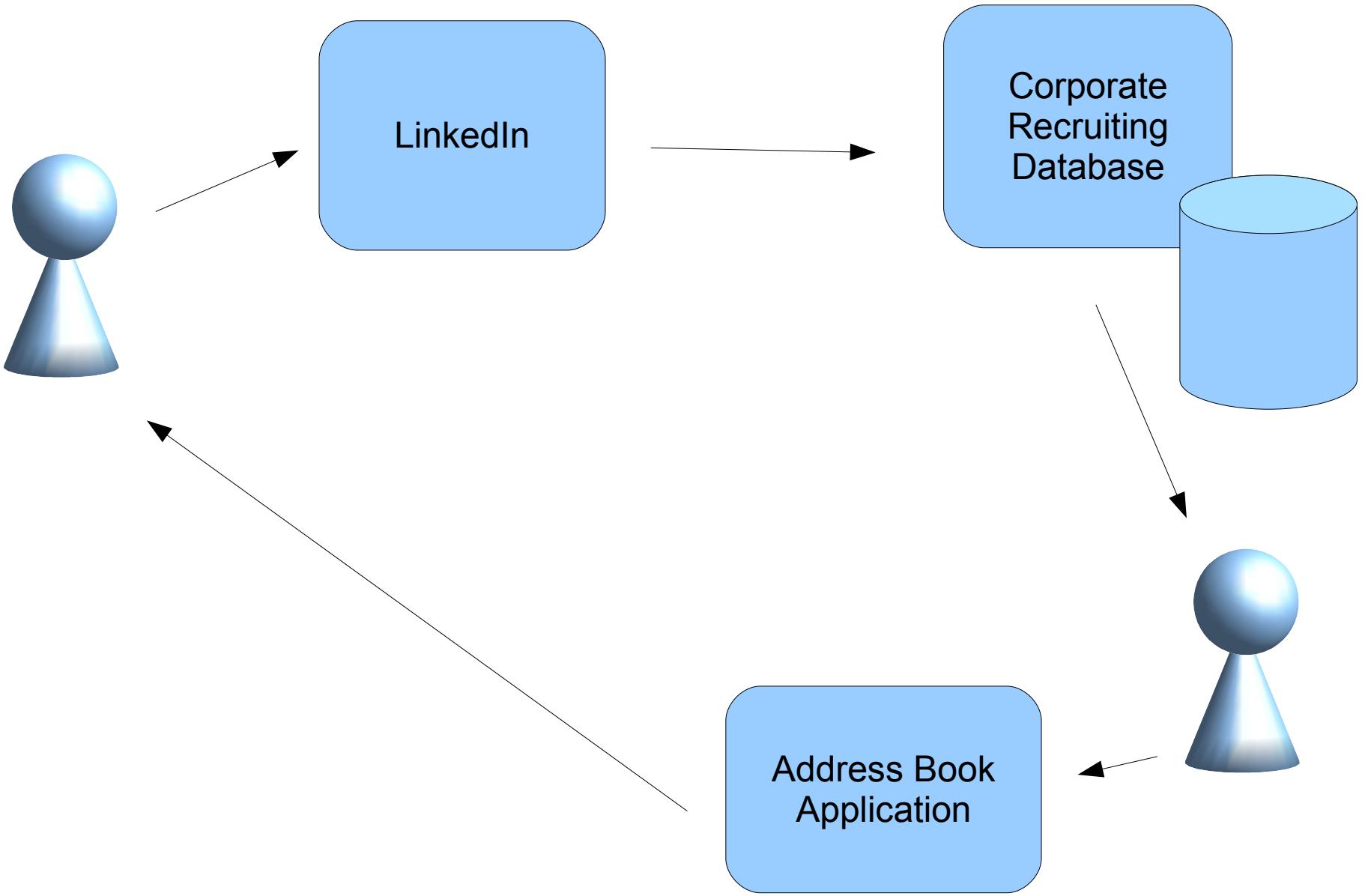
# microformat styling

```
div.vcard {
 width: 26em;
 margin: 0 auto;
 padding: 2em 2em 3em 2em;
 line-height: 1.5em;
 border-top: 1px solid #fff;
 background: url(img/bg.gif) no-repeat
bottom right;
}
```

courtesy Dan Cederholm

# microformat styling

# semantic data via microformats



# microformats

- use HTML markup for semantics
- fully compliant HTML, renders correctly
- parallels historical development
- in common use already

layout

# plain text reformatting

- **input**
  - some text with long lines, line breaks, paragraphs
- **output**
  - nicely formatted text, filled paragraphs, page breaks
  - word breaking? hyphenation?
- **examples**
  - fmt — simple text formatter
  - par — handles quoting

# plain text reformatting

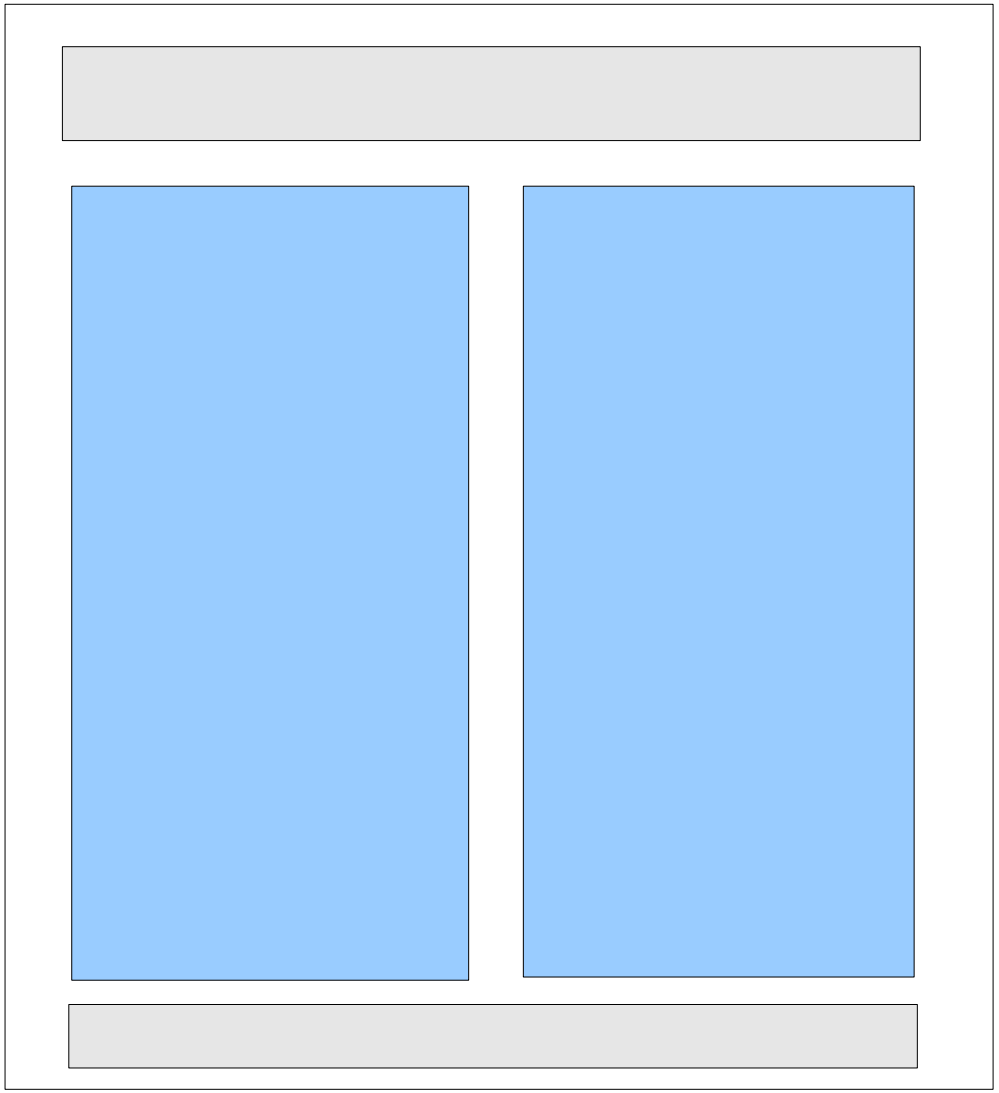
```
$ cat alice.txt
Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and
of having nothing to do: once or twice she had peeped into the book her sister w
as reading, but it had no pictures or conversations in it, `and what is the use
of a book,' thought Alice `without pictures or conversation?'

So she was considering in her own mind (as well as she could, for the hot day ma
de her feel very sleepy and stupid), whether the pleasure of making a daisy-chai
n would be worth the trouble of getting up and picking the daisies, when suddenl
y a White Rabbit with pink eyes ran close by her.
```

```
$ par 40f < alice.txt
Down the Rabbit-Hole

Alice was beginning to get very tired of
sitting by her sister on the bank, and
of having nothing to do: once or twice
she had peeped into the book her sister
was reading, but it had no pictures or
conversations in it, `and what is the
use of a book,' thought Alice `without
pictures or conversation?'

So she was considering in her own
mind (as well as she could, for the
hot day made her feel very sleepy
and stupid), whether the pleasure
of making a daisy-chain would be
worth the trouble of getting up and
picking the daisies, when suddenly
a White Rabbit with pink eyes ran
close by her.
```

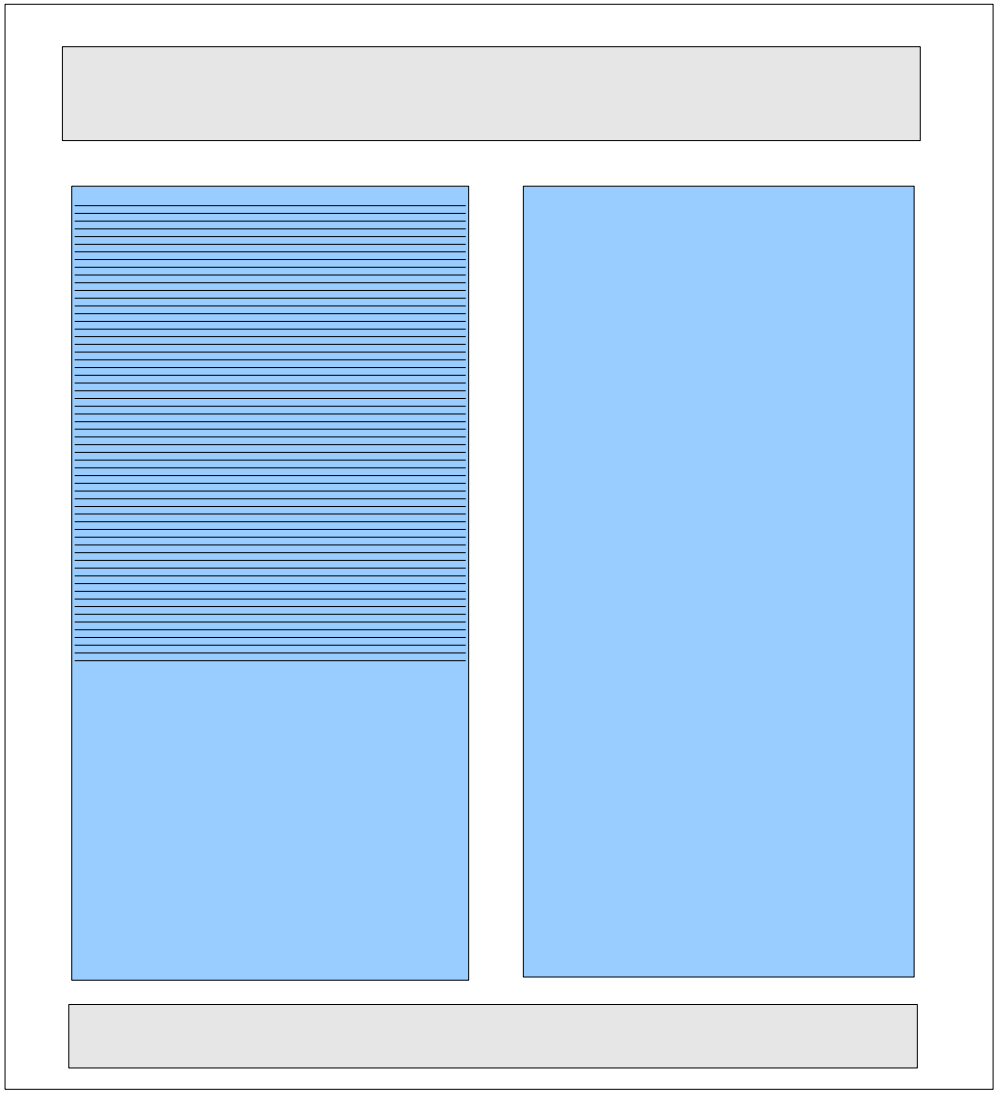
# plain text reformatting

- **questions**
  - algorithm?
  - hyphenation?
  - right justification?
  - what do you do with really long words?

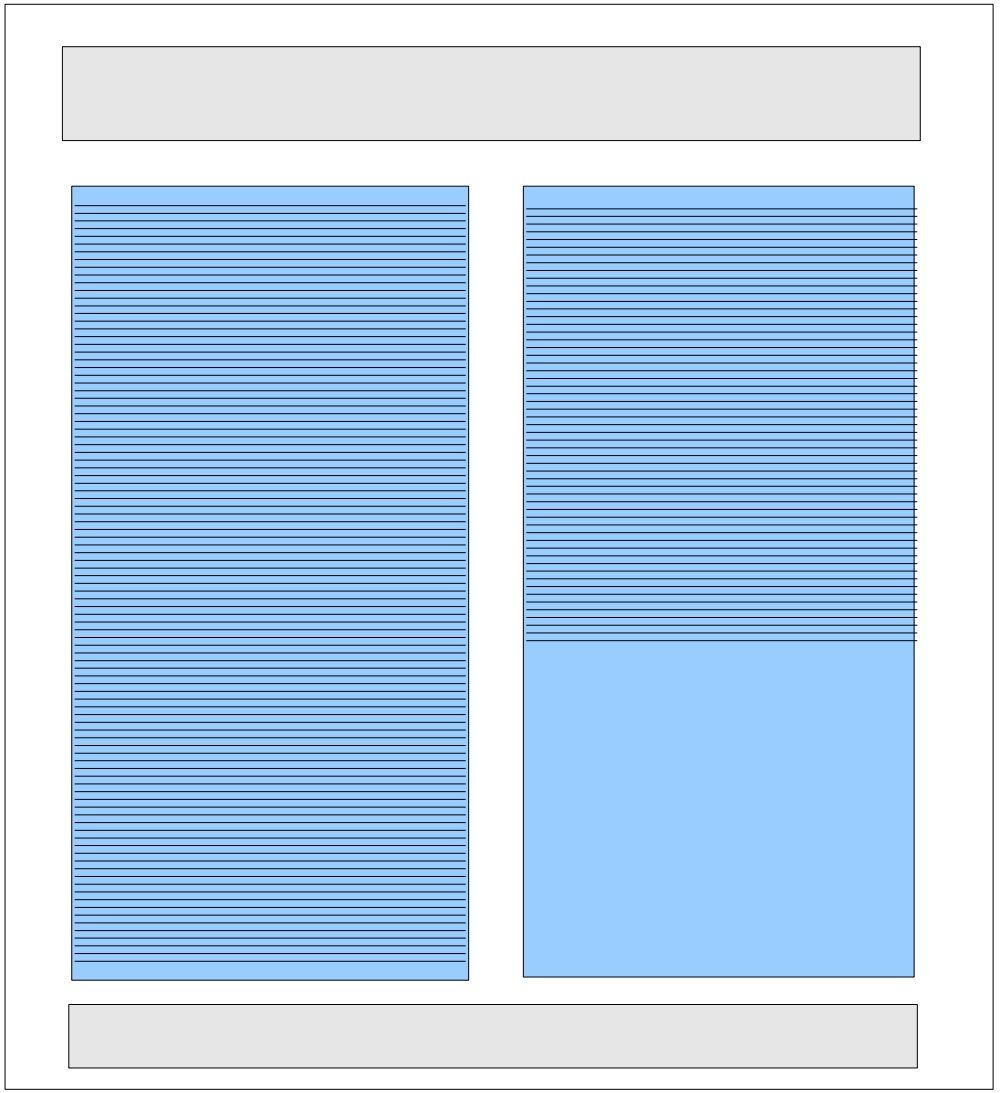
# typesetting



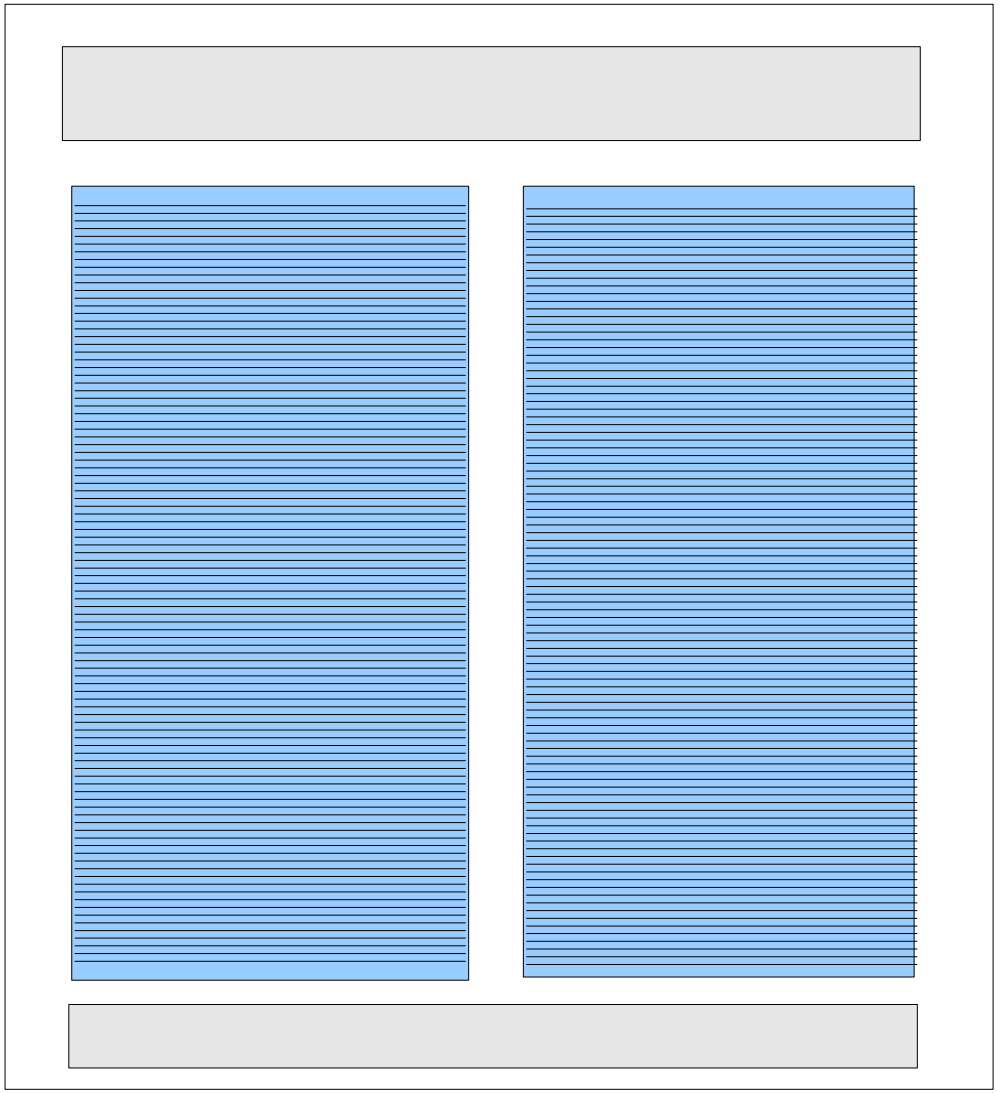
# typesetting



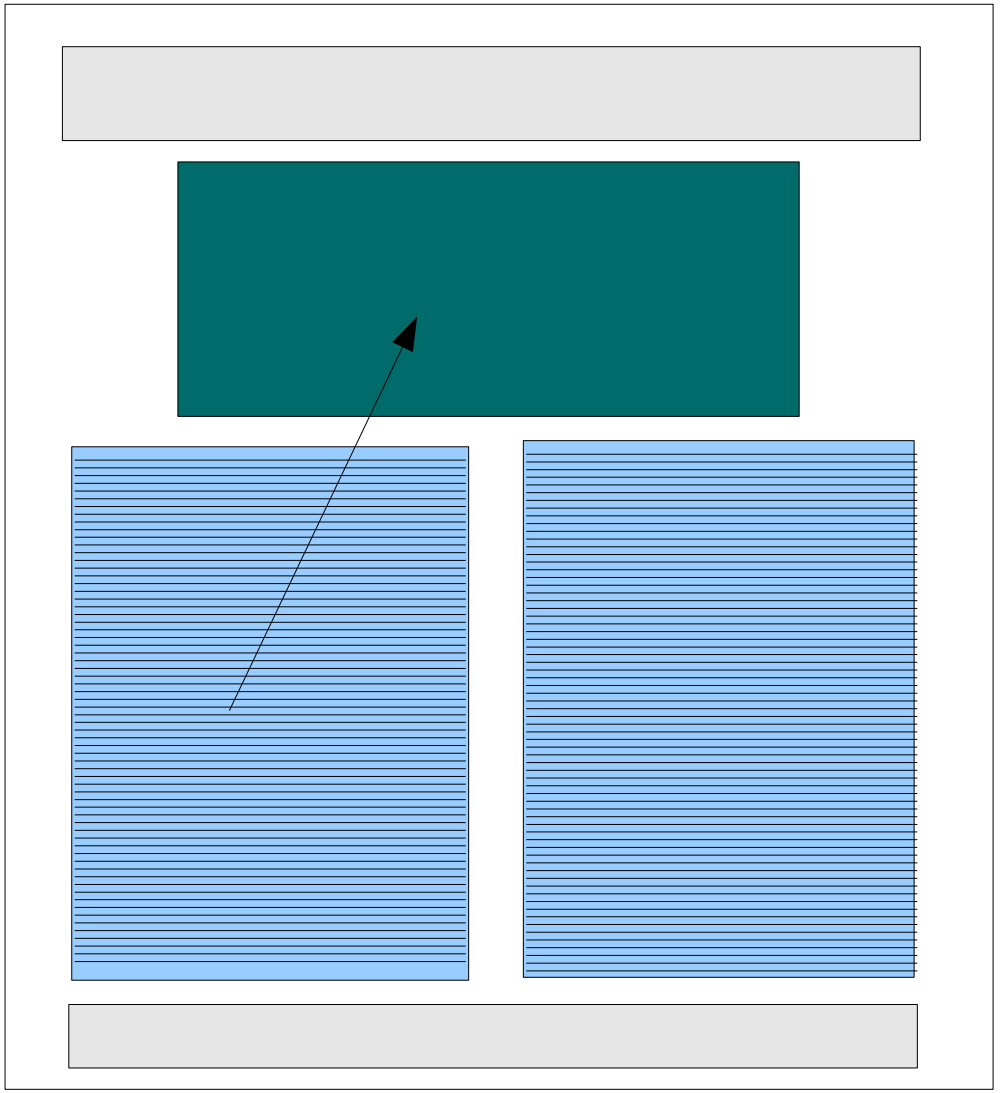
# typesetting



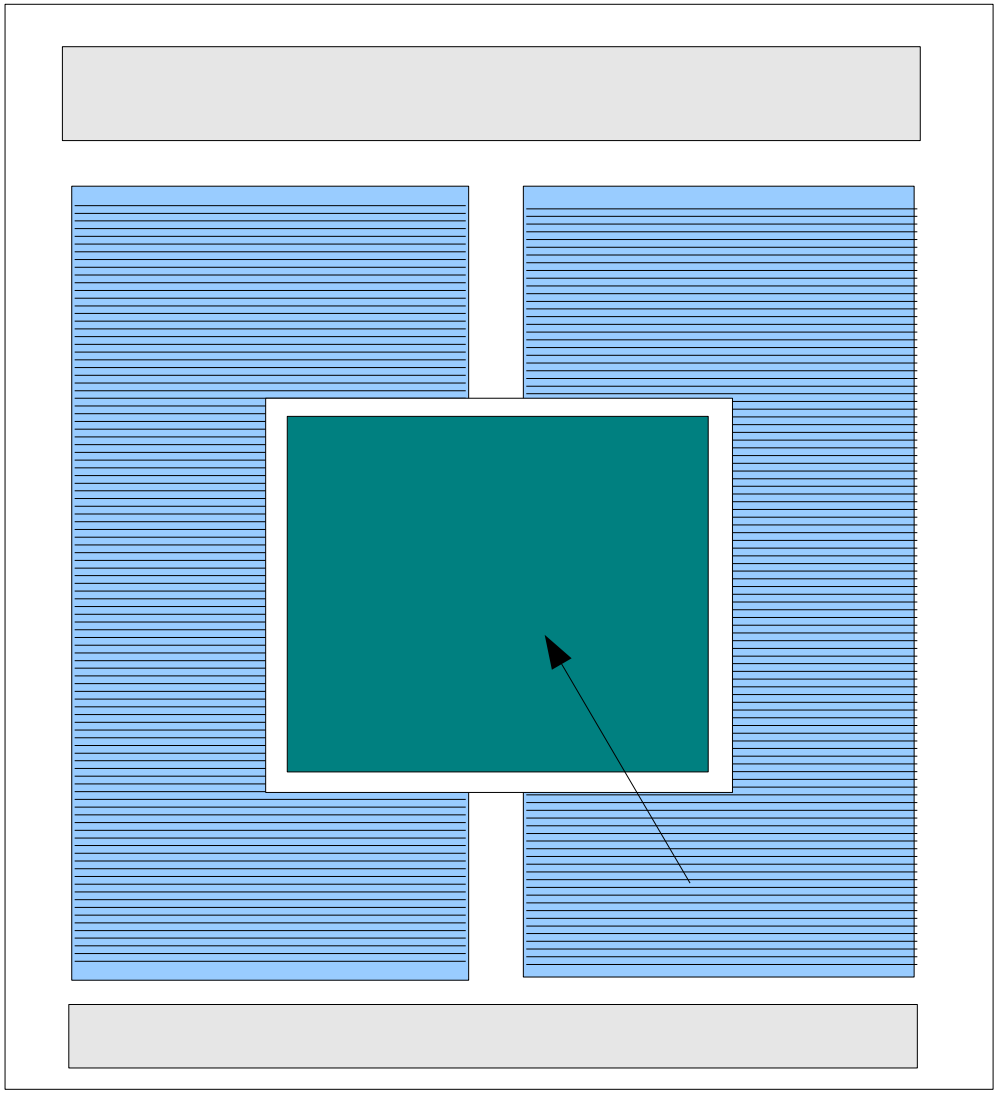
# typesetting



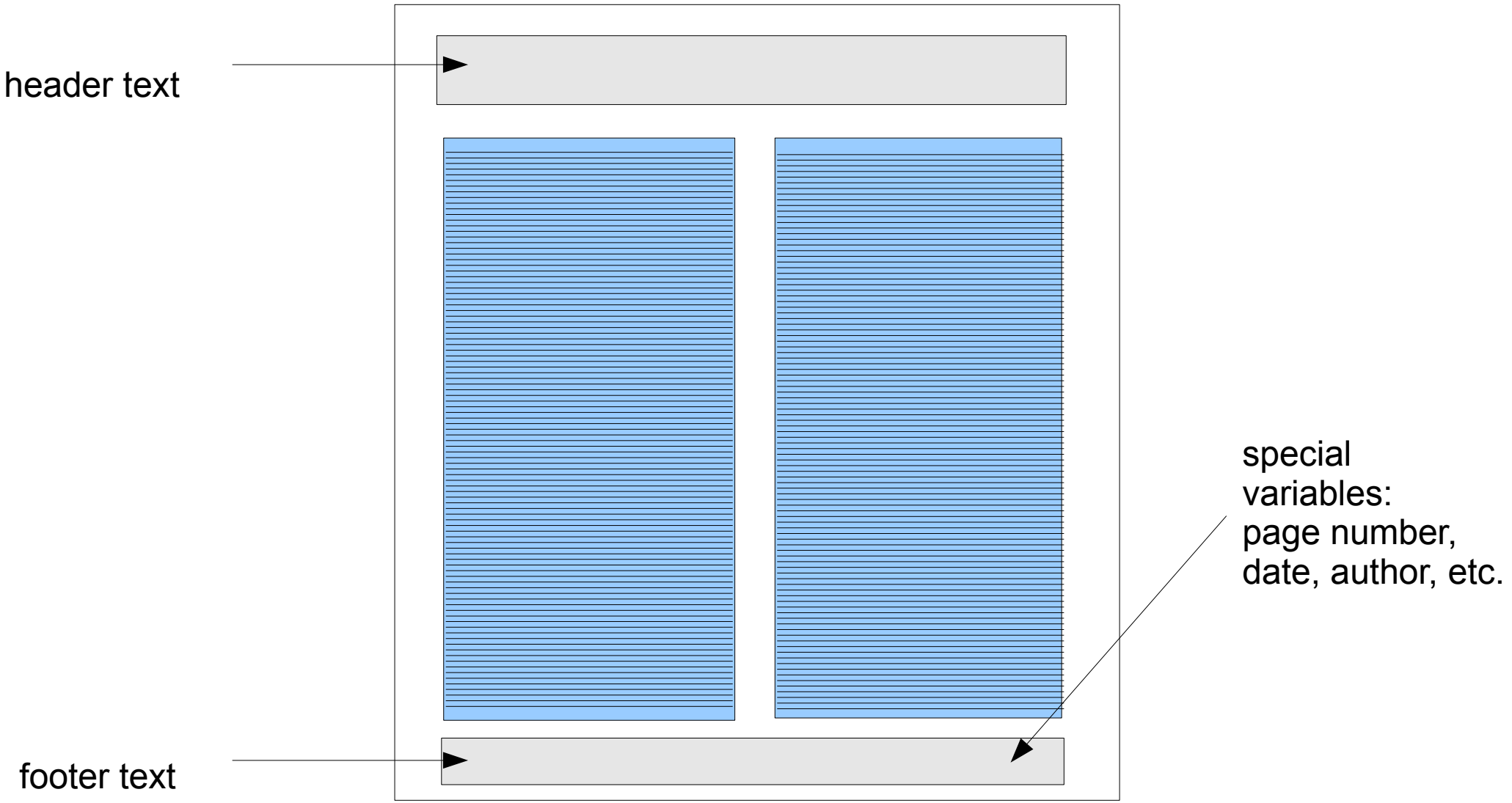
# floats



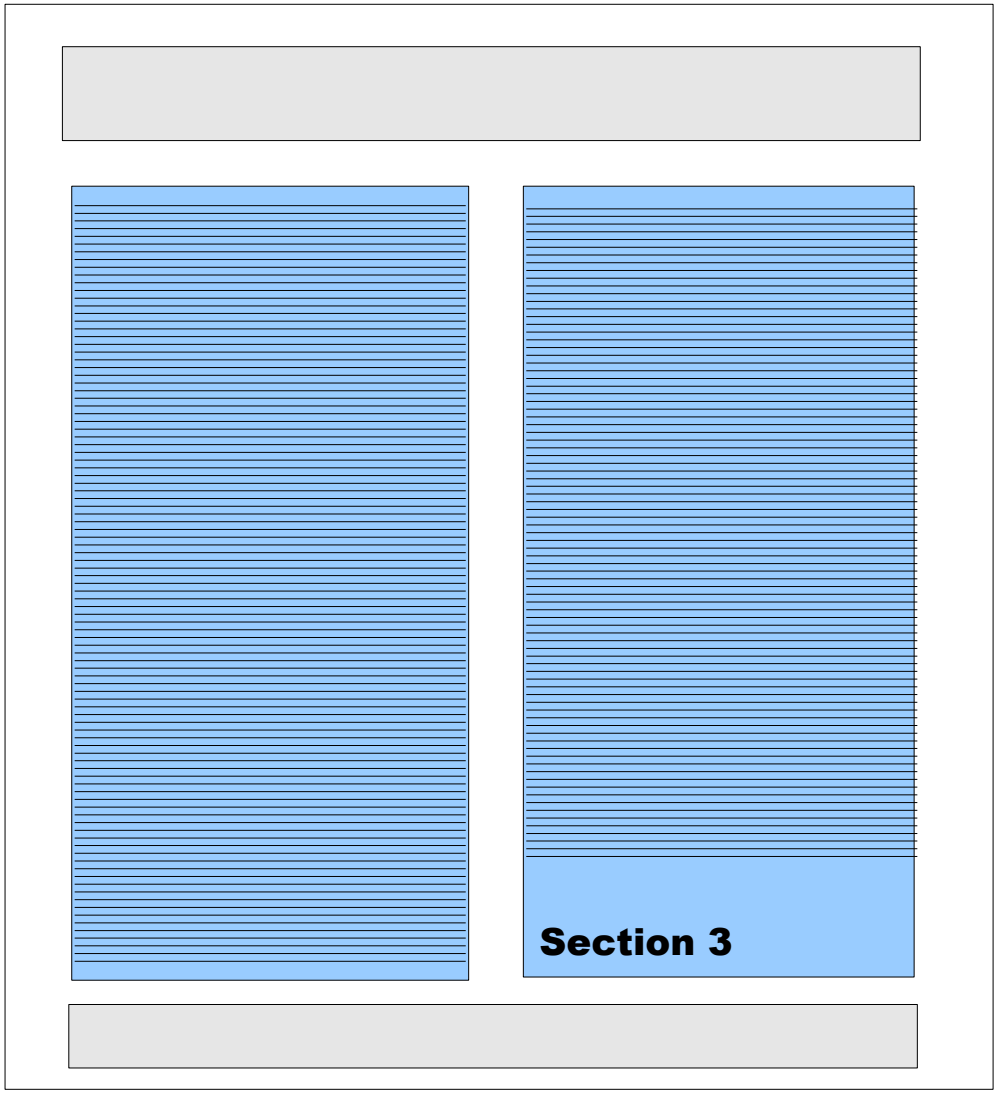
# floats



# headers / footers



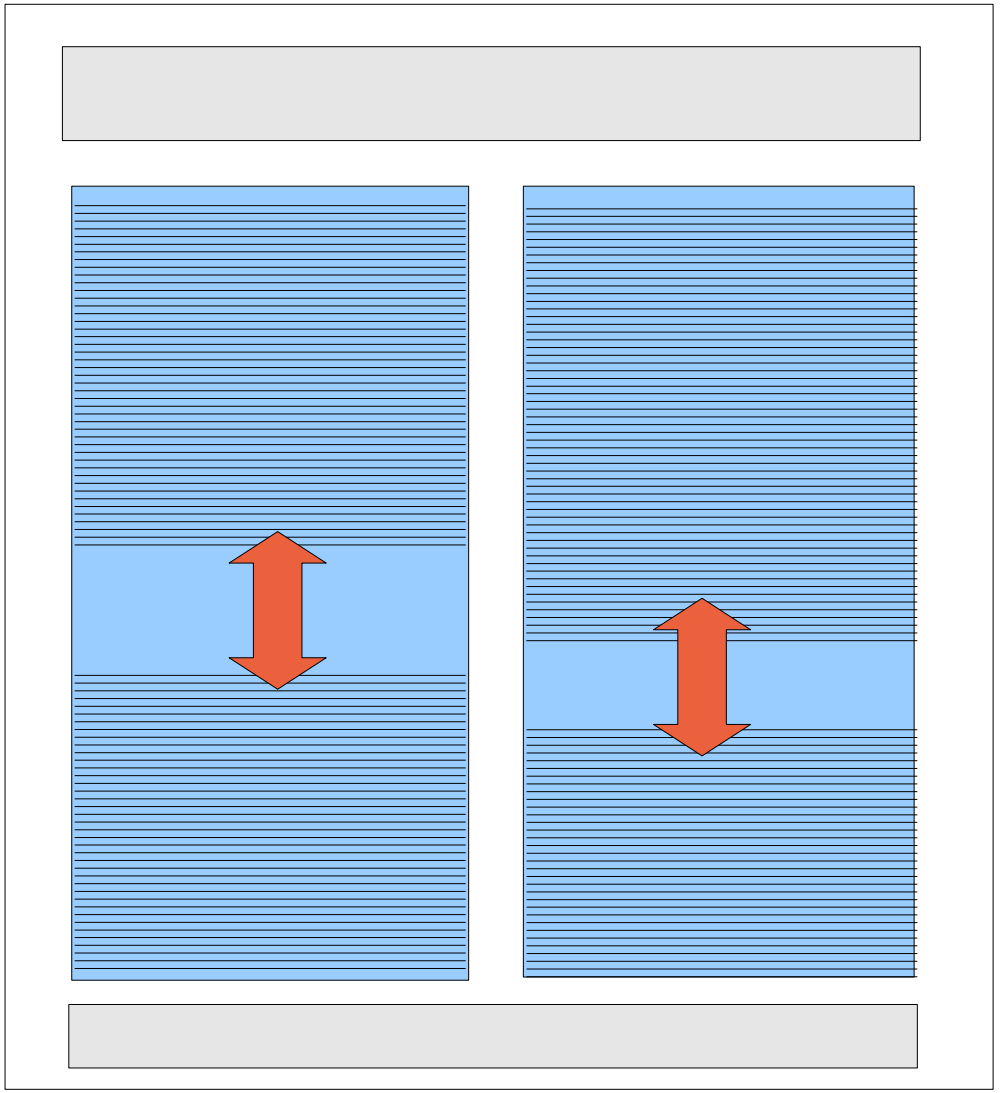
# widows, orphans, …



# glue

target size
minimum size
maximum size
cost of changing size

used by optimizer to
make better tradeoffs



# page layout

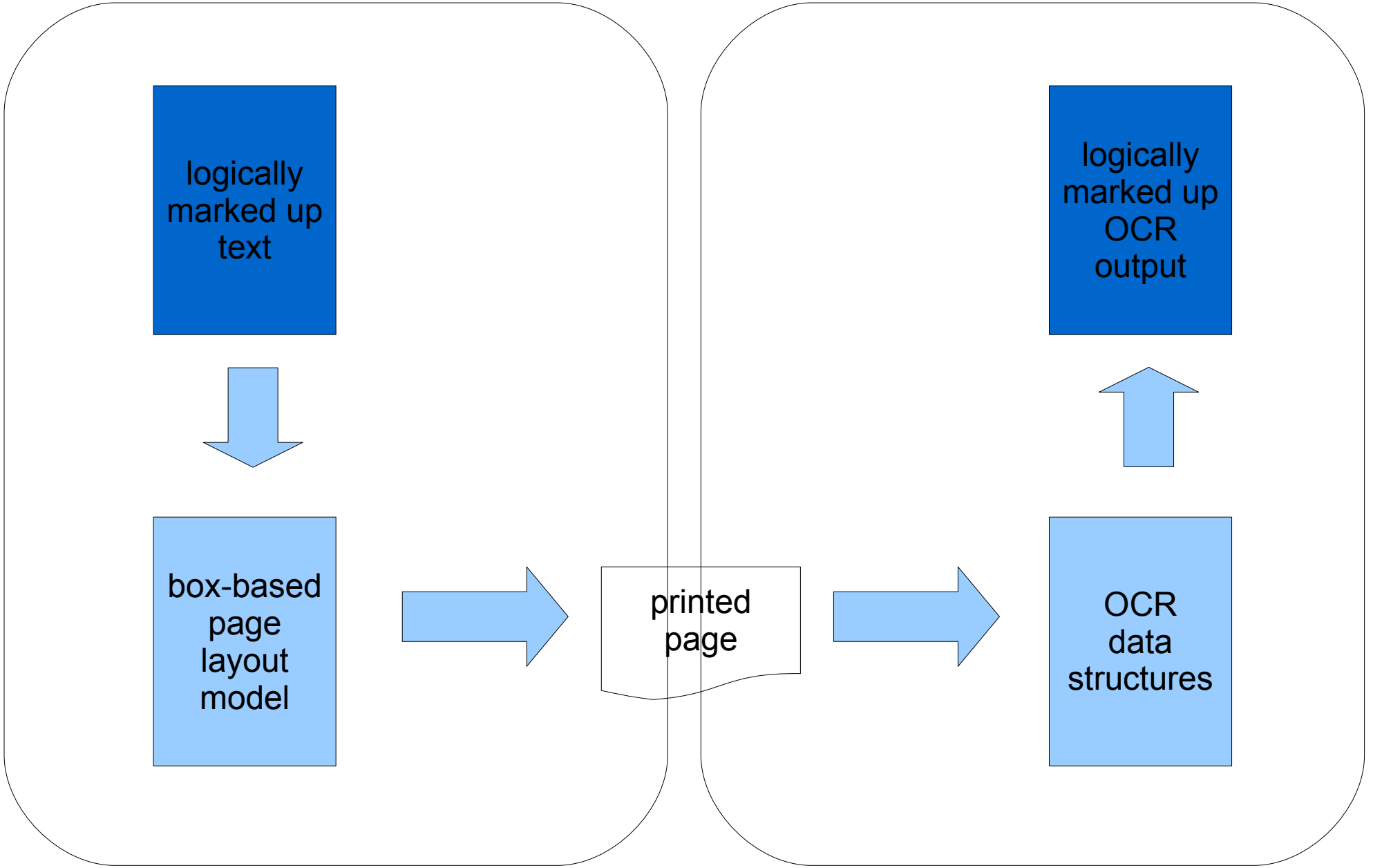
- **page layouts are defined in terms of boxes**
- **boxes get filled with source text**
- **floats get inserted as needed**
- **formatting is a global optimization problem**
  - floats affect available space
  - widows, orphans, line breaks, etc. may require reformatting
  - flexible glue allows optimizer to work

hOCR

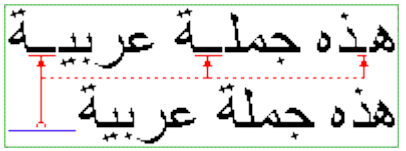
# formatting vs recognition

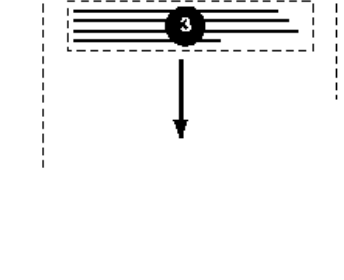
- **formatting language (TeX, etc.)**
  - describe how text is supposed to be laid out
  - instructions and hints to a formatting program
- **OCR description languages**
  - describe the visual appearance of laid out text
  - used for interpreting layouts

# communications & type setting model



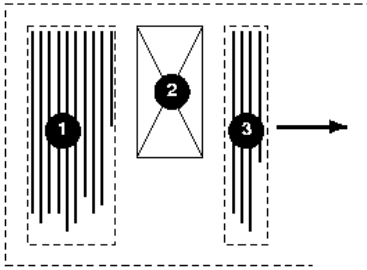
# HTML / CSS3 markup

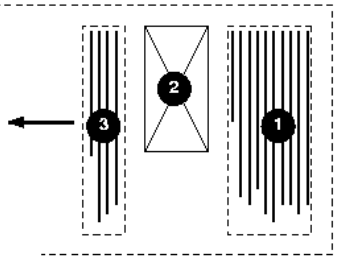



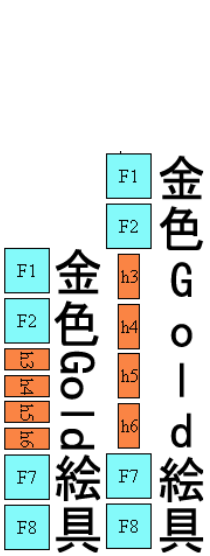



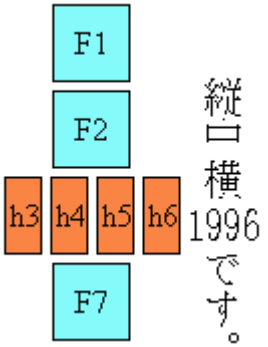


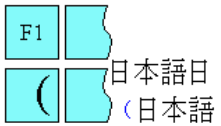


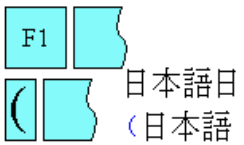


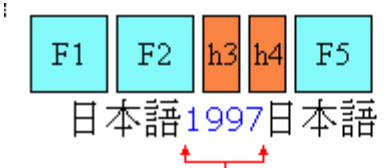



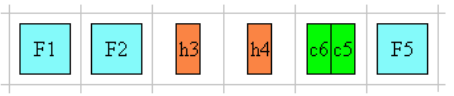




Mixed glyph layout in fixed grid

For example:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| こ | れ | は | 日 | 本 | 語 | の | 文 | 章 |
| で | す | 。 | T | h | i | s | | |
| i | s | | a | n | | E | n | - |

Fixed grid applied to mixed text

| | | | |
|---|---|---|---|
| こ | 本 | 語 | れ |
| れ | 語 | の | は |
| は | の | 文 | 日 |
| 章 | 文 | 。 | This is an English sentence. |

# hOCR Example

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN">

<html>
<head>
  <meta name="generator" content="HTML Tidy for Mac OS X (vers 1st December 2004), see www.w3.org">
  <meta name='ocr-id' content='OCRopus Revision: 312'>
  <meta name='ocr-recognized' content='lines text'>
  <meta name='DC.creator' content='Lewis Carroll'>
  <meta name='DC.title' content='Alice in Wonderland'>
  <meta name='DC.publisher' content='Macmillan'>

  <title>Alice in Wonderland</title>
</head>

<body>
  <div class='ocr_page' title='file ../data-ocr-test/alice_1.png'>
    <h3><span class='ocr_line' title='bbox 467 525 1386 588'>1 Down the Rabbit-Hole</span></h3>

    <p class='ocr_par'>
    <span class='ocr_line' title='bbox 461 648 2077 707'>Alice was beginning to get very tired of sitt
    <span class='ocr_line' title='bbox 461 708 2079 766'>and of having nothing to do: once or twice sh
    <span class='ocr_line' title='bbox 460 770 2078 826'>sister was reading, but it had no pictures or
    <span class='ocr_line' title='bbox 459 829 1906 884'>the use of a book,' thought Alice 'without pi
    </p>

    <p class='ocr_par'>
    <span class='ocr_line' title='bbox 533 888 2077 946'>So she was considering in her own mind (as we
    <span class='ocr_line' title='bbox 459 947 2075 1007'>day made her feel very sleepy and stupid),
    <span class='ocr_line' title='bbox 459 1008 2075 1067'>daisy-chain would be worth the trouble of
    <span class='ocr_line' title='bbox 458 1069 1819 1127'>when suddenly a White Rabbit with pink eyes
    </p>
```

standards-compliant HTML

compatible with other embedded HTML formats

separation of presentation and OCR markup

embedded geometric information

full access to Unicode and CSS3

# hOCR processing is simple

```
import sys,os,string,re
from xml.dom.ext.reader import HtmlLib
from xml.xpath import Evaluate as xquery

def get_text(node):
    textnodes = xquery(".//text()",node)
    s = string.join([node.nodeValue for node in textnodes])
    return re.sub(r'\s+',' ',s)

if len(sys.argv)>1: stream = open(sys.argv[1])
else: stream = sys.stdin
doc = HtmlLib.Reader().fromString(stream.read())
lines = xquery("//*[@class='ocr_line']",doc.documentElement)

for line in lines:
    print get_text(line)
```

# hOCR levels of markup

- **logical structure**
  - documents, sections, paragraphs, ...
- **per-page layout**
  - columns, floats, images, ...
- **engine specific**
  - boxes, words, ...
- **style, script, language, writing direction**
  - CSS-standard fonts, languages, etc.
- **metadata**
  - indicate capabilities (absence from doc insuff.)

# additional hOCR information

- **geometric information**
  - bounding boxes or polygons
  - polygons compressed for character segmentation
- **font / style information**
  - standards CSS font/style information
  - recommended inline
- **segmentation / recognition alternatives**
  - per character in elements
  - larger units via <INS> / <DEL> revision tags

# overview of hOCR markup elements

**logical**

* ocr_document
  * ocr_linear
    * ocr_title
    * ocr_author
    * ocr_abstract
    * ocr_part [H1]
      * ocr_chapter [H1]
        * ocr_section [H2]
          * ocr_sub*section [H3,H4]
            * ocr_display
            * ocr_blockquote [BLOCKQUOTE]
            * ocr_par [P]

**engine-specific**

- ocrx_block
- ocrx_line
- ocrx_word

- x_font s
  - OCR-engine specific font n
- x_fsize n
  - OCR-engine specific font size
- x_boxes b1x0 b1y0 b1x1 b1y1 b2x0 b2y0 b2x1 b2y1 ...
  - OCR-engine specific boxes associated with each character
- x_confs c1 c2 c3 ...
  - OCR-engine specific character confidences
- x_wconf n
  - OCR-engine specific confidence for the entire contained substring

**page layout**

- ocr_page
  - ocr_carea ("ocr content area
    - ocr_line [SPAN]
  - (floats)
  - ocr_separator (any separator
  - ocr_noise (any noise element

- ocr_float
  - ocr_separator
  - ocr_textfloat
  - ocr_textimage
  - ocr_image
    - ocr_linedrawing – s SVG (even if it is a
    - ocr_photo – somet
  - ocr_header
  - ocr_footer
  - ocr_pageno
  - ocr_table

**special**

```
<SPAN class="alternatives">
<INS class="alt" title="nlp 0.3">hello</INS>
<DEL class="alt" title="nlp 1.1">hallo</DEL>
</SPAN>
```

```
<span class="ocr_cinfo" title="nlp 1.7 2.3 3.9 2.7; seg 9 11 7.8,-2 15 3">hello</span>
```