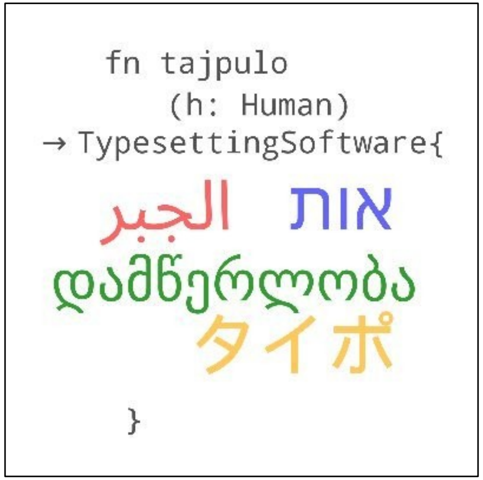
digitaler Schriftsatz

8th of November 2024

https://lukas-prokop.at/talks/2024-11-08_druckzeug

Who am I

- tajpulo
- DruckZeug member since 2014
- in Graz since 2010
- privately: https://lukas-prokop.at/
- typographically: https://who.typho.org/tajpulo
- I do digital typesetting (one day a week)



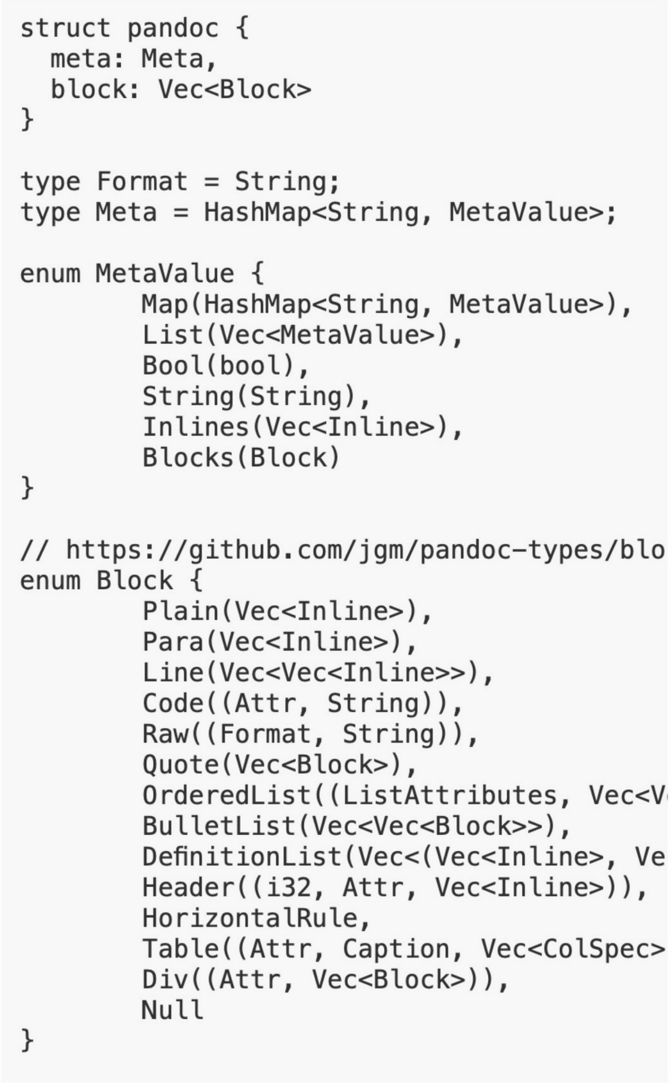
```
   fn tajpulo
        (h: Human)
→ TypesettingSoftware{
  الجبر   אות
 დამწერლობა
        タイポ

     }
```

digitaler Schriftsatz

Vortrag
von Lukas Prokop

Freitag, 8. November 2024
19–20:30 Uhr
Spektral Graz
Lendkai 45, 8020 Graz

Bitte um Anmeldung unter
office@druckzeug.at
Der Vortrag ist kostenlos
Sprache: Folien EN, Vortrag DE,
Fragen EN/DE



```
struct pandoc {
  meta: Meta,
  block: Vec<Block>
}

type Format = String;
type Meta = HashMap<String, MetaValue>;

enum MetaValue {
        Map(HashMap<String, MetaValue>),
        List(Vec<MetaValue>),
        Bool(bool),
        String(String),
        Inlines(Vec<Inline>),
        Blocks(Block)
}

// https://github.com/jgm/pandoc-types/blo
enum Block {
        Plain(Vec<Inline>),
        Para(Vec<Inline>),
        Line(Vec<Vec<Inline>>),
        Code((Attr, String)),
        Raw((Format, String)),
        Quote(Vec<Block>),
        OrderedList((ListAttributes, Vec<V
        BulletList(Vec<Vec<Block>>),
        DefinitionList(Vec<(Vec<Inline>, Ve
        Header((i32, Attr, Vec<Inline>)),
        HorizontalRule,
        Table((Attr, Caption, Vec<ColSpec>
        Div((Attr, Vec<Block>)),
        Null
}
```

Lukas stellt für DruckZeug-Mitglieder seinen Bereich „digitaler Textsatz“ vor. Er beschäftigt sich damit, wie digitale Dokumente im gesamten Prozess vom Eintippen bis zur Erstellung des Ausgabeformats wie PDF verarbeitet werden. Hintergrund ist, dass er die Freie Software Typho in diesem Bereich schreibt.
Er wird im Vortrag über Unicode, HTML, PDF und OpenType sprechen und technische Gegebenheiten allgemeinverständlich herunterbrechen. Der Vortrag ist auf Mitglieder des Vereins DruckZeug ausgerichtet, Freunde sind willkommen.

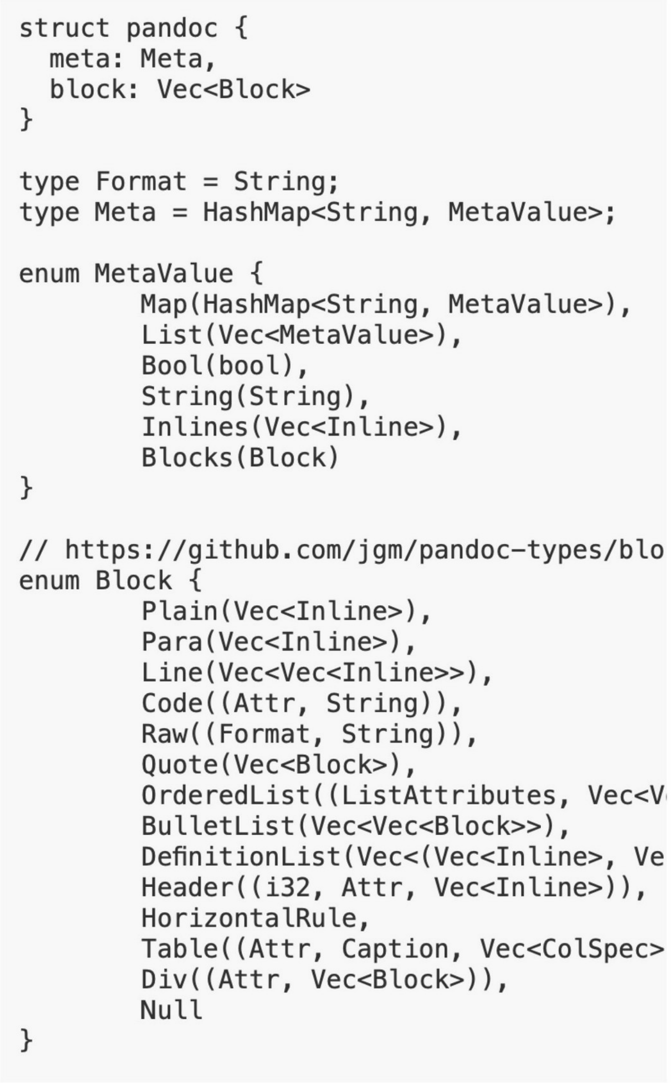
DruckZeug
BLEISATZ & BUCHDRUCK IN GRAZ

digitaler Schriftsatz

Vortrag
von Lukas Prokop

Freitag, 8. November 2024
19–20:30 Uhr
Spektral Graz
Lendkai 45, 8020 Graz

Bitte um Anmeldung unter
office@druckzeug.at
Der Vortrag ist kostenlos.
Sprache: Folien EN, Vortrag DE,
Fragen EN/DE



```
struct pandoc {
  meta: Meta,
  block: Vec<Block>
}

type Format = String;
type Meta = HashMap<String, MetaValue>;

enum MetaValue {
        Map(HashMap<String, MetaValue>),
        List(Vec<MetaValue>),
        Bool(bool),
        String(String),
        Inlines(Vec<Inline>),
        Blocks(Block)
}

// https://github.com/jgm/pandoc-types/blo
enum Block {
        Plain(Vec<Inline>),
        Para(Vec<Inline>),
        Line(Vec<Vec<Inline>>),
        Code((Attr, String)),
        Raw((Format, String)),
        Quote(Vec<Block>),
        OrderedList((ListAttributes, Vec<V
        BulletList(Vec<Vec<Block>>),
        DefinitionList(Vec<(Vec<Inline>, Ve
        Header((i32, Attr, Vec<Inline>)),
        HorizontalRule,
        Table((Attr, Caption, Vec<ColSpec>
        Div((Attr, Vec<Block>)),
        Null
}
```

Lukas stellt für DruckZeug-Mitglieder seinen Bereich „digitaler Textsatz“ vor. Er beschäftigt sich damit, wie digitale Dokumente im gesamten Prozess vom Eintippen bis zur Erstellung des Ausgabeformats wie PDF verarbeitet werden. Hintergrund ist, dass er die Freie Software Typho in diesem Bereich schreibt.
Er wird im Vortrag über Unicode, HTML, PDF und OpenType sprechen und technische Gegebenheiten allgemeinverständlich herunterbrechen. Der Vortrag ist auf Mitglieder des Vereins DruckZeug ausgerichtet, Freunde sind willkommen.

DruckZeug
BLEISATZ & BUCHDRUCK IN GRAZ

Chapter 1.
Understanding the domain

Definition

> “ Typesetting is the composition of **text** for publication, display, or distribution by means of arranging **physical type** (or sort) in mechanical systems or **glyphs** in digital systems representing characters (letters and other symbols).
>
> – *EN Wikipedia “Typesetting” (emphasis mine*)

Definition

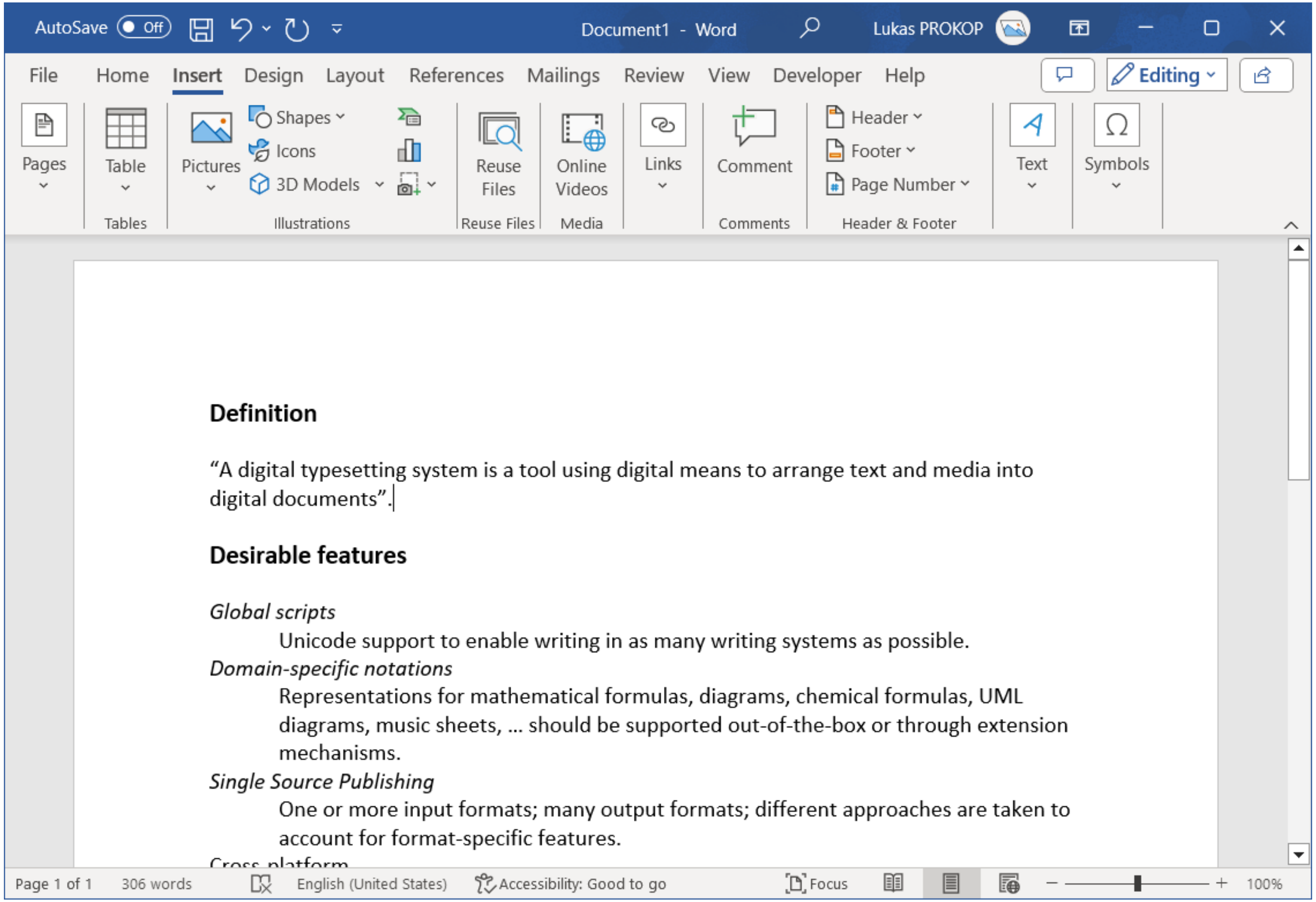
> A digital typesetting system is any system using **digital means** to **arrange text and media** into digital documents
>
> – *my definition (https://arewedigitaltypesettingyet.com/)*

Definition

> A digital typesetting system is any system using **digital means** to **arrange text and media** into digital documents.
>
> – *my definition (https://arewedigitaltypesettingyet.com/)*

> A **plaintext** digital typesetting system uses Unicode scalars to specify text and its arrangement in digital documents.
>
> – *my definition (https://arewedigitaltypesettingyet.com/)*

> A **FOSS** system satisfies the OpenSource criteria by OSI
>
> – *my definition (c.f. https://opensource.org/osd)*



Definition

"A digital typesetting system is a tool using digital means to arrange text and media into digital documents".

Desirable features

Global scripts

Unicode support to enable writing in as many writing systems as possible.

Domain-specific notations

Representations for mathematical formulas, diagrams, chemical formulas, UML diagrams, music sheets, … should be supported out-of-the-box or through extension mechanisms.

Single Source Publishing

One or more input formats; many output formats; different approaches are taken to account for format-specific features.

Cross-platform

digital typesetting system

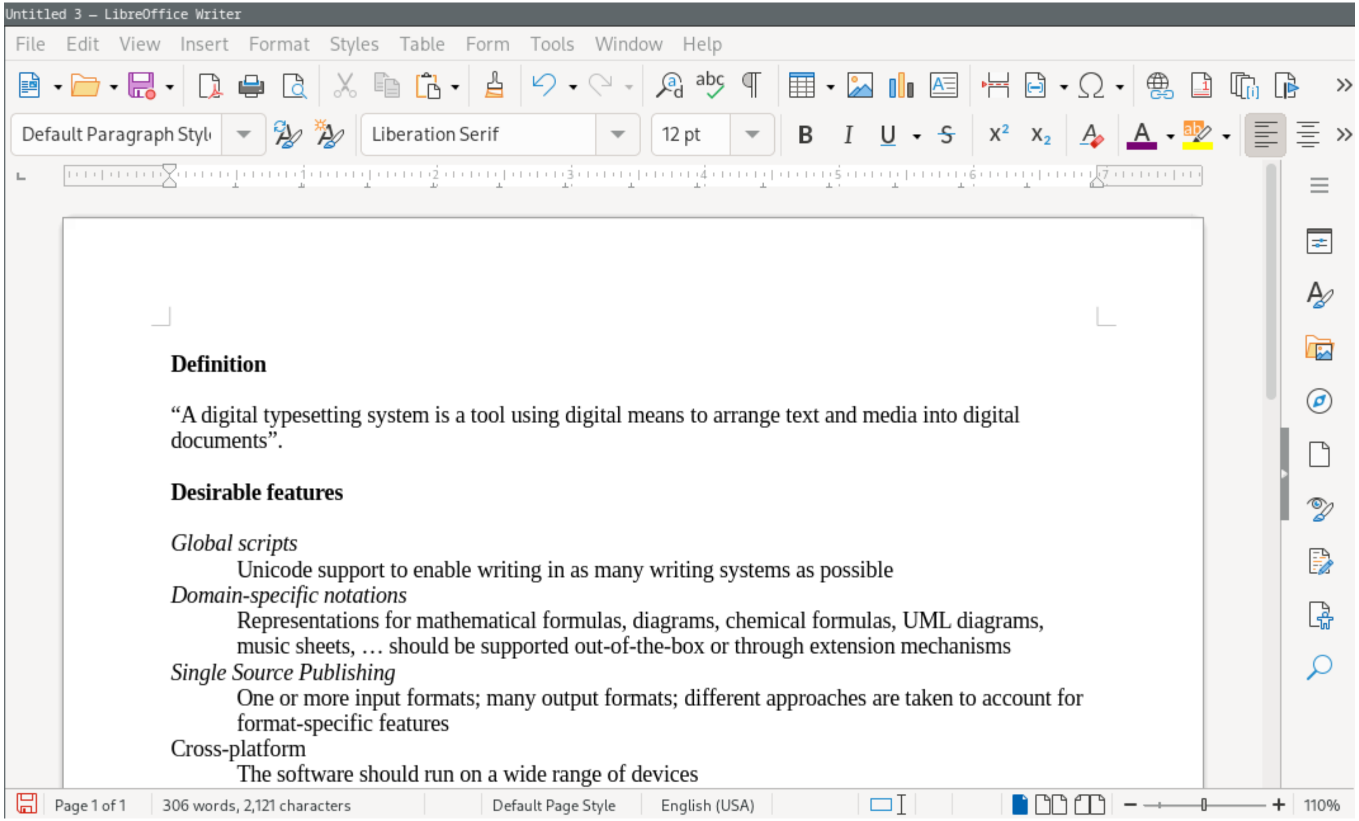
✓

plaintext

❌

FOSS

❌


Untitled 3 – LibreOffice Writer
File Edit View Insert Format Styles Table Form Tools Window Help
Default Paragraph Style
Liberation Serif
12 pt
Page 1 of 1
306 words, 2,121 characters
Default Page Style
English (USA)
110%

Definition

"A digital typesetting system is a tool using digital means to arrange text and media into digital documents".

Desirable features

Global scripts
  Unicode support to enable writing in as many writing systems as possible

Domain-specific notations
  Representations for mathematical formulas, diagrams, chemical formulas, UML diagrams, music sheets, … should be supported out-of-the-box or through extension mechanisms

Single Source Publishing
  One or more input formats; many output formats; different approaches are taken to account for format-specific features

Cross-platform
  The software should run on a wide range of devices

digital typesetting system

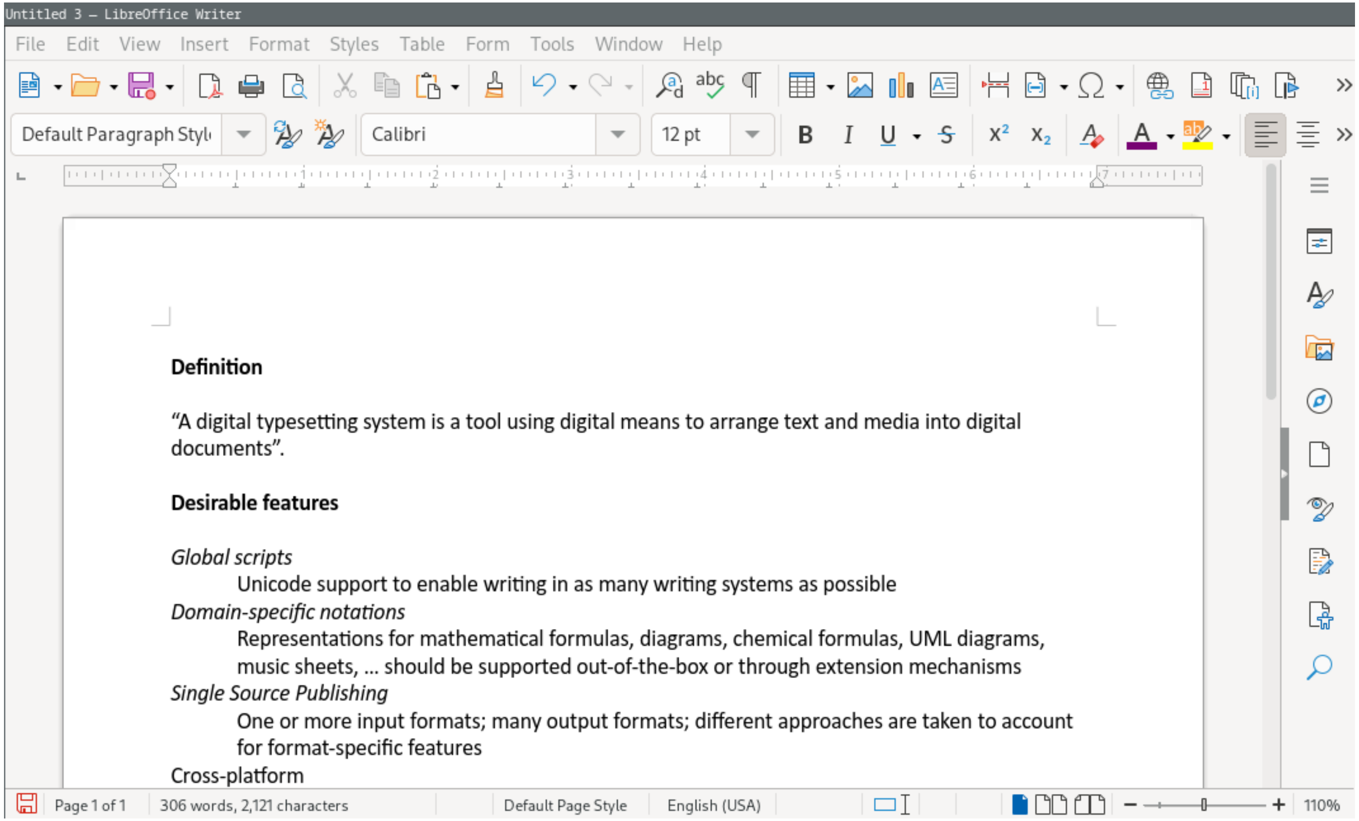
✔

plaintext

❌

FOSS

✔


Untitled 3 – LibreOffice Writer
File Edit View Insert Format Styles Table Form Tools Window Help
Default Paragraph Styl
Calibri
12 pt
Page 1 of 1
306 words, 2,121 characters
Default Page Style
English (USA)
110%

Definition

“A digital typesetting system is a tool using digital means to arrange text and media into digital documents”.

Desirable features

Global scripts

Unicode support to enable writing in as many writing systems as possible

Domain-specific notations

Representations for mathematical formulas, diagrams, chemical formulas, UML diagrams, music sheets, ... should be supported out-of-the-box or through extension mechanisms

Single Source Publishing

One or more input formats; many output formats; different approaches are taken to account for format-specific features

Cross-platform

digital typesetting system

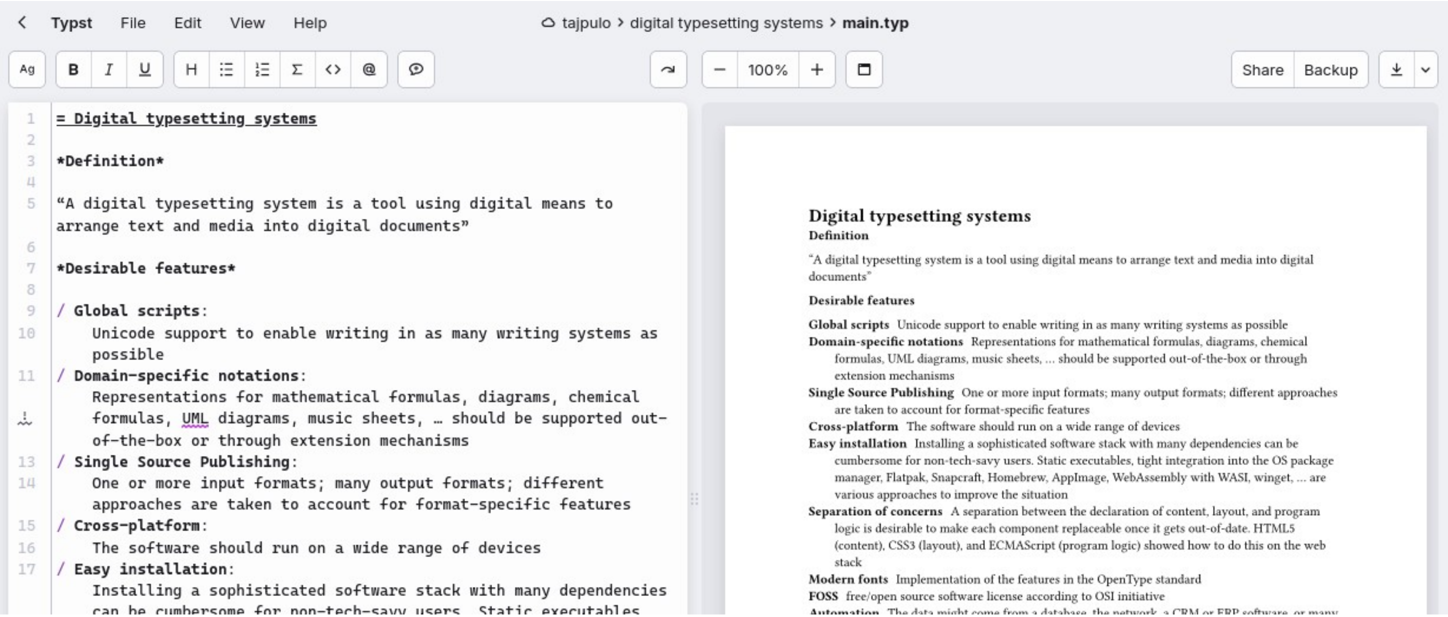
✓

plaintext

✗

FOSS

✓



```
= Digital typesetting systems

*Definition*

“A digital typesetting system is a tool using digital means to arrange text and media into digital documents”

*Desirable features*

/ Global scripts:
  Unicode support to enable writing in as many writing systems as possible
/ Domain-specific notations:
  Representations for mathematical formulas, diagrams, chemical formulas, UML diagrams, music sheets, … should be supported out-of-the-box or through extension mechanisms
/ Single Source Publishing:
  One or more input formats; many output formats; different approaches are taken to account for format-specific features
/ Cross-platform:
  The software should run on a wide range of devices
/ Easy installation:
  Installing a sophisticated software stack with many dependencies can be cumbersome for non-tech-savy users. Static executables
```

Digital typesetting systems

Definition

“A digital typesetting system is a tool using digital means to arrange text and media into digital documents”

Desirable features

Global scripts Unicode support to enable writing in as many writing systems as possible
Domain-specific notations Representations for mathematical formulas, diagrams, chemical formulas, UML diagrams, music sheets, … should be supported out-of-the-box or through extension mechanisms
Single Source Publishing One or more input formats; many output formats; different approaches are taken to account for format-specific features
Cross-platform The software should run on a wide range of devices
Easy installation Installing a sophisticated software stack with many dependencies can be cumbersome for non-tech-savy users. Static executables, tight integration into the OS package manager, Flatpak, Snapcraft, Homebrew, AppImage, WebAssembly with WASI, winget, … are various approaches to improve the situation
Separation of concerns A separation between the declaration of content, layout, and program logic is desirable to make each component replaceable once it gets out-of-date. HTML5 (content), CSS3 (layout), and ECMAScript (program logic) showed how to do this on the web stack
Modern fonts Implementation of the features in the OpenType standard
FOSS free/open source software license according to OSI initiative
Automation The data might come from a database, the network, a CRM or ERP software, or many

digital typesetting system

✓

plaintext

✓

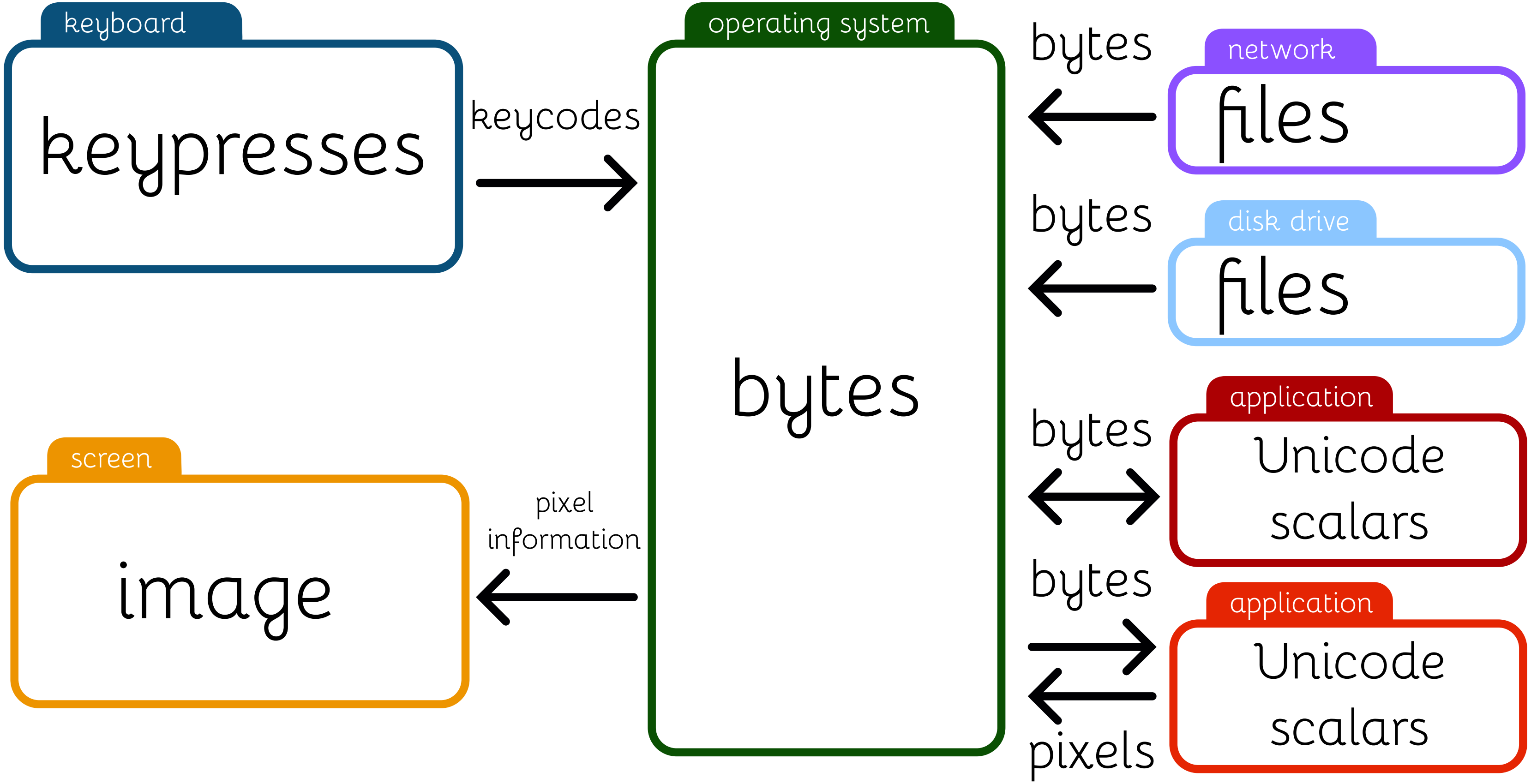
FOSS

✓

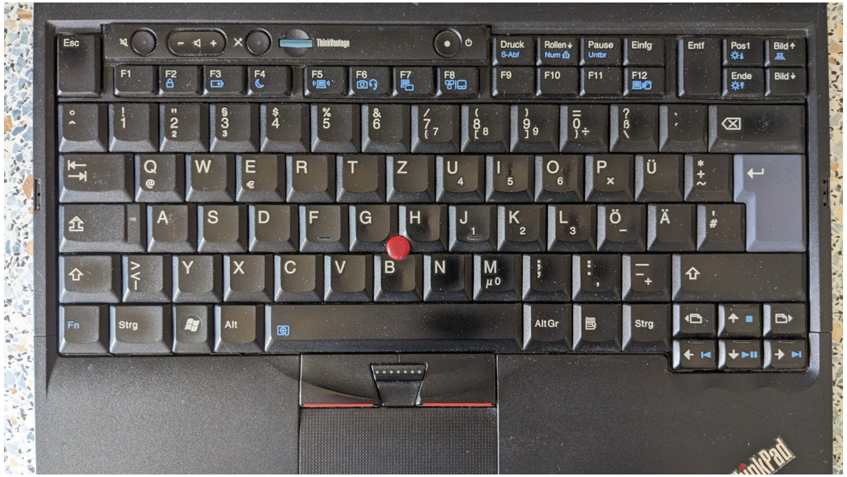
Non-FOSS products

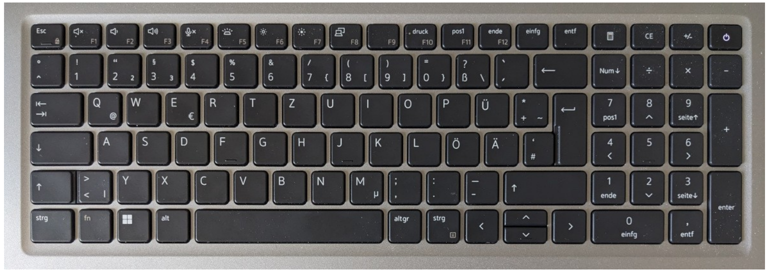
- Adobe Creative Cloud: Adobe Acrobat, Adobe DreamWeaver, Adobe Fonts, Adobe FrameMaker, Adobe Fresco, Adobe Illustrator, Adobe Illustrator Draw, Adobe InCopy, Adobe InDesign, Adobe XD (?–today)
- XML Professional Publisher (1980–today)
- Arbortext Advanced Print Publisher (1980–today)
- Microsoft Word (1983–today)
- QuarkXPress (1987–today)
- Quark CopyDesk (1991–today)
- Antennahouse Formatter XSL-FO or CSS (2000–today)
- Prince (2003–today)
- Apple Pages (2005–today)
- PDFReactor (2006?–today)
- DocBook XSL (2012?–today)
- typeset.sh by Paddle (2020–today)
- BFO Publisher (2022–today)
- …

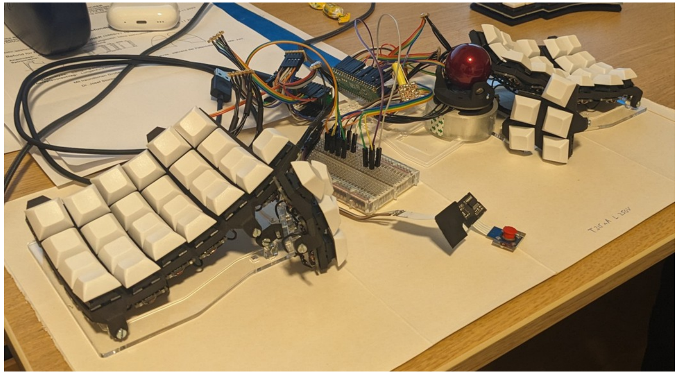
Keypress to screen



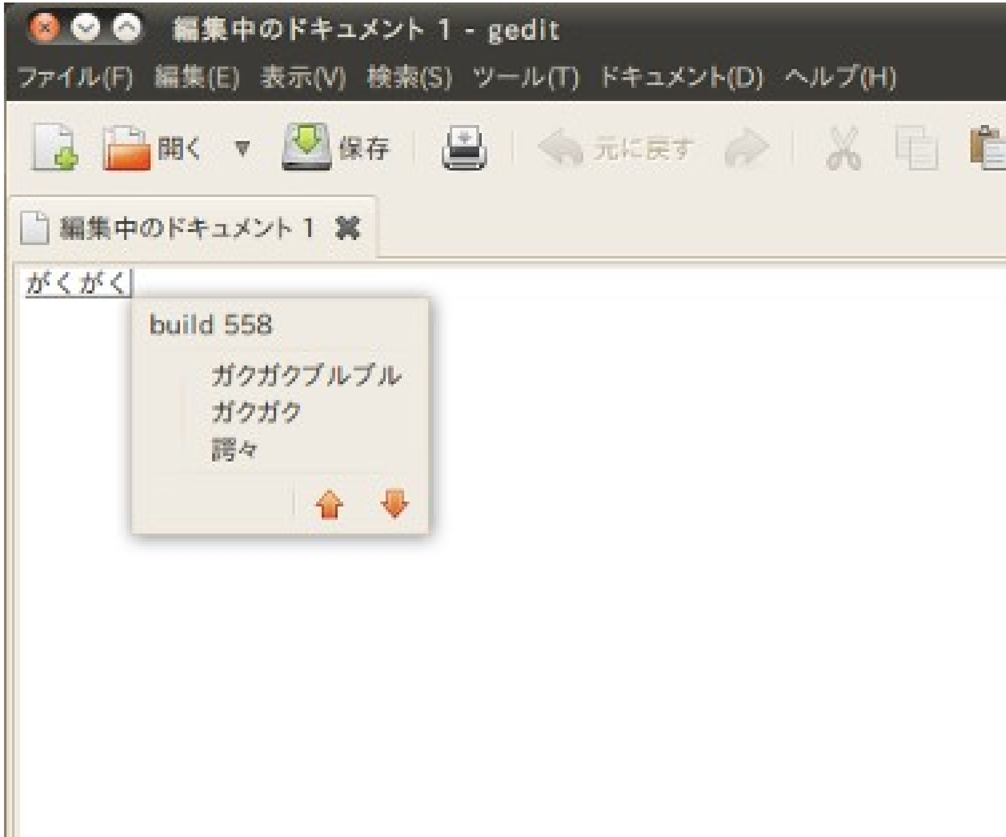
Typing in typesetting







Typing in typesetting


編集中のドキュメント 1 - gedit
ファイル(F) 編集(E) 表示(V) 検索(S) ツール(T) ドキュメント(D) ヘルプ(H)
開く
保存
元に戻す
編集中のドキュメント 1

がくがく

build 558

ガクガクブルブル
ガクガク
諤々

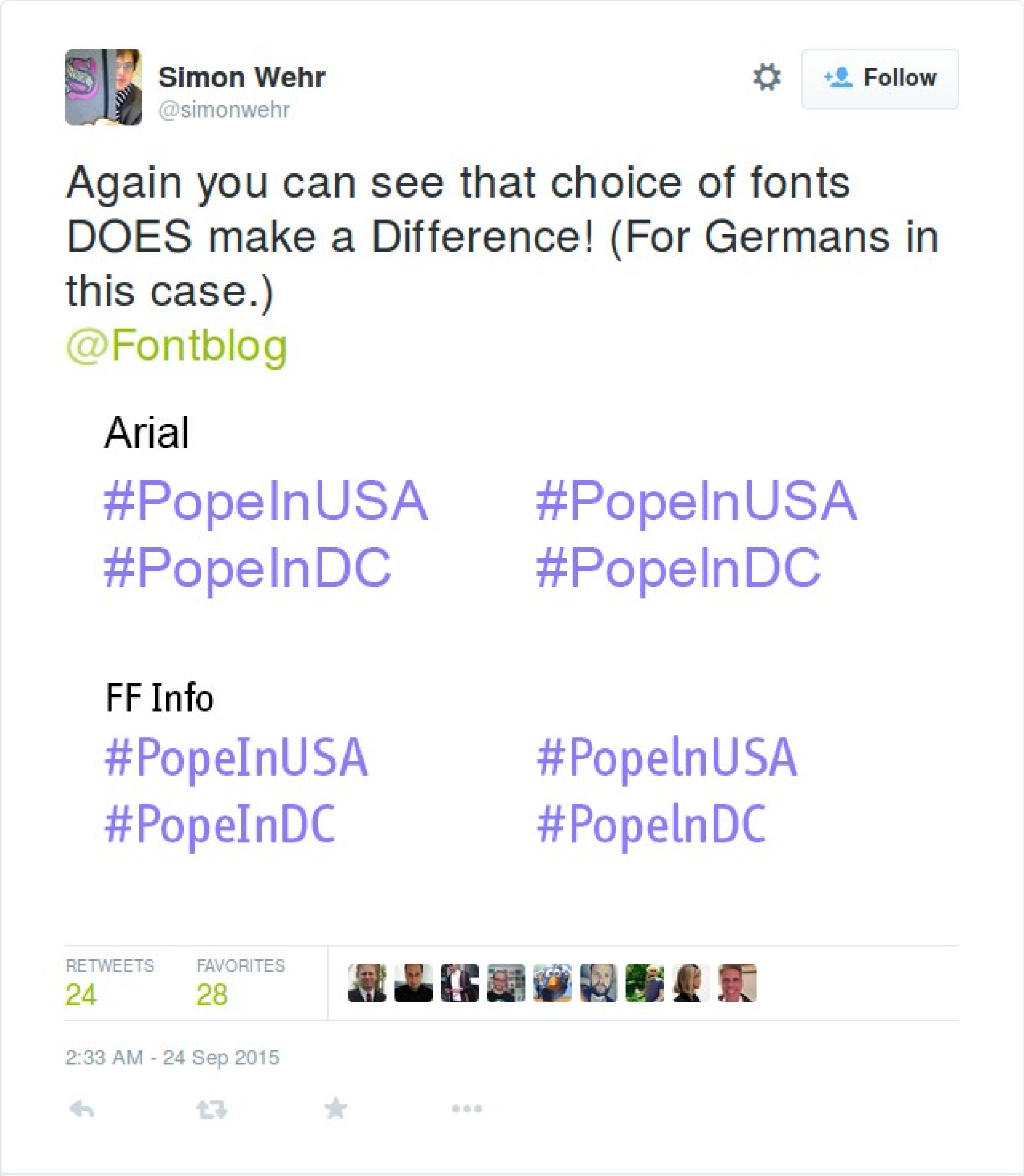
The domain

1) read input files
2) understand the sequence of characters
3) select appropriate fonts
4) select appropriate glyphs from the fonts
5) arrange on page
6) write output files

The domain

1) read input files (JPEG/PNG/SVG/ML/OpenType/…)
2) understand the sequence of characters (Unicode)
3) select appropriate fonts
4) select appropriate glyphs from the fonts (HarfBuzz)
5) arrange on page (CSS/line-breaking/hyphenation/…)
6) write output files (HTML5/CSS/PDF/EPUB/…)

Chapter 2.
Fun~~damentals~~



Simon Wehr
@simonwehr

Follow

Again you can see that choice of fonts DOES make a Difference! (For Germans in this case.)
@Fontblog

Arial

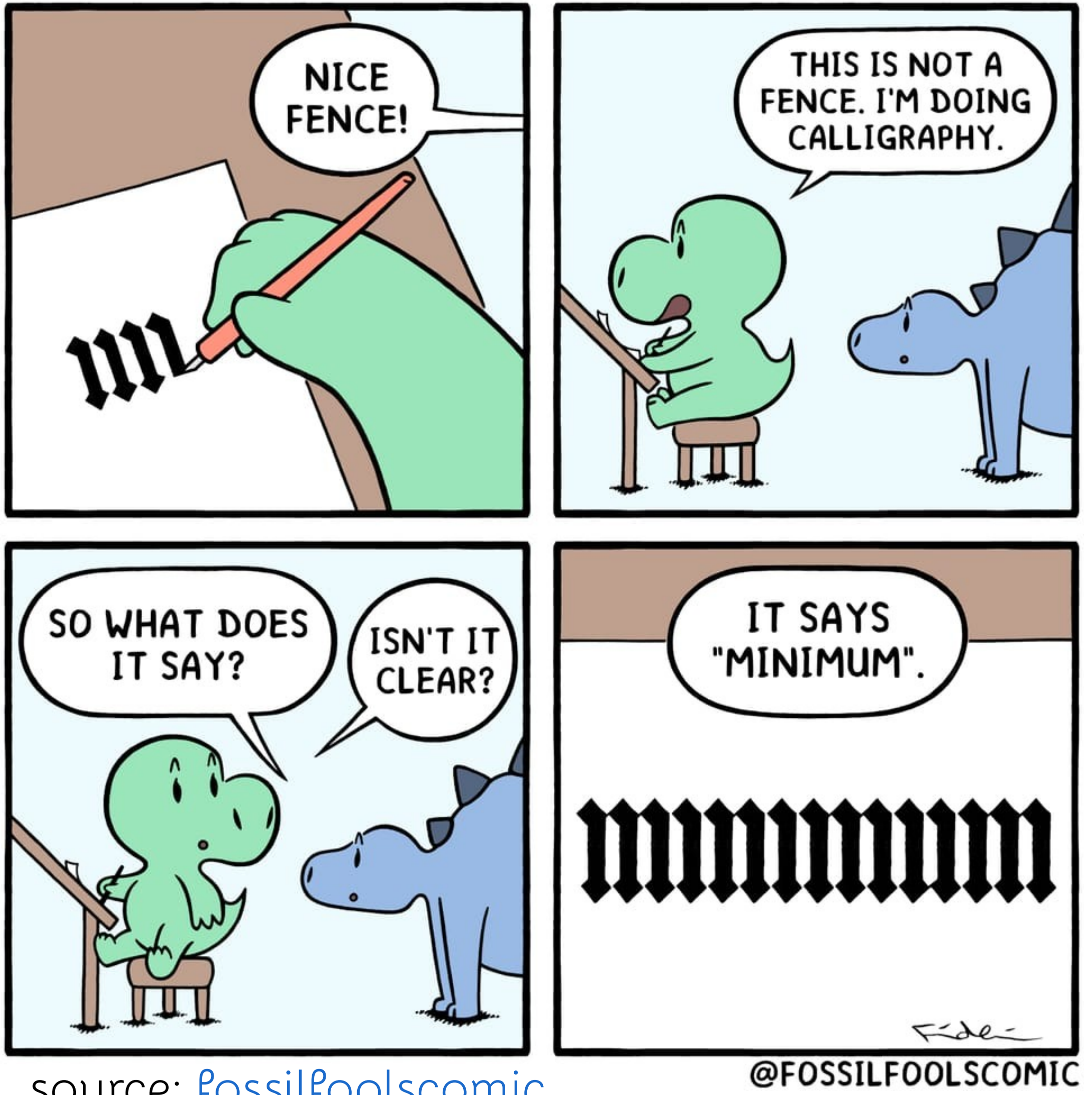
#PopeInUSA #PopelnUSA
#PopeInDC #PopelnDC

FF Info

#PopeInUSA #PopelnUSA
#PopeInDC #PopelnDC

RETWEETS 24 FAVORITES 28

2:33 AM - 24 Sep 2015

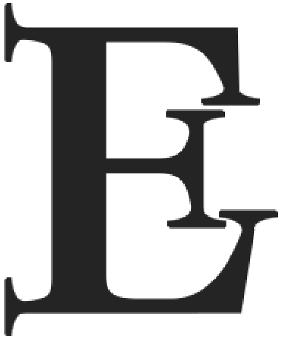


source: fossilfoolscomic

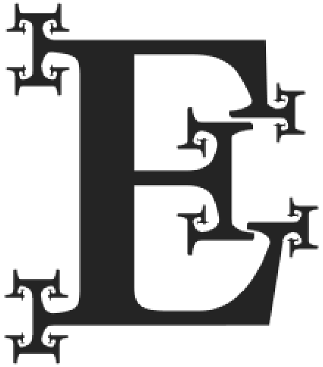
Sans-serif

Serif

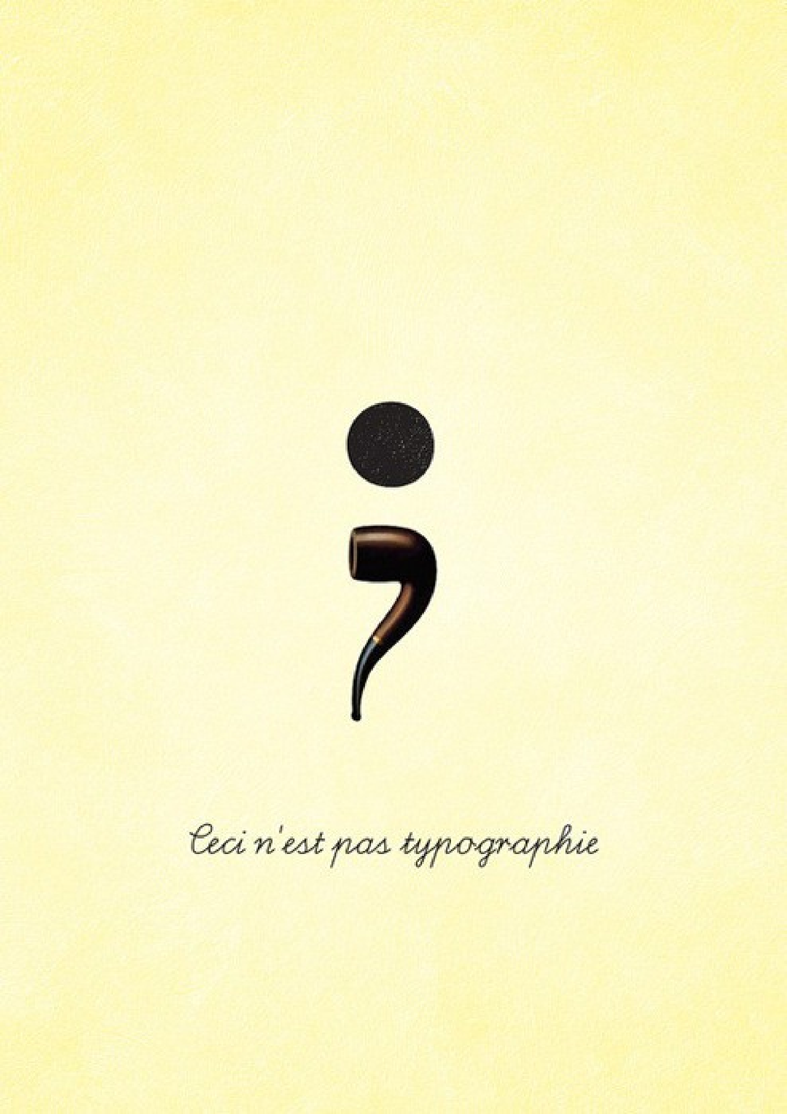
Extra-serif



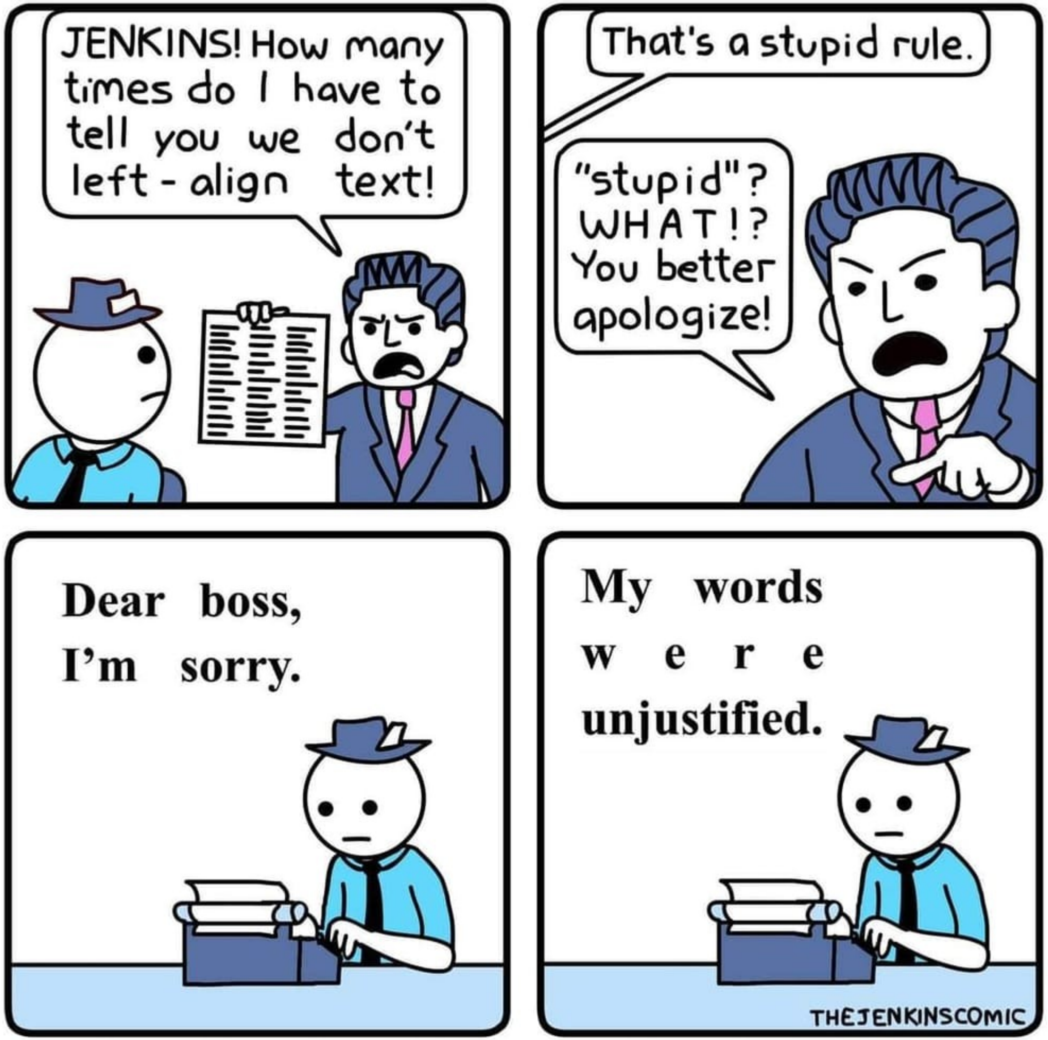
Fractal-Serif



source: samim.io?



source: ffffound on pinterest?


JENKINS! How many times do I have to tell you we don't left - align text!
That's a stupid rule.
"stupid"? WHAT!? You better apologize!
Dear boss, I'm sorry.
My words w e r e unjustified.
THEJENKINSCOMIC

source: thejenkinscomic

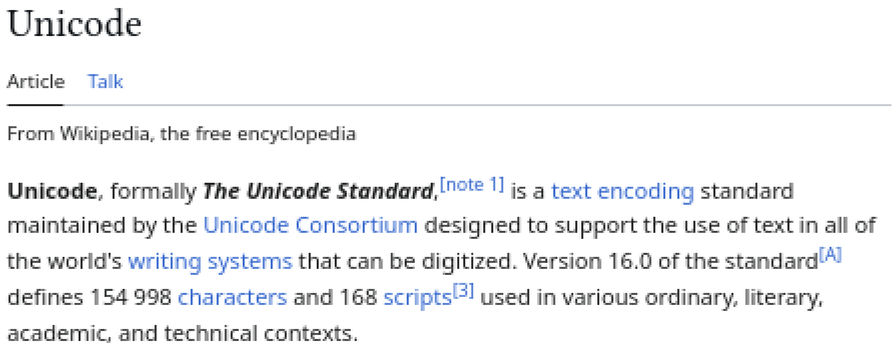
Pink Elephant

Chapter 3. Fundamentals

Section (1/3).

Unicode

Unicode



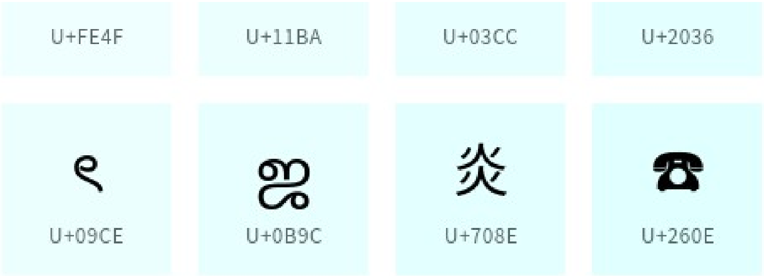
Unicode

Article Talk

From Wikipedia, the free encyclopedia

Unicode, formally ***The Unicode Standard***,[note 1] is a text encoding standard maintained by the Unicode Consortium designed to support the use of text in all of the world's writing systems that can be digitized. Version 16.0 of the standard[A] defines 154 998 characters and 168 scripts[3] used in various ordinary, literary, academic, and technical contexts.

UNICODE



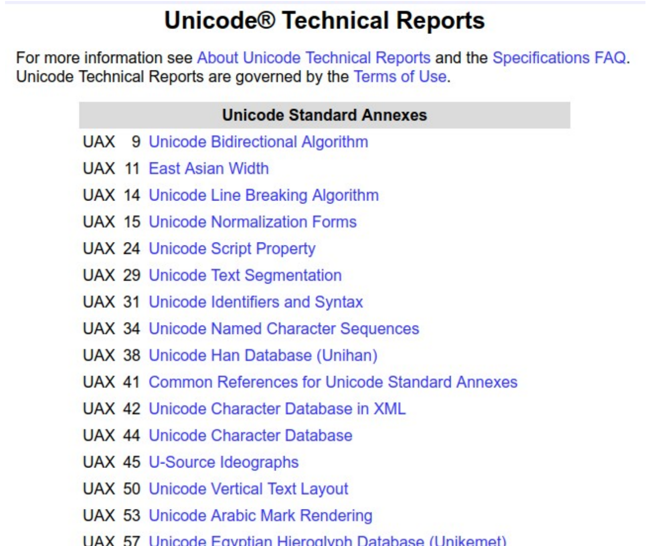
U+FE4F U+11BA U+03CC U+2036

ৎ U+09CE

ஜ U+0B9C

炎 U+708E

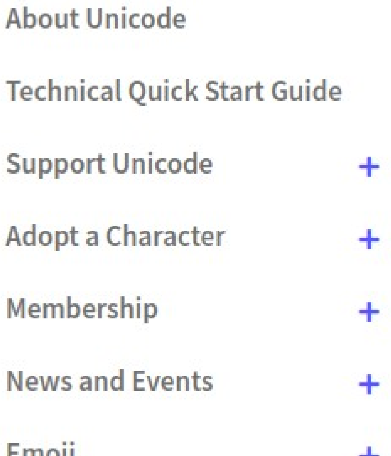
☎ U+260E



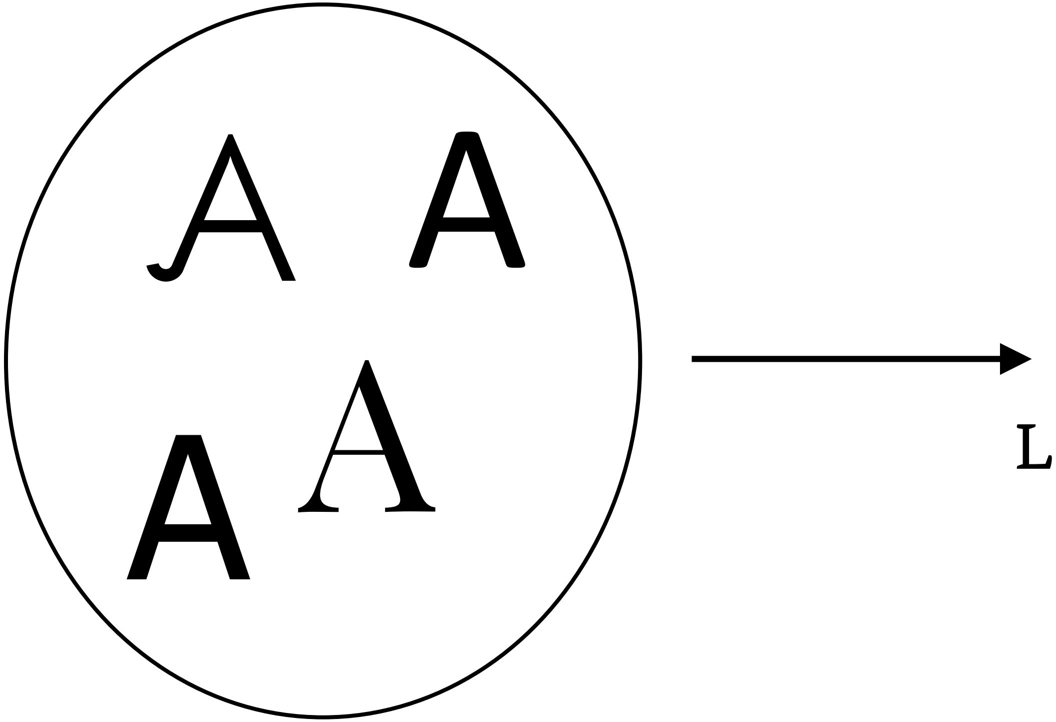
Unicode® Technical Reports

For more information see About Unicode Technical Reports and the Specifications FAQ.
Unicode Technical Reports are governed by the Terms of Use.

Unicode Standard Annexes		
UAX	9	Unicode Bidirectional Algorithm
UAX	11	East Asian Width
UAX	14	Unicode Line Breaking Algorithm
UAX	15	Unicode Normalization Forms
UAX	24	Unicode Script Property
UAX	29	Unicode Text Segmentation
UAX	31	Unicode Identifiers and Syntax
UAX	34	Unicode Named Character Sequences
UAX	38	Unicode Han Database (Unihan)
UAX	41	Common References for Unicode Standard Annexes
UAX	42	Unicode Character Database in XML
UAX	44	Unicode Character Database
UAX	45	U-Source Ideographs
UAX	50	Unicode Vertical Text Layout
UAX	53	Unicode Arabic Mark Rendering
UAX	57	Unicode Egyptian Hieroglyph Database (Unikemet)



Unicode

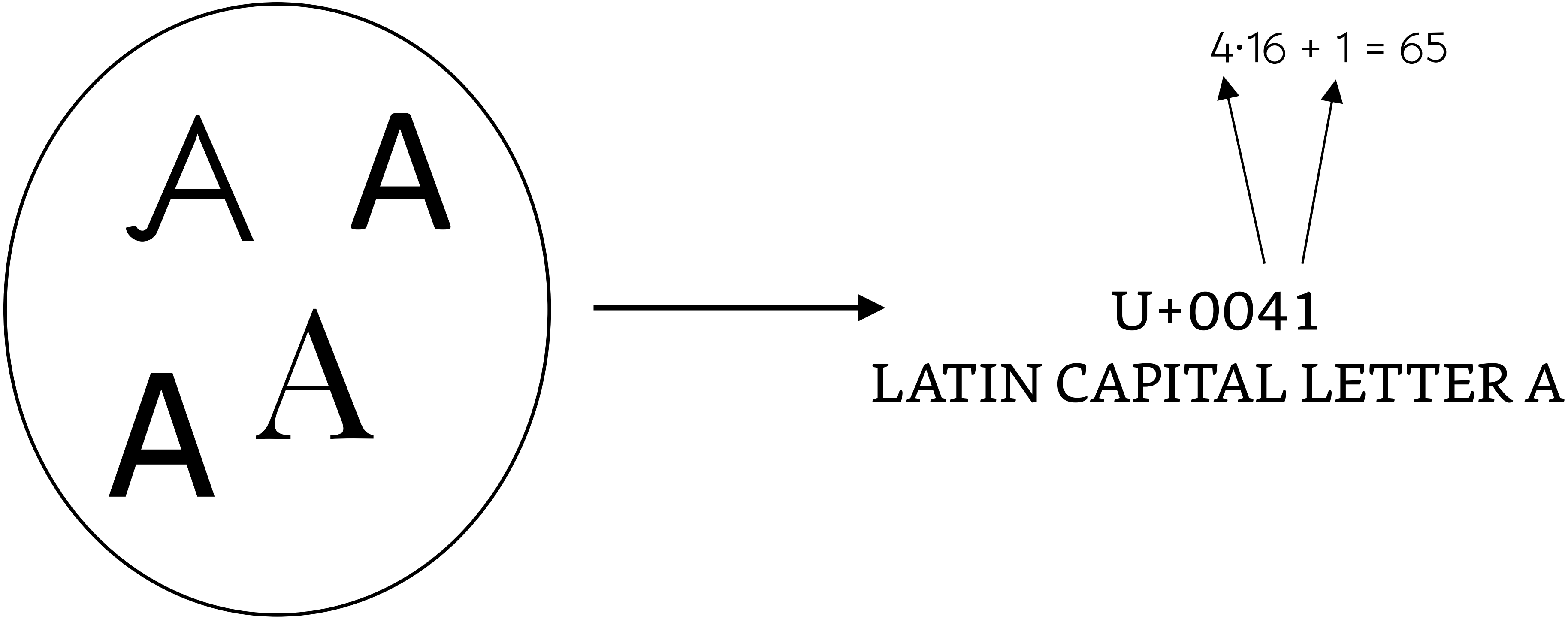


U+0041
LATIN CAPITAL LETTER A

element of a writing system

Unicode identifier

Unicode



4·16 + 1 = 65

U+0041

LATIN CAPITAL LETTER A

element of a writing system

Unicode identifier

Unicode

A

U+0041 LATIN CAPITAL LETTER A

B

U+0042 LATIN CAPITAL LETTER B

C

U+0043 LATIN CAPITAL LETTER C

...

Unicode

A

U+0041 LATIN CAPITAL LETTER A

B

U+0042 LATIN CAPITAL LETTER B

C

U+0043 LATIN CAPITAL LETTER C

…

U+2026 HORIZONTAL ELLIPSIS

Unicode

木

U+6728 CJK UNIFIED IDEOGRAPH-6728

𒀍

U+1200D CUNEIFORM SIGN AB TIMES GAL

ع

U+0639 ARABIC LETTER AIN

Unicode

鬱

U+9B31 CJK UNIFIED IDEOGRAPH-9B31

💩

U+1F4A9 PILE OF POO

﷽

U+FDFD ARABIC LIGATURE BISMILLAH AR-RAHMAN AR-RAHEEM
(basmala)

Unicode

Combining characters:

🥷

U+1F977 NINJA

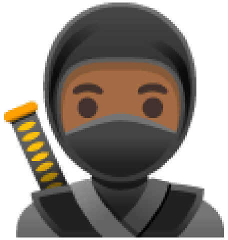
🥷🏾

U+1F977 NINJA

U+1F3FE EMOJI MODIFIER FITZPATRICK TYPE-5

Unicode

Combining characters:



U+1F977 NINJA

U+1F977 NINJA

U+1F3FE EMOJI MODIFIER FITZPATRICK TYPE-5

Unicode

Combining characters:

🥷

U+1F977 NINJA

🥷🏾

U+1F977 NINJA

↑

U+1F3FE EMOJI MODIFIER
FITZPATRICK TYPE-5

Unicode

Combining characters:

U+1F977 NINJA

�

↑

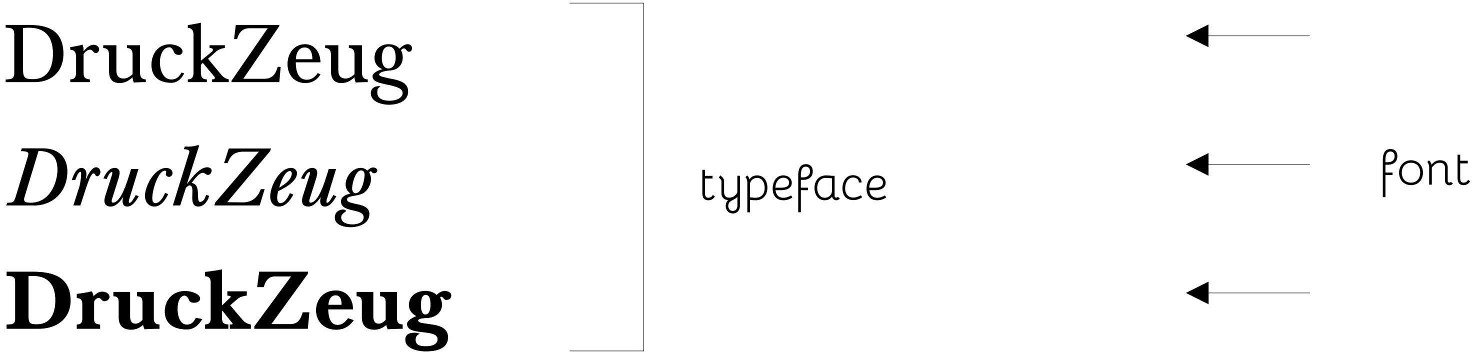
U+1F3FE EMOJI MODIFIER FITZPATRICK TYPE-5

U+1F977 NINJA

Section (2/3).

OpenType

OpenType



DruckZeug

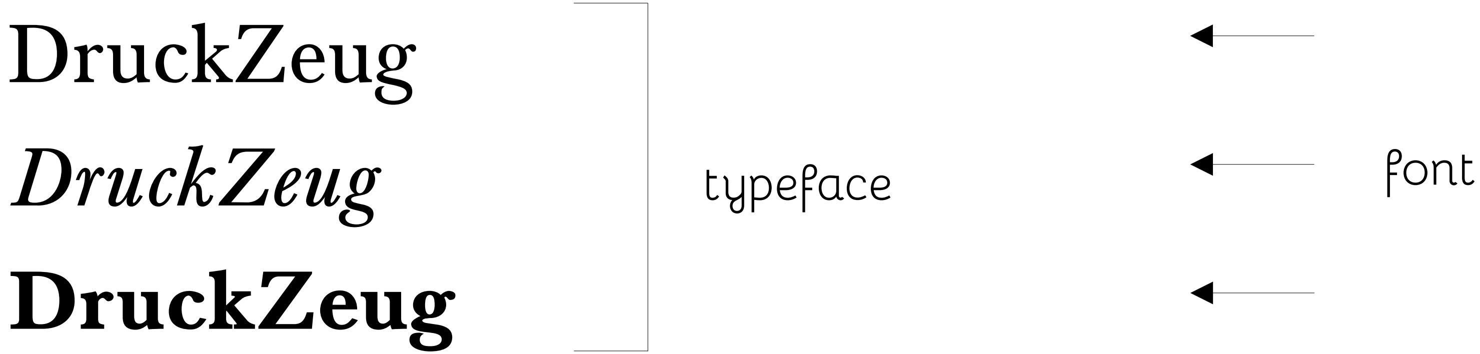
DruckZeug

DruckZeug

typeface

font

OpenType



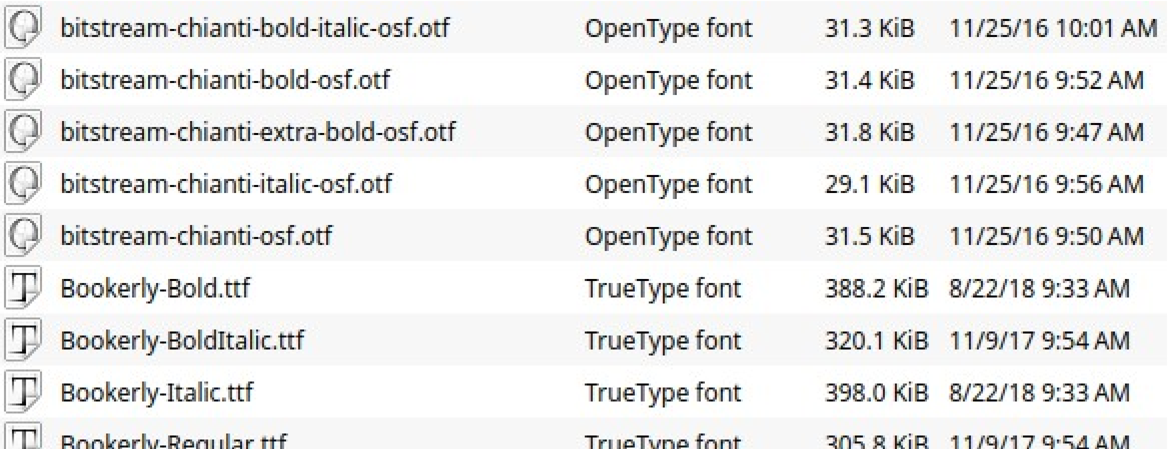
DruckZeug

DruckZeug

DruckZeug

typeface

font



bitstream-chianti-bold-italic-osf.otf	OpenType font	31.3 KiB	11/25/16 10:01 AM
bitstream-chianti-bold-osf.otf	OpenType font	31.4 KiB	11/25/16 9:52 AM
bitstream-chianti-extra-bold-osf.otf	OpenType font	31.8 KiB	11/25/16 9:47 AM
bitstream-chianti-italic-osf.otf	OpenType font	29.1 KiB	11/25/16 9:56 AM
bitstream-chianti-osf.otf	OpenType font	31.5 KiB	11/25/16 9:50 AM
Bookerly-Bold.ttf	TrueType font	388.2 KiB	8/22/18 9:33 AM
Bookerly-BoldItalic.ttf	TrueType font	320.1 KiB	11/9/17 9:54 AM
Bookerly-Italic.ttf	TrueType font	398.0 KiB	8/22/18 9:33 AM
Bookerly-Regular.ttf	TrueType font	305.8 KiB	11/9/17 9:54 AM

OpenType file

- A database
  - GlyphOrder table
  - head table
  - hhea table
  - maxp table
  - OS/2 table
  - name table
  - cmap table
  - post table
  - CFF table
  - GPOS table
  - hmtx table
  - ...

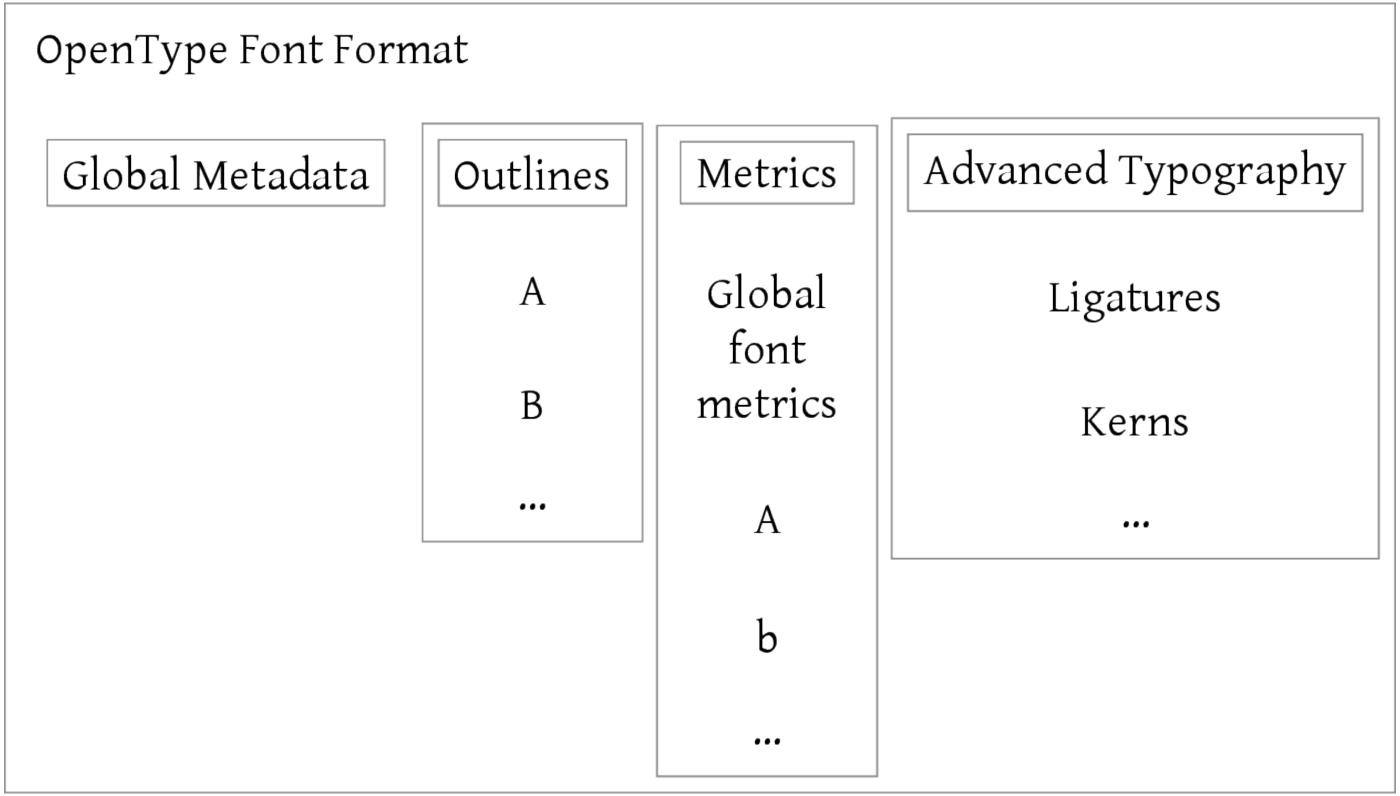
python software “fonttools”:

```
$ fonttools ttx ~/.fonts/Fontin-Regular.otf
Dumping "Fontin-Regular.otf" to "Fontin-Regular.ttx"...
Dumping 'GlyphOrder' table...
Dumping 'head' table...
```

Web application by Roel Nieskens/PixelAmbacht:

Wakamafondue

OpenType file



OpenType Font Format

Global Metadata

Outlines

A

B

...

Metrics

Global font metrics

A

b

...

Advanced Typography

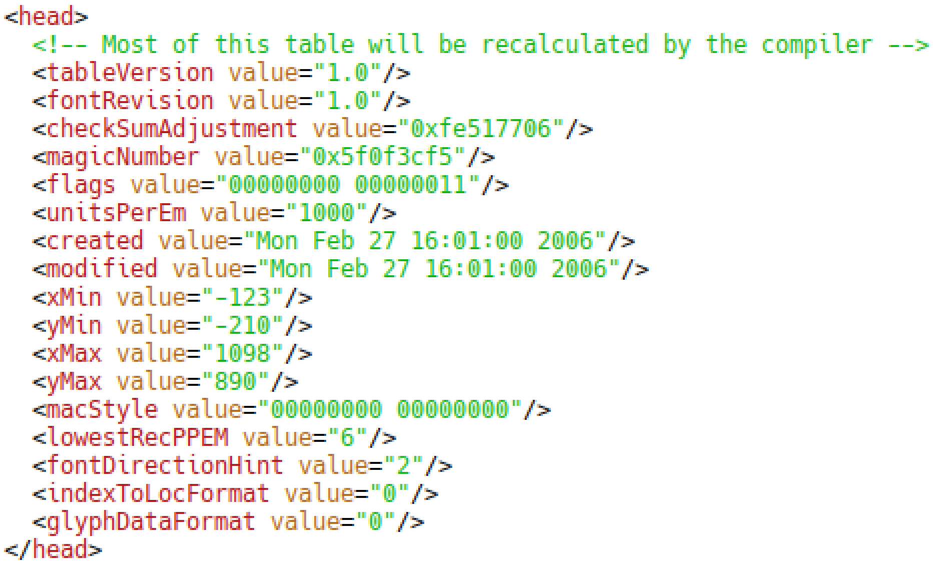
Ligatures

Kerns

...

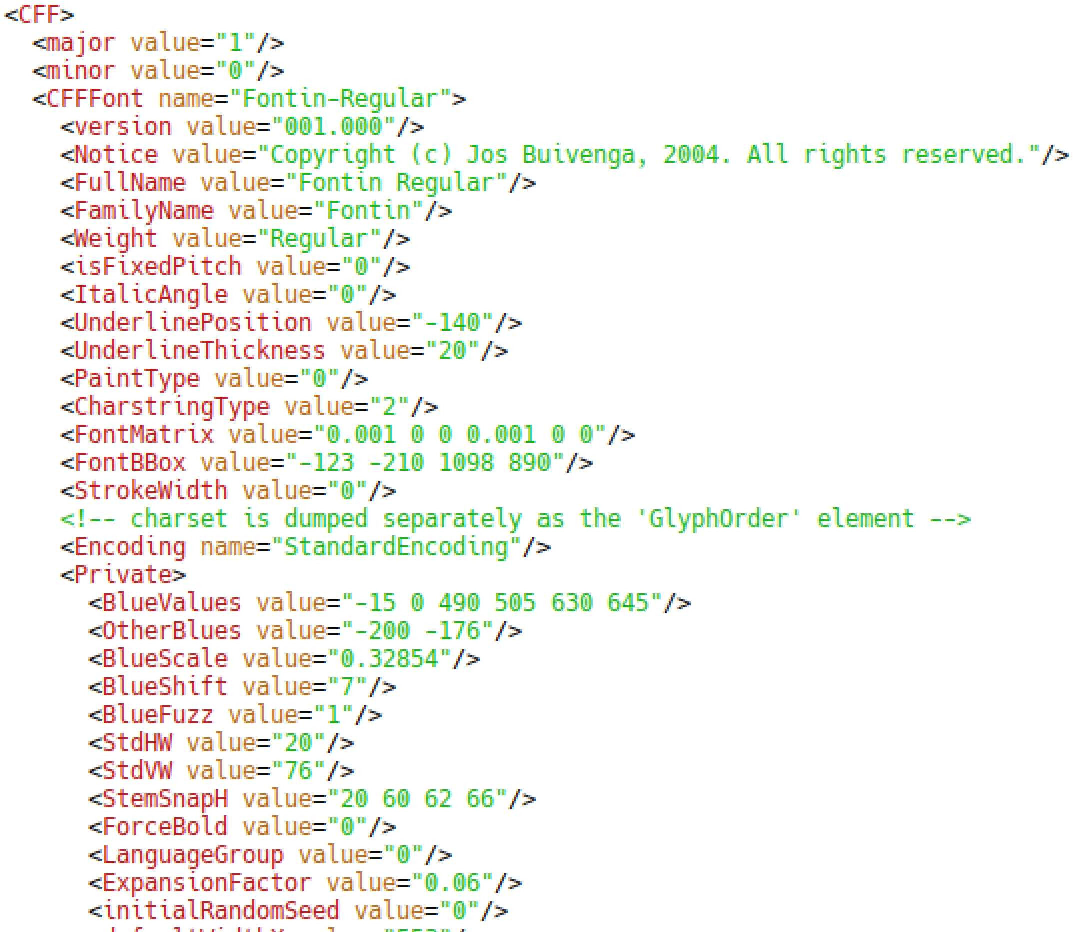
source: "Font and Layout for Global Scripts" by Simon Cozens

OpenType file



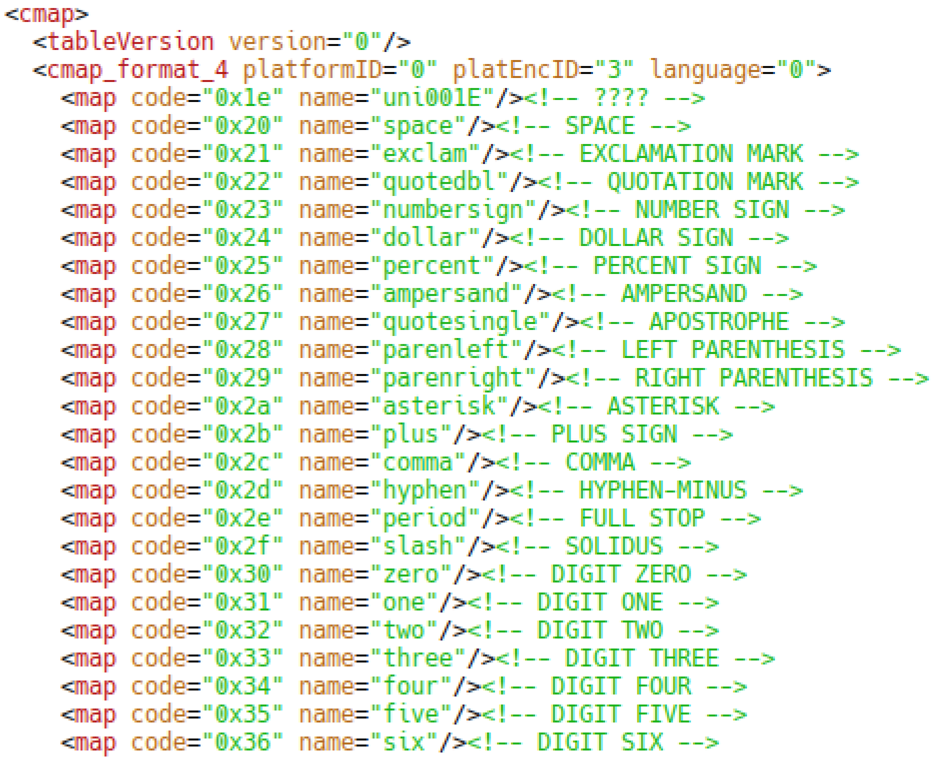
```
<head>
  <!-- Most of this table will be recalculated by the compiler -->
  <tableVersion value="1.0"/>
  <fontRevision value="1.0"/>
  <checkSumAdjustment value="0xfe517706"/>
  <magicNumber value="0x5f0f3cf5"/>
  <flags value="00000000 00000011"/>
  <unitsPerEm value="1000"/>
  <created value="Mon Feb 27 16:01:00 2006"/>
  <modified value="Mon Feb 27 16:01:00 2006"/>
  <xMin value="-123"/>
  <yMin value="-210"/>
  <xMax value="1098"/>
  <yMax value="890"/>
  <macStyle value="00000000 00000000"/>
  <lowestRecPPEM value="6"/>
  <fontDirectionHint value="2"/>
  <indexToLocFormat value="0"/>
  <glyphDataFormat value="0"/>
</head>
```

OpenType file



```
<CFF>
  <major value="1"/>
  <minor value="0"/>
  <CFFFont name="Fontin-Regular">
    <version value="001.000"/>
    <Notice value="Copyright (c) Jos Buivenga, 2004. All rights reserved."/>
    <FullName value="Fontin Regular"/>
    <FamilyName value="Fontin"/>
    <Weight value="Regular"/>
    <isFixedPitch value="0"/>
    <ItalicAngle value="0"/>
    <UnderlinePosition value="-140"/>
    <UnderlineThickness value="20"/>
    <PaintType value="0"/>
    <CharstringType value="2"/>
    <FontMatrix value="0.001 0 0 0.001 0 0"/>
    <FontBBox value="-123 -210 1098 890"/>
    <StrokeWidth value="0"/>
    <!-- charset is dumped separately as the 'GlyphOrder' element -->
    <Encoding name="StandardEncoding"/>
    <Private>
      <BlueValues value="-15 0 490 505 630 645"/>
      <OtherBlues value="-200 -176"/>
      <BlueScale value="0.32854"/>
      <BlueShift value="7"/>
      <BlueFuzz value="1"/>
      <StdHW value="20"/>
      <StdVW value="76"/>
      <StemSnapH value="20 60 62 66"/>
      <ForceBold value="0"/>
      <LanguageGroup value="0"/>
      <ExpansionFactor value="0.06"/>
      <initialRandomSeed value="0"/>
```

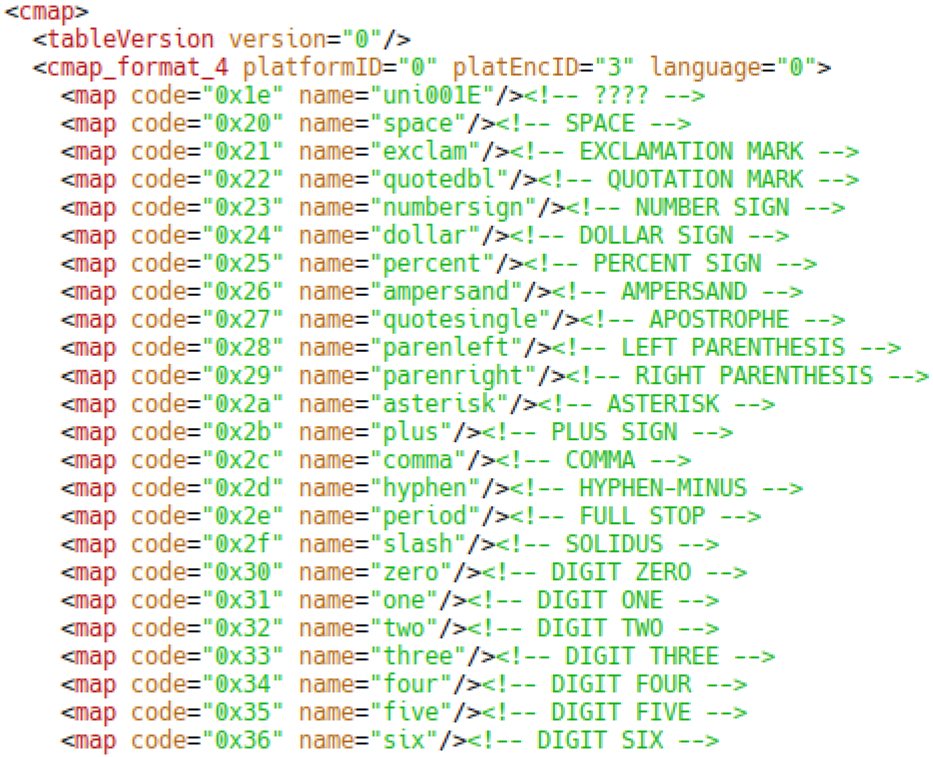
OpenType file



```
<cmap>
  <tableVersion version="0"/>
  <cmap_format_4 platformID="0" platEncID="3" language="0">
    <map code="0x1e" name="uni001E"/><!-- ???? -->
    <map code="0x20" name="space"/><!-- SPACE -->
    <map code="0x21" name="exclam"/><!-- EXCLAMATION MARK -->
    <map code="0x22" name="quotedbl"/><!-- QUOTATION MARK -->
    <map code="0x23" name="numbersign"/><!-- NUMBER SIGN -->
    <map code="0x24" name="dollar"/><!-- DOLLAR SIGN -->
    <map code="0x25" name="percent"/><!-- PERCENT SIGN -->
    <map code="0x26" name="ampersand"/><!-- AMPERSAND -->
    <map code="0x27" name="quotesingle"/><!-- APOSTROPHE -->
    <map code="0x28" name="parenleft"/><!-- LEFT PARENTHESIS -->
    <map code="0x29" name="parenright"/><!-- RIGHT PARENTHESIS -->
    <map code="0x2a" name="asterisk"/><!-- ASTERISK -->
    <map code="0x2b" name="plus"/><!-- PLUS SIGN -->
    <map code="0x2c" name="comma"/><!-- COMMA -->
    <map code="0x2d" name="hyphen"/><!-- HYPHEN-MINUS -->
    <map code="0x2e" name="period"/><!-- FULL STOP -->
    <map code="0x2f" name="slash"/><!-- SOLIDUS -->
    <map code="0x30" name="zero"/><!-- DIGIT ZERO -->
    <map code="0x31" name="one"/><!-- DIGIT ONE -->
    <map code="0x32" name="two"/><!-- DIGIT TWO -->
    <map code="0x33" name="three"/><!-- DIGIT THREE -->
    <map code="0x34" name="four"/><!-- DIGIT FOUR -->
    <map code="0x35" name="five"/><!-- DIGIT FIVE -->
    <map code="0x36" name="six"/><!-- DIGIT SIX -->
```

U+0036 DIGIT SIX → six

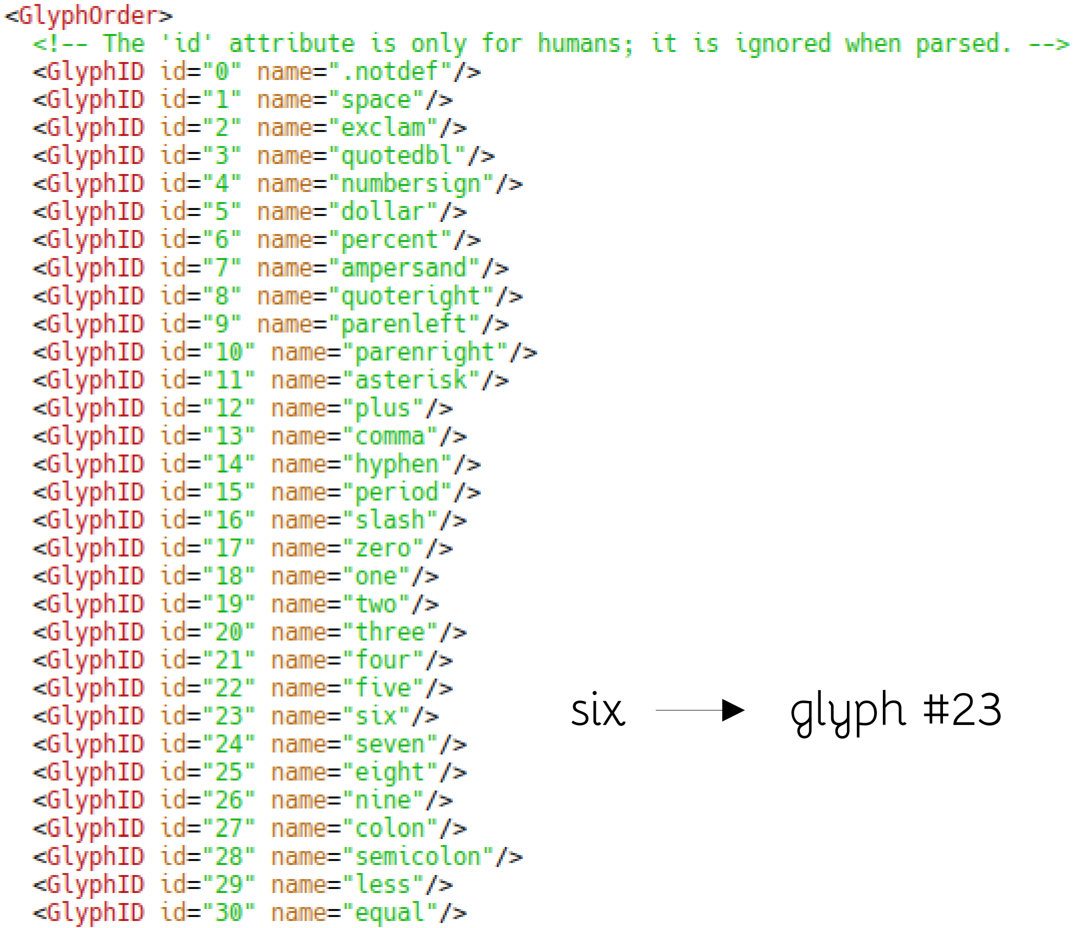
OpenType file



```
<cmap>
  <tableVersion version="0"/>
  <cmap_format_4 platformID="0" platEncID="3" language="0">
    <map code="0x1e" name="uni001E"/><!-- ???? -->
    <map code="0x20" name="space"/><!-- SPACE -->
    <map code="0x21" name="exclam"/><!-- EXCLAMATION MARK -->
    <map code="0x22" name="quotedbl"/><!-- QUOTATION MARK -->
    <map code="0x23" name="numbersign"/><!-- NUMBER SIGN -->
    <map code="0x24" name="dollar"/><!-- DOLLAR SIGN -->
    <map code="0x25" name="percent"/><!-- PERCENT SIGN -->
    <map code="0x26" name="ampersand"/><!-- AMPERSAND -->
    <map code="0x27" name="quotesingle"/><!-- APOSTROPHE -->
    <map code="0x28" name="parenleft"/><!-- LEFT PARENTHESIS -->
    <map code="0x29" name="parenright"/><!-- RIGHT PARENTHESIS -->
    <map code="0x2a" name="asterisk"/><!-- ASTERISK -->
    <map code="0x2b" name="plus"/><!-- PLUS SIGN -->
    <map code="0x2c" name="comma"/><!-- COMMA -->
    <map code="0x2d" name="hyphen"/><!-- HYPHEN-MINUS -->
    <map code="0x2e" name="period"/><!-- FULL STOP -->
    <map code="0x2f" name="slash"/><!-- SOLIDUS -->
    <map code="0x30" name="zero"/><!-- DIGIT ZERO -->
    <map code="0x31" name="one"/><!-- DIGIT ONE -->
    <map code="0x32" name="two"/><!-- DIGIT TWO -->
    <map code="0x33" name="three"/><!-- DIGIT THREE -->
    <map code="0x34" name="four"/><!-- DIGIT FOUR -->
    <map code="0x35" name="five"/><!-- DIGIT FIVE -->
    <map code="0x36" name="six"/><!-- DIGIT SIX -->
```

U+0036 DIGIT SIX → six

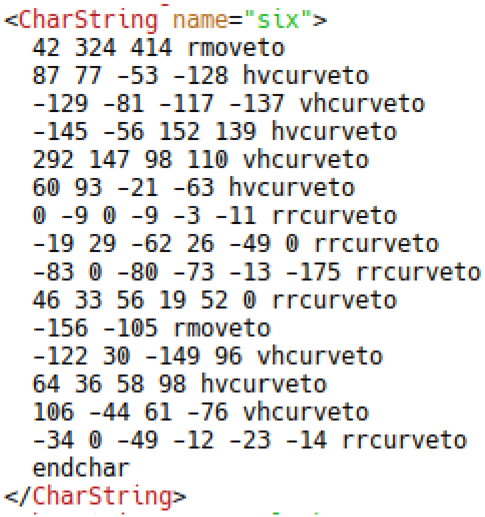
OpenType file



```
<GlyphOrder>
  <!-- The 'id' attribute is only for humans; it is ignored when parsed. -->
  <GlyphID id="0" name=".notdef"/>
  <GlyphID id="1" name="space"/>
  <GlyphID id="2" name="exclam"/>
  <GlyphID id="3" name="quotedbl"/>
  <GlyphID id="4" name="numbersign"/>
  <GlyphID id="5" name="dollar"/>
  <GlyphID id="6" name="percent"/>
  <GlyphID id="7" name="ampersand"/>
  <GlyphID id="8" name="quoteright"/>
  <GlyphID id="9" name="parenleft"/>
  <GlyphID id="10" name="parenright"/>
  <GlyphID id="11" name="asterisk"/>
  <GlyphID id="12" name="plus"/>
  <GlyphID id="13" name="comma"/>
  <GlyphID id="14" name="hyphen"/>
  <GlyphID id="15" name="period"/>
  <GlyphID id="16" name="slash"/>
  <GlyphID id="17" name="zero"/>
  <GlyphID id="18" name="one"/>
  <GlyphID id="19" name="two"/>
  <GlyphID id="20" name="three"/>
  <GlyphID id="21" name="four"/>
  <GlyphID id="22" name="five"/>
  <GlyphID id="23" name="six"/>
  <GlyphID id="24" name="seven"/>
  <GlyphID id="25" name="eight"/>
  <GlyphID id="26" name="nine"/>
  <GlyphID id="27" name="colon"/>
  <GlyphID id="28" name="semicolon"/>
  <GlyphID id="29" name="less"/>
  <GlyphID id="30" name="equal"/>
```

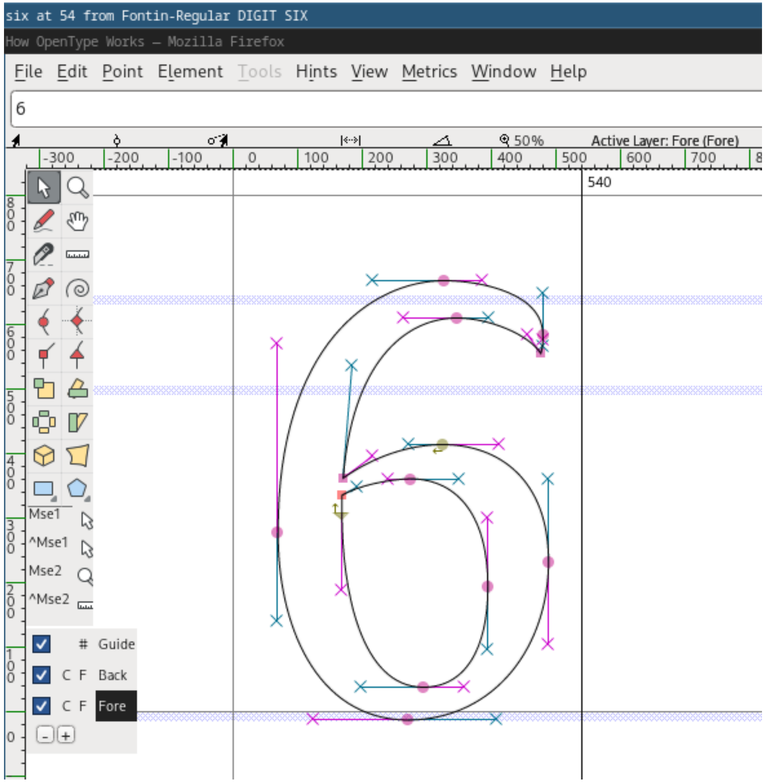
six → glyph #23

OpenType file



```
<CharString name="six">
  42 324 414 rmoveto
  87 77 -53 -128 hvcurveto
  -129 -81 -117 -137 vhcurveto
  -145 -56 152 139 hvcurveto
  292 147 98 110 vhcurveto
  60 93 -21 -63 hvcurveto
  0 -9 0 -9 -3 -11 rrcurveto
  -19 29 -62 26 -49 0 rrcurveto
  -83 0 -80 -73 -13 -175 rrcurveto
  46 33 56 19 52 0 rrcurveto
  -156 -105 rmoveto
  -122 30 -149 96 vhcurveto
  64 36 58 98 hvcurveto
  106 -44 61 -76 vhcurveto
  -34 0 -49 -12 -23 -14 rrcurveto
  endchar
</CharString>
```

in CFF table

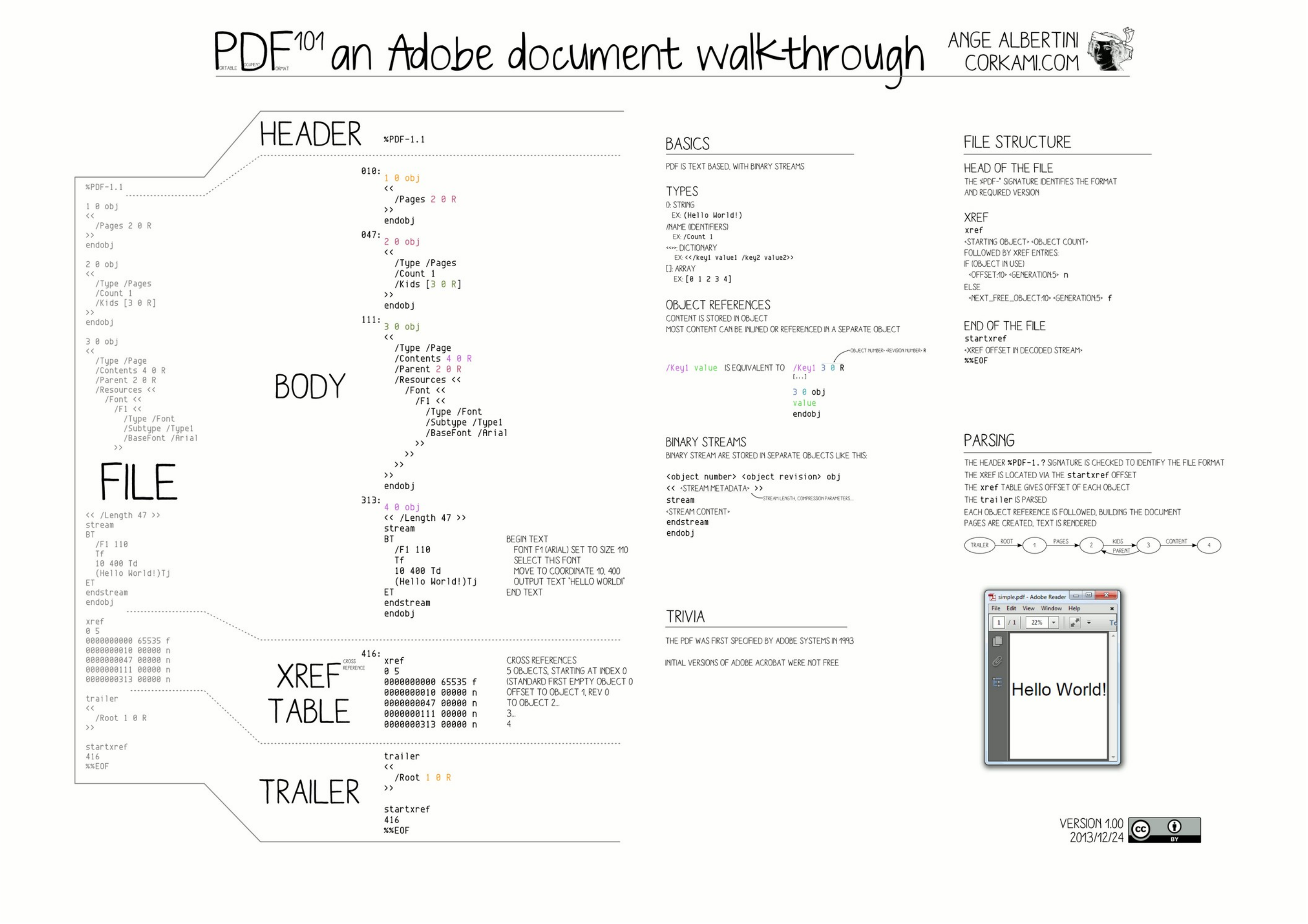


opened with FontForge

Section (3/3).

PDF

PDF file



PDF 101 an Adobe document walkthrough

PORTABLE DOCUMENT FORMAT

ANGE ALBERTINI
CORKAMI.COM

FILE

```
%PDF-1.1

1 0 obj
<<
  /Pages 2 0 R
>>
endobj

2 0 obj
<<
  /Type /Pages
  /Count 1
  /Kids [3 0 R]
>>
endobj

3 0 obj
<<
  /Type /Page
  /Contents 4 0 R
  /Parent 2 0 R
  /Resources <<
    /Font <<
      /F1 <<
        /Type /Font
        /Subtype /Type1
        /BaseFont /Arial
      >>
<< /Length 47 >>
stream
BT
  /F1 110
  Tf
  10 400 Td
  (Hello World!)Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000047 00000 n
0000000111 00000 n
0000000313 00000 n

trailer
<<
  /Root 1 0 R
>>

startxref
416
%%EOF
```

HEADER

```
%PDF-1.1
```

BODY

```
010:
    1 0 obj
    <<
      /Pages 2 0 R
    >>
    endobj
047:
    2 0 obj
    <<
      /Type /Pages
      /Count 1
      /Kids [3 0 R]
    >>
    endobj
111:
    3 0 obj
    <<
      /Type /Page
      /Contents 4 0 R
      /Parent 2 0 R
      /Resources <<
        /Font <<
          /F1 <<
            /Type /Font
            /Subtype /Type1
            /BaseFont /Arial
          >>
        >>
      >>
    >>
    endobj
313:
    4 0 obj
    << /Length 47 >>
    stream
    BT                        BEGIN TEXT
      /F1 110                   FONT F1 (ARIAL) SET TO SIZE 110
      Tf                        SELECT THIS FONT
      10 400 Td                 MOVE TO COORDINATE 10, 400
      (Hello World!)Tj          OUTPUT TEXT "HELLO WORLD!"
    ET                        END TEXT
    endstream
    endobj
```

XREF TABLE (CROSS REFERENCE)

```
416:
    xref                      CROSS REFERENCES
    0 5                       5 OBJECTS, STARTING AT INDEX 0
    0000000000 65535 f        (STANDARD FIRST EMPTY OBJECT 0
    0000000010 00000 n        OFFSET TO OBJECT 1, REV 0
    0000000047 00000 n        TO OBJECT 2...
    0000000111 00000 n        3...
    0000000313 00000 n        4
```

TRAILER

```
    trailer
    <<
      /Root 1 0 R
    >>

    startxref
    416
    %%EOF
```

BASICS

PDF IS TEXT BASED, WITH BINARY STREAMS

TYPES

(): STRING
  EX: (Hello World!)
/NAME (IDENTIFIERS)
  EX: /Count 1
<<>>: DICTIONARY
  EX: <</key1 value1 /key2 value2>>
[]: ARRAY
  EX: [0 1 2 3 4]

OBJECT REFERENCES

CONTENT IS STORED IN OBJECT
MOST CONTENT CAN BE INLINED OR REFERENCED IN A SEPARATE OBJECT

```
/Key1 value  IS EQUIVALENT TO  /Key1 3 0 R
                               [...]

                               3 0 obj
                               value
                               endobj
```

<OBJECT NUMBER> <REVISION NUMBER> R

BINARY STREAMS

BINARY STREAM ARE STORED IN SEPARATE OBJECTS LIKE THIS:

```
<object number> <object revision> obj
<< <STREAM METADATA> >>
stream
<STREAM CONTENT>
endstream
endobj
```

STREAM LENGTH, COMPRESSION PARAMETERS...

TRIVIA

THE PDF WAS FIRST SPECIFIED BY ADOBE SYSTEMS IN 1993

INITIAL VERSIONS OF ADOBE ACROBAT WERE NOT FREE

FILE STRUCTURE

HEAD OF THE FILE

THE "%PDF-" SIGNATURE IDENTIFIES THE FORMAT
AND REQUIRED VERSION

XREF

xref
<STARTING OBJECT> <OBJECT COUNT>
FOLLOWED BY XREF ENTRIES:
IF (OBJECT IN USE)
  <OFFSET:10> <GENERATION:5> n
ELSE
  <NEXT_FREE_OBJECT:10> <GENERATION:5> f

END OF THE FILE

startxref
<XREF OFFSET IN DECODED STREAM>
%%EOF

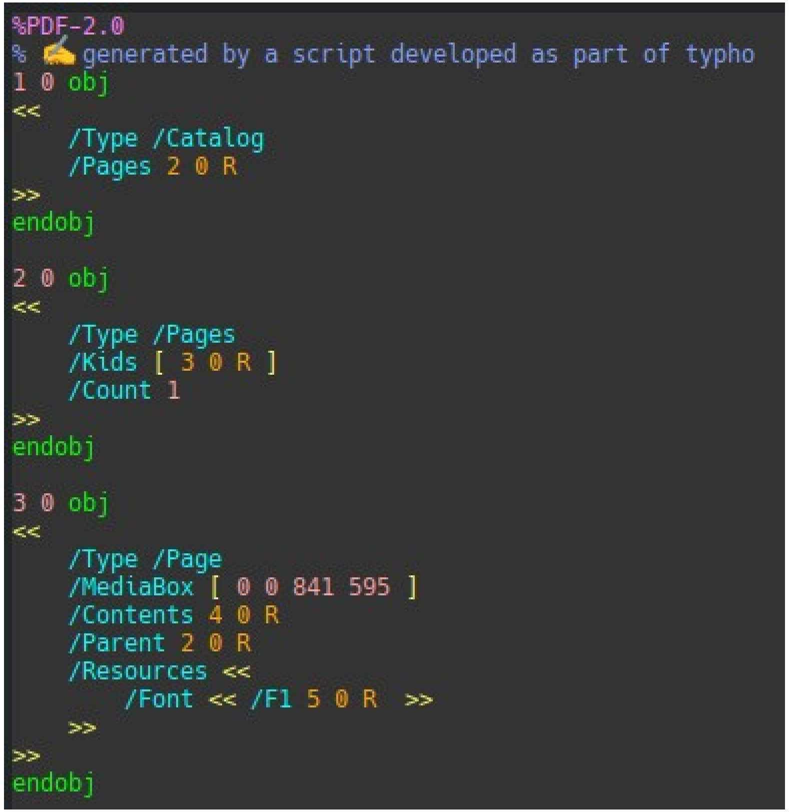
PARSING

THE HEADER %PDF-1.? SIGNATURE IS CHECKED TO IDENTIFY THE FILE FORMAT
THE XREF IS LOCATED VIA THE startxref OFFSET
THE xref TABLE GIVES OFFSET OF EACH OBJECT
THE trailer IS PARSED
EACH OBJECT REFERENCE IS FOLLOWED, BUILDING THE DOCUMENT
PAGES ARE CREATED, TEXT IS RENDERED

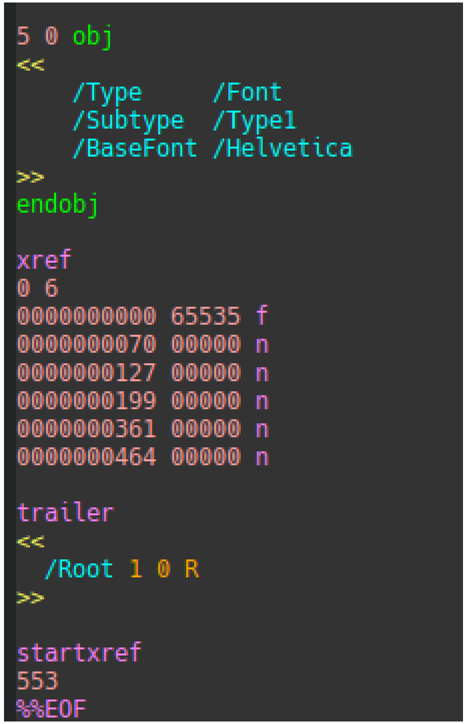
VERSION 1.00
2013/12/24

source: Ange Albertini

Creating a PDF



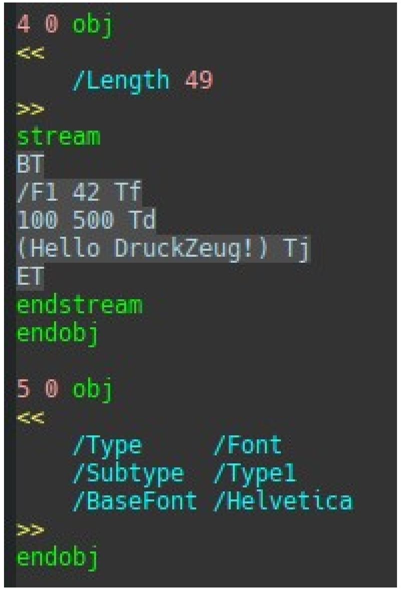
```
%PDF-2.0
% ✍ generated by a script developed as part of typho
1 0 obj
<<
    /Type /Catalog
    /Pages 2 0 R
>>
endobj

2 0 obj
<<
    /Type /Pages
    /Kids [ 3 0 R ]
    /Count 1
>>
endobj

3 0 obj
<<
    /Type /Page
    /MediaBox [ 0 0 841 595 ]
    /Contents 4 0 R
    /Parent 2 0 R
    /Resources <<
        /Font << /F1 5 0 R  >>
    >>
>>
endobj
```



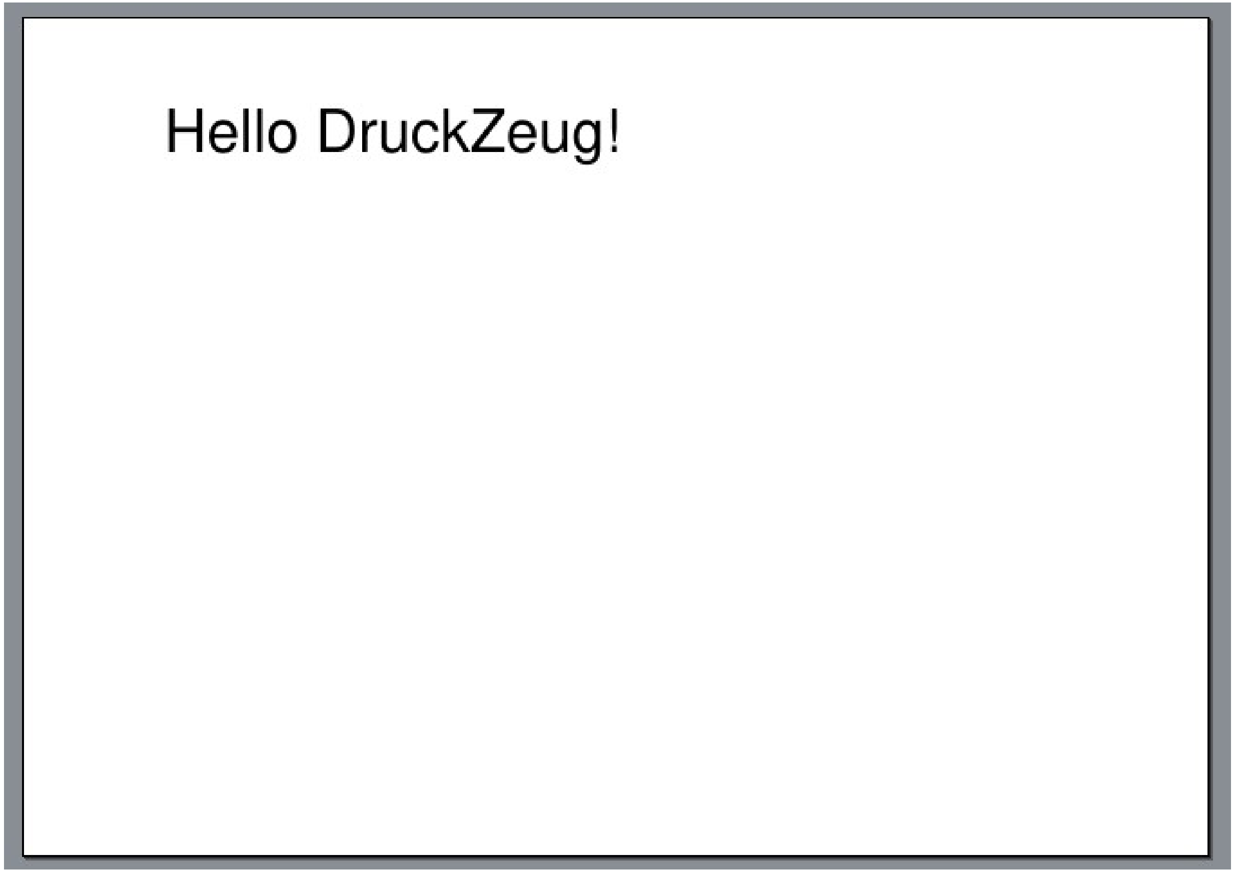
```
5 0 obj
<<
    /Type     /Font
    /Subtype  /Type1
    /BaseFont /Helvetica
>>
endobj

xref
0 6
0000000000 65535 f
0000000070 00000 n
0000000127 00000 n
0000000199 00000 n
0000000361 00000 n
0000000464 00000 n

trailer
<<
  /Root 1 0 R
>>

startxref
553
%%EOF
```

https://gendignoux.com/blog/2016/10/04/pdf-basics.html
https://news.speedata.de/2024/03/19/insidepdf-01/

Creating a PDF

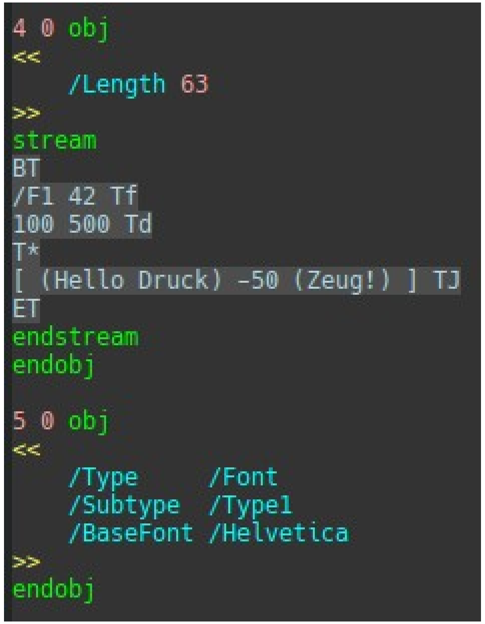


```
4 0 obj
<<
    /Length 49
>>
stream
BT
/F1 42 Tf
100 500 Td
(Hello DruckZeug!) Tj
ET
endstream
endobj

5 0 obj
<<
    /Type     /Font
    /Subtype  /Type1
    /BaseFont /Helvetica
>>
endobj
```

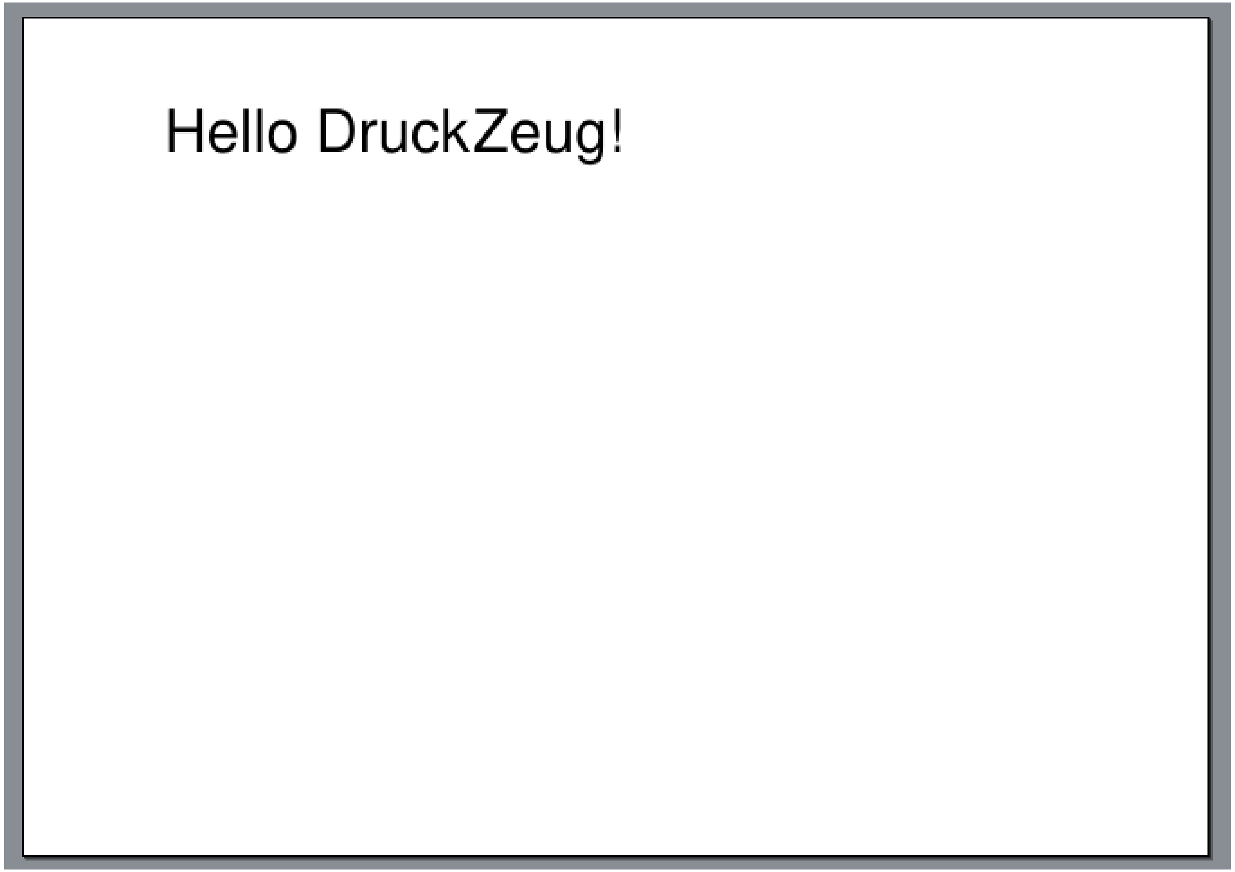


Hello DruckZeug!

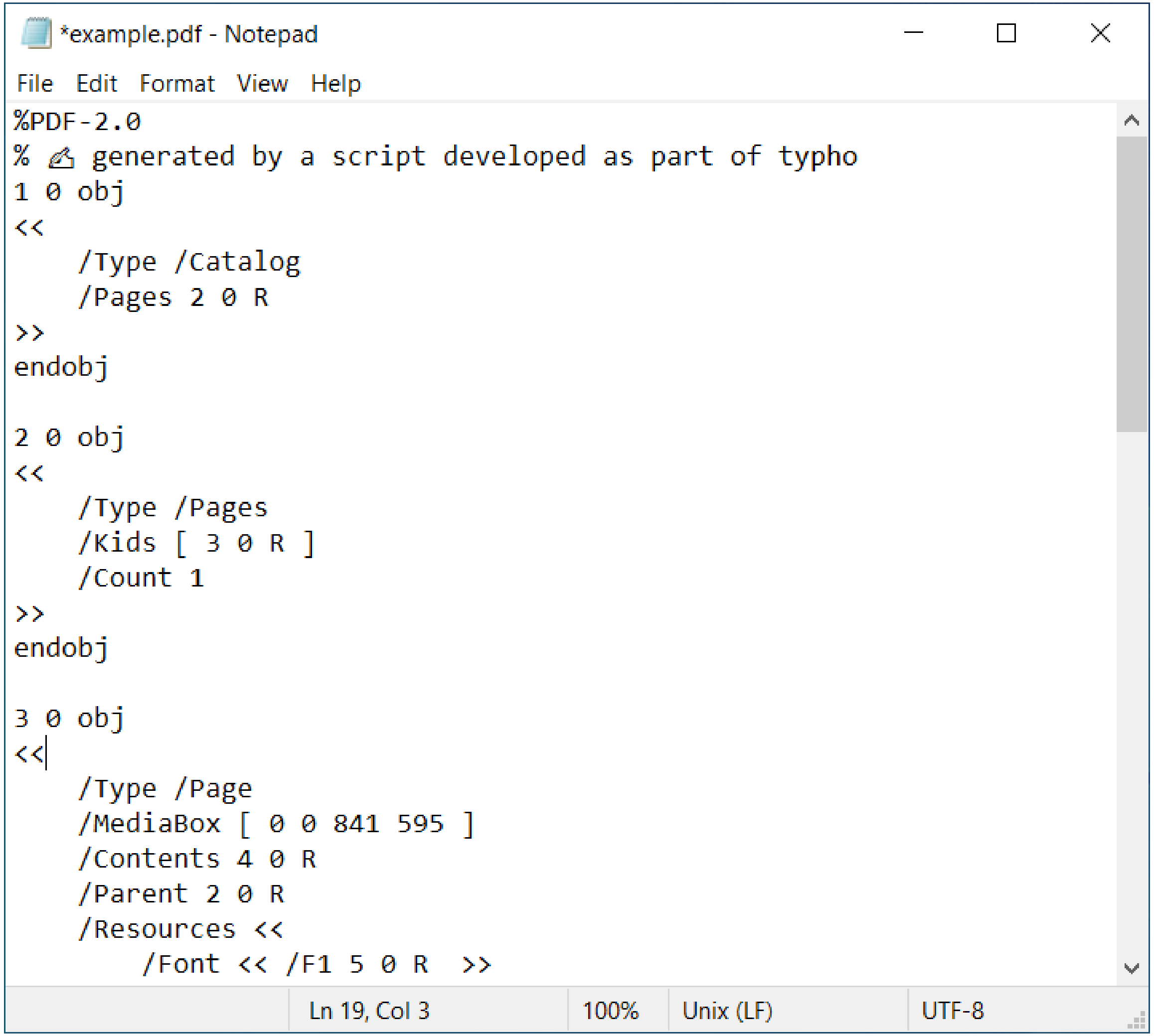
Creating a PDF



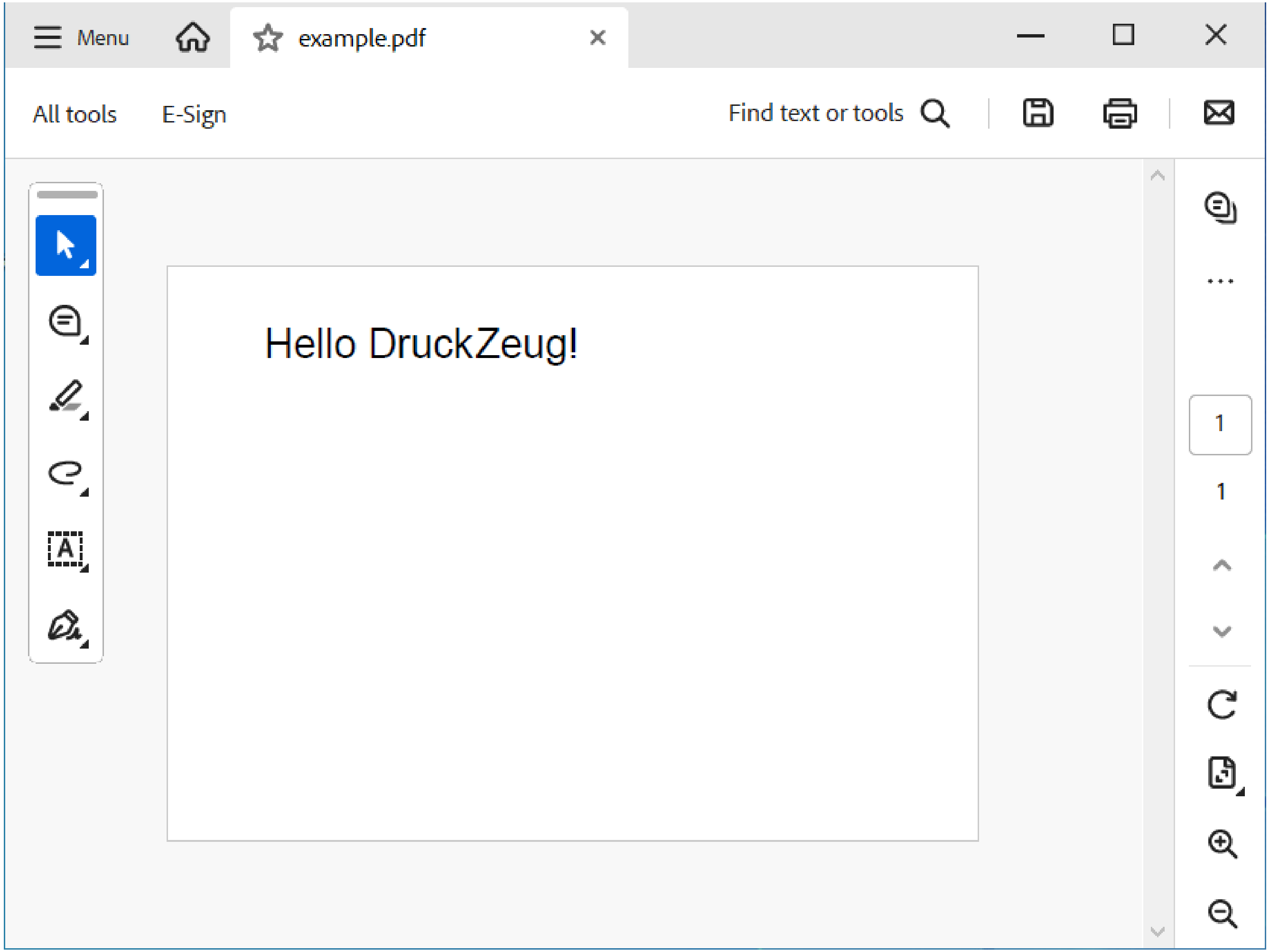
```
4 0 obj
<<
    /Length 63
>>
stream
BT
/F1 42 Tf
100 500 Td
T*
[ (Hello Druck) -50 (Zeug!) ] TJ
ET
endstream
endobj

5 0 obj
<<
    /Type     /Font
    /Subtype  /Type1
    /BaseFont /Helvetica
>>
endobj
```



Hello DruckZeug!



```
%PDF-2.0
% ✍ generated by a script developed as part of typho
1 0 obj
<<
    /Type /Catalog
    /Pages 2 0 R
>>
endobj

2 0 obj
<<
    /Type /Pages
    /Kids [ 3 0 R ]
    /Count 1
>>
endobj

3 0 obj
<<
    /Type /Page
    /MediaBox [ 0 0 841 595 ]
    /Contents 4 0 R
    /Parent 2 0 R
    /Resources <<
        /Font << /F1 5 0 R  >>
```

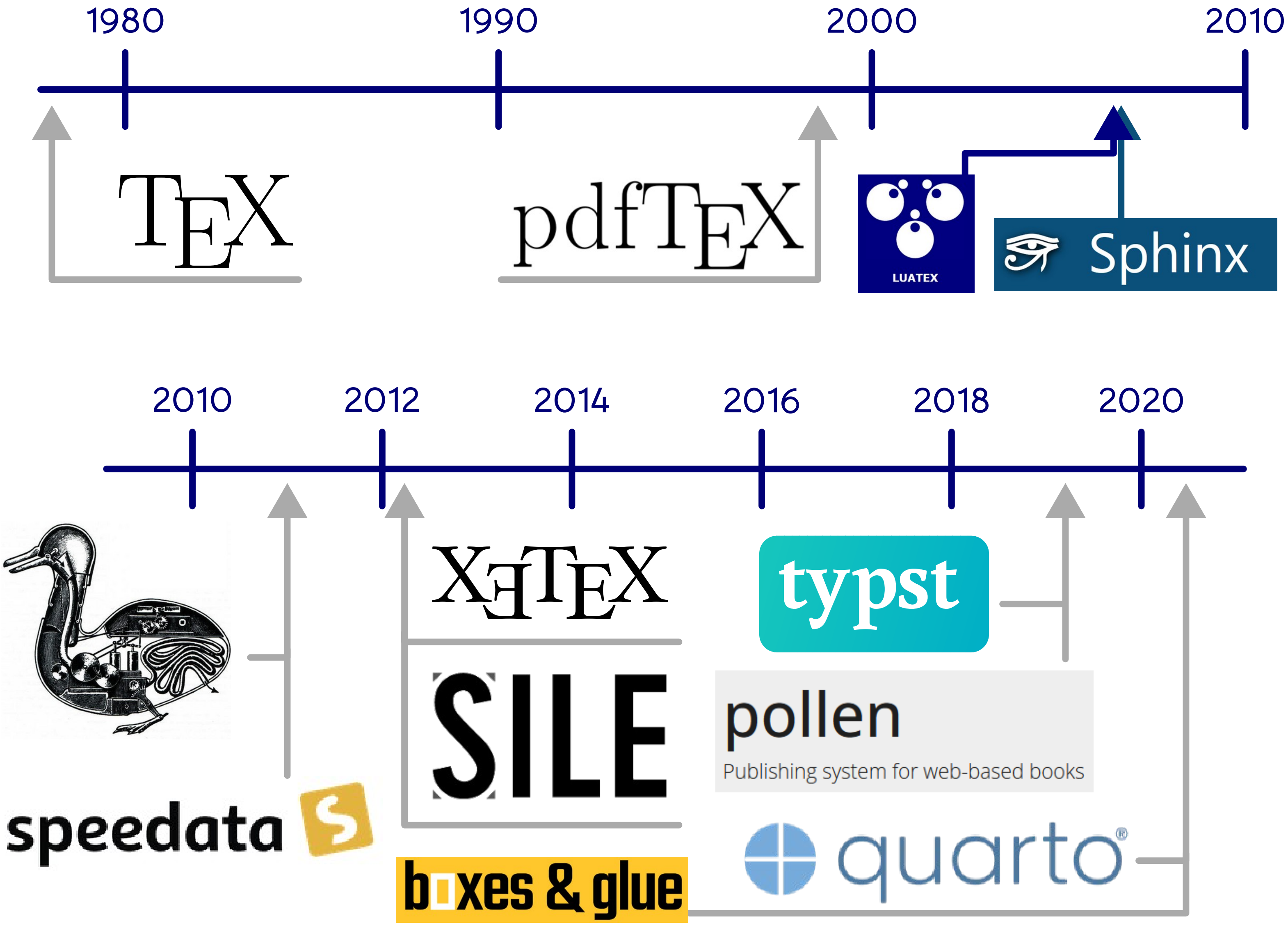


Hello DruckZeug!

Fundamentals – Summary

- Unicode defines what “text” is
  (for all writing systems in use)
- OpenType define what a “font” is
- Given a font and text, you can determine which glyphs you want to place where (“text shaping”)
- You need to put the arrangement into a digital document (e.g. PDF)

Chapter 4.
Status Quo



1980 1990 2000 2010

TeX

pdfTeX

LUATEX

Sphinx

2010 2012 2014 2016 2018 2020

speedata

XeTeX

SILE

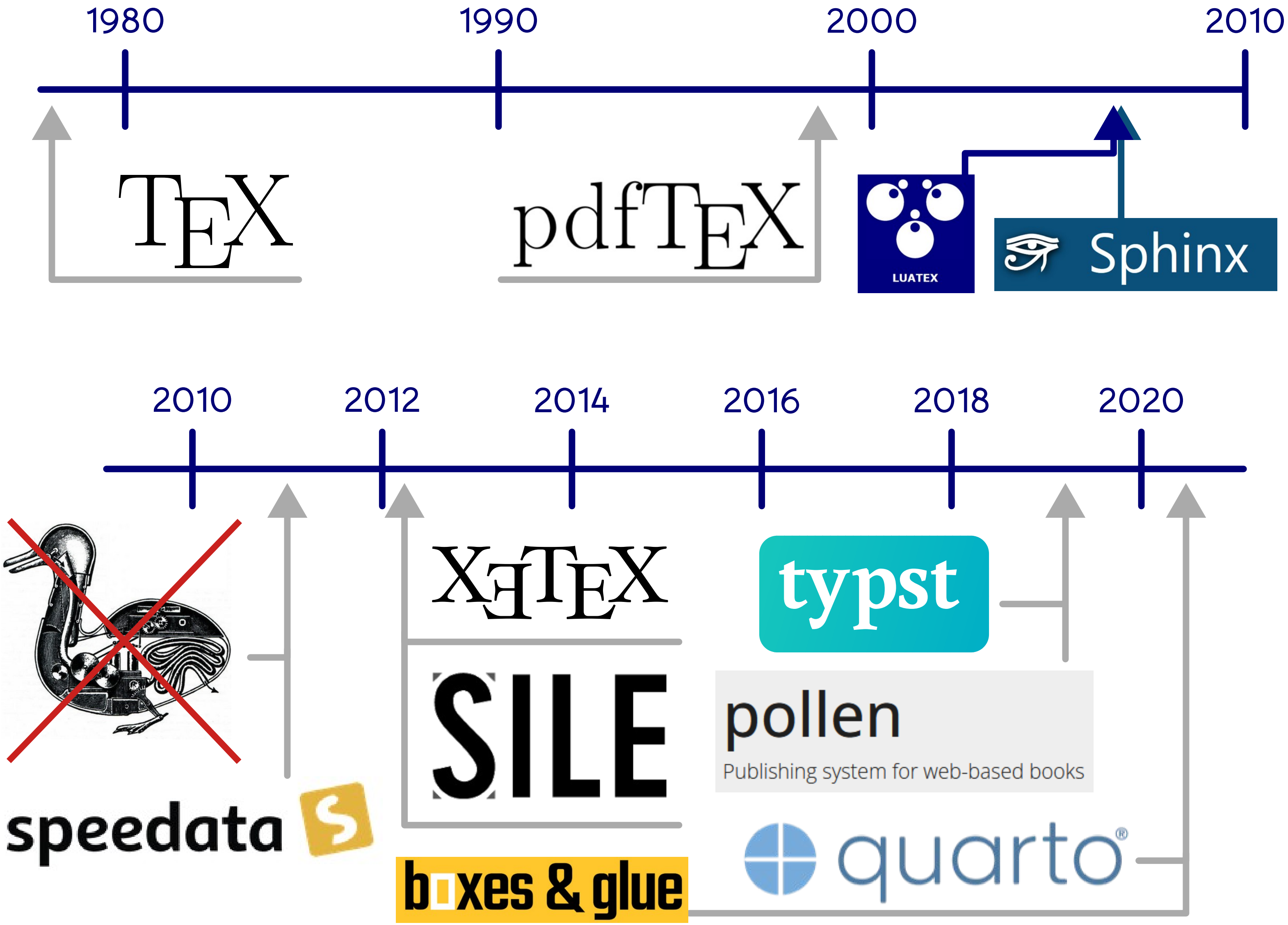
boxes & glue

typst

pollen

Publishing system for web-based books

quarto



1980
1990
2000
2010
TEX
pdfTEX
LUATEX
Sphinx
2010
2012
2014
2016
2018
2020
speedata
XƎTEX
SILE
boxes & glue
typst
pollen
Publishing system for web-based books
quarto®

Chapter 5.
Quo vadis?

Goals

- encode information in an accessible/processable way
- develop the best typesetting engine with excellent Unicode and OpenType support
- provide user-friendly tools to work with text/fonts/documents
- community: find abstraction boundaries!
- community: unite and work together!

Done

- bibparser: parser for .bib files
- opstr: string operations on the CLI
- litua: process trees of markup documents
- arewedigitaltypesettingyet:
  website with an overview over existing projects

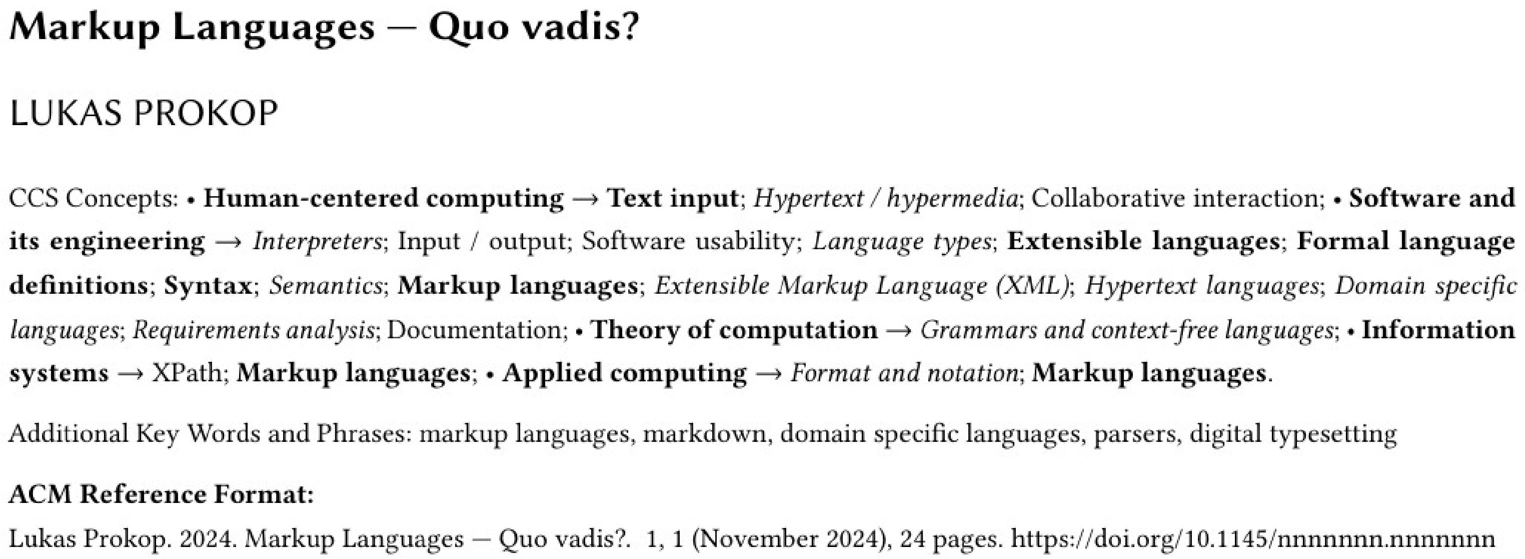
umbrella project: “typho”

Typho – digital typesetting

Digital typesetting for everyone

4 followers https://typho.org/ @tajpulo

Very soon…



Markup Languages — Quo vadis?

LUKAS PROKOP

CCS Concepts: • **Human-centered computing** → **Text input**; *Hypertext / hypermedia*; Collaborative interaction; • **Software and its engineering** → *Interpreters*; Input / output; Software usability; *Language types*; **Extensible languages**; **Formal language definitions**; **Syntax**; *Semantics*; **Markup languages**; *Extensible Markup Language (XML)*; *Hypertext languages*; *Domain specific languages*; *Requirements analysis*; Documentation; • **Theory of computation** → *Grammars and context-free languages*; • **Information systems** → XPath; **Markup languages**; • **Applied computing** → *Format and notation*; **Markup languages**.

Additional Key Words and Phrases: markup languages, markdown, domain specific languages, parsers, digital typesetting

ACM Reference Format:

Lukas Prokop. 2024. Markup Languages — Quo vadis?. 1, 1 (November 2024), 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Thank you

- Do you like my work?
  - github sponsors for typho organization

  - Club „Verein zur Förderung von digitalem Textsatz“

    Empfänger Lukas Prokop
    IBAN AT721912050010921510

Addendum 1. Justification

Thai script

วิวัฒนาการอักษรไทย

ราว พ.ศ. 400 ไทยได้อพยพจากถิ่นเดิมมาตั้งภูมิลำเนาอยู่ใกล้อาณาเขตมอญ ซึ่งกำลังเป็นชาติที่เจริญรุ่งเรืองในสมัยนั้น เริ่มแรกคงเริ่มเลียนแบบตัวอักษรมาจากมอญ ต่อมาราว พ.ศ. 1500 เมื่อขอมขยายอำนาจเข้ามาในดินแดนของคนไทยซึ่งตั้งอยู่บริเวณริมแม่น้ำยม และได้ปกครองเมืองเชลียงและเมืองสุโขทัย ไทยก็เริ่มดัดแปลงอักษรที่มีอยู่เดิมให้คล้ายกับอักษรขอมหวัด

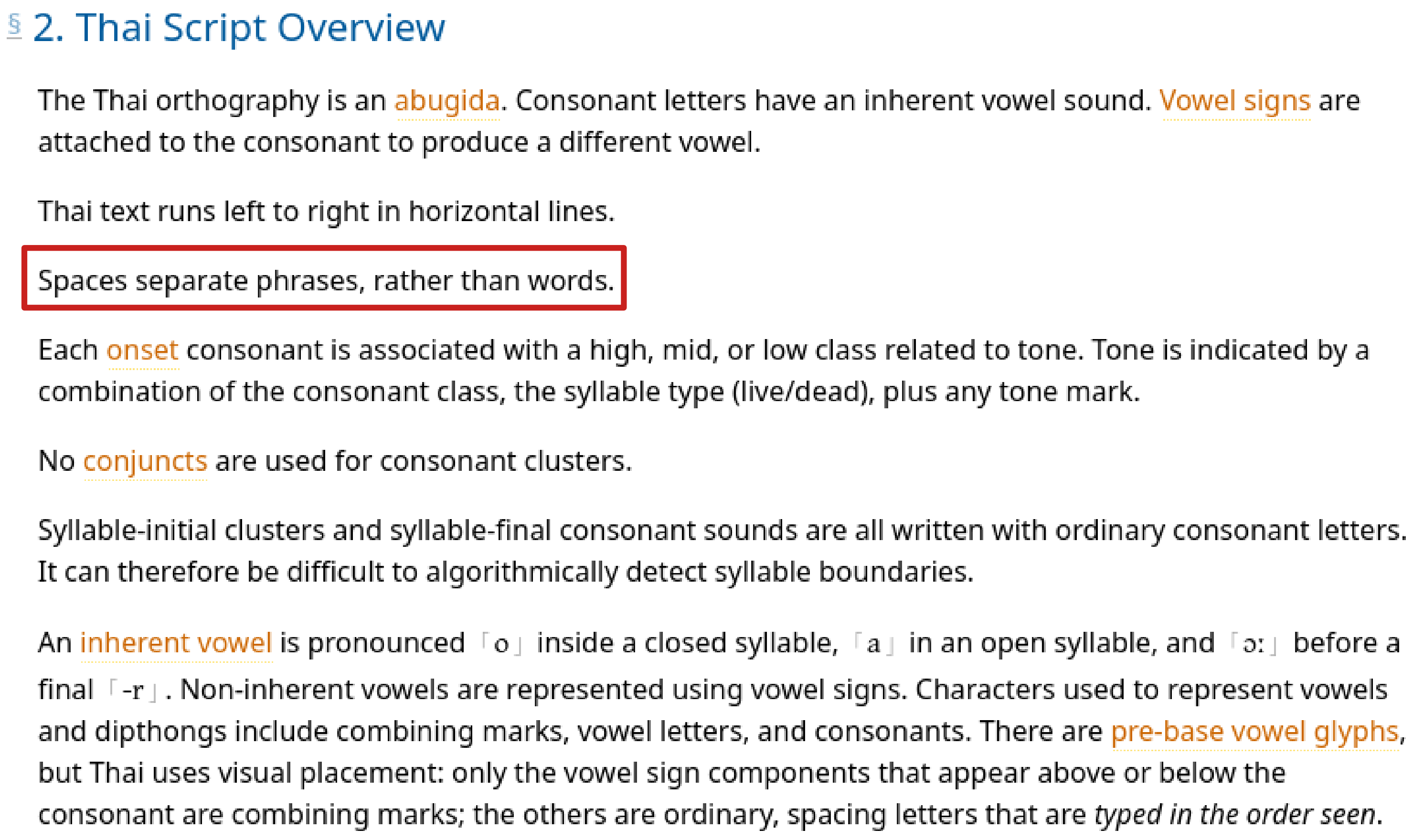
อักษรมอญและอักษรขอมที่ไทยนำมาดัดแปลงใช้นั้นล้วนเป็นอักษรที่รับและแปลงรูปมาจากอักษรพราหมี ของพวกพราหมณ์ซึ่งแพร่หลายในอินเดียตอนเหนือ และอักษรสันสกฤตในสมัยราชวงศ์ปัลลวะ ซึ่งแพร่หลายบริเวณอินเดียตอนใต้ อักษรอินเดียทั้งคู่นี้ต่างก็รับแบบมาจากอักษรฟินิเชียนอีกชั้นหนึ่ง อักษรเฟนีเซียนับได้ว่าเป็นอักษรที่เก่าแก่ที่สุด และเป็นแม่แบบตัวอักษรของชาติต่างๆ ทั้งในเอเชียและยุโรป

ราว พ.ศ. 1826 พ่อขุนรามคำแหงทรงประดิษฐ์อักษรไทยที่เรียกกันว่า ”ลายสือไทย” ขึ้น ซึ่งได้เค้ารูปมาจากอักษรมอญและอักษรขอมที่มีอยู่เดิม ทำให้อักษรไทยมีลักษณะคล้ายคลึงกับอักษรทั้งสอง แม้บางตัวจะไม่คล้ายกัน แต่ก็สามารถรู้ได้ว่าดัดแปลงมาจากอักษรตัวไหน

อักษรไทยมีการปรับปรุงอยู่เรื่อยๆ ในสมัยพญาฦๅไทราว พ.ศ. 1900 มีการแก้ไขตัวอักษรให้ผิดเพี้ยนไปบ้างเล็กน้อย โดยเฉพาะการเพิ่มเชิงที่ตัว ญ ซึ่งใช้ติดต่อเรื่อยมาจนทุกวันนี้ คาดว่าน่าจะเอาอย่างมาจากเขมร ในสมัยสมเด็จพระนารายณ์มหาราช ราว พ.ศ. 2223 ตัวอักษรเริ่มมีทรวดทรงดีขึ้น แต่ก็ไม่ทิ้งเค้าเดิม มีบางตัวเท่านั้นที่แก้ไขผิดไปจากเดิม คือตัว ฎ และ ธ ซึ่งเหมือนกับที่ใช้อยู่ในปัจจุบัน นักวิชาการจำนวนหนึ่งเชื่อว่าในสมัยสมเด็จพระนารายณ์มหาราช ตัวอักษรและการใช้งานมีความคล้ายคลึงกับในปัจจุบันมากที่สุด

by TH Wikipedia

Thai script



§ 2. Thai Script Overview

The Thai orthography is an abugida. Consonant letters have an inherent vowel sound. Vowel signs are attached to the consonant to produce a different vowel.

Thai text runs left to right in horizontal lines.

Spaces separate phrases, rather than words.

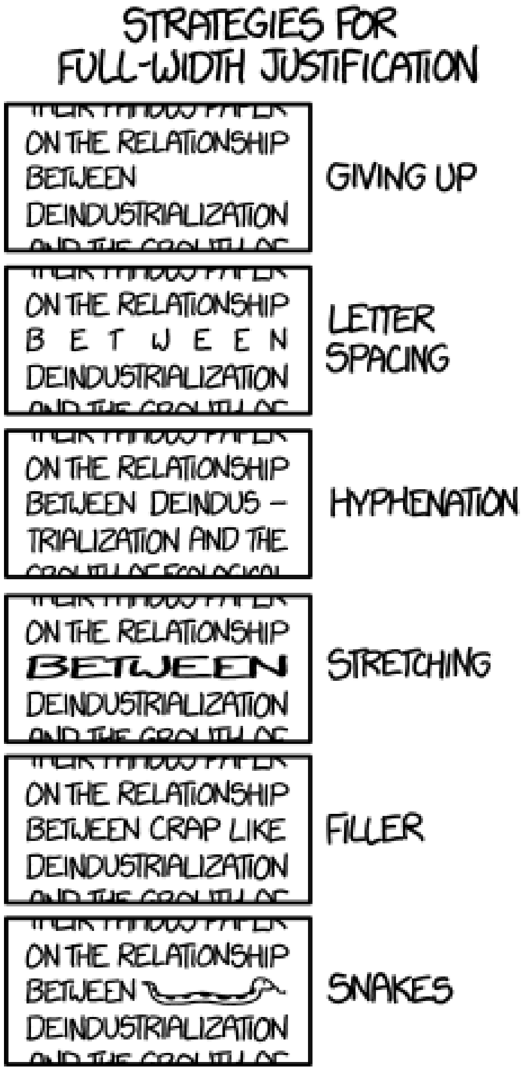
Each onset consonant is associated with a high, mid, or low class related to tone. Tone is indicated by a combination of the consonant class, the syllable type (live/dead), plus any tone mark.

No conjuncts are used for consonant clusters.

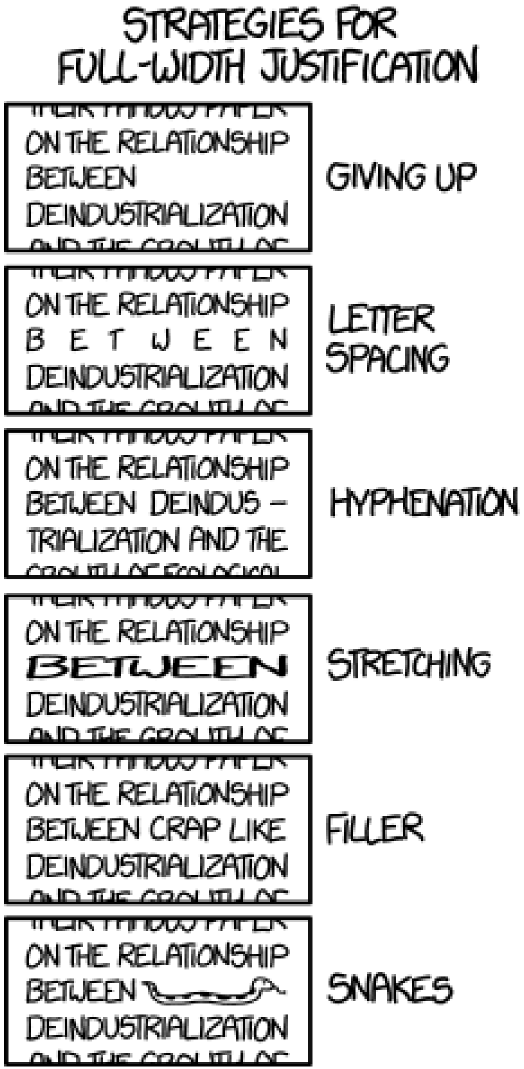
Syllable-initial clusters and syllable-final consonant sounds are all written with ordinary consonant letters. It can therefore be difficult to algorithmically detect syllable boundaries.

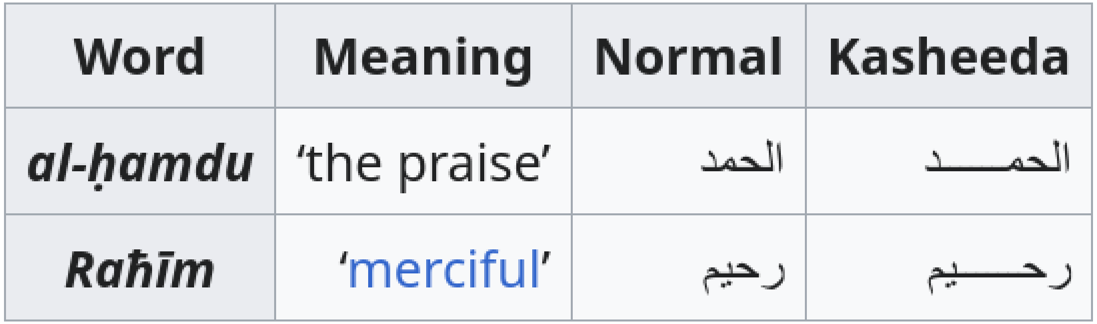
An inherent vowel is pronounced 「o」 inside a closed syllable, 「a」 in an open syllable, and 「ɔː」 before a final 「-r」. Non-inherent vowels are represented using vowel signs. Characters used to represent vowels and dipthongs include combining marks, vowel letters, and consonants. There are pre-base vowel glyphs, but Thai uses visual placement: only the vowel sign components that appear above or below the consonant are combining marks; the others are ordinary, spacing letters that are *typed in the order seen*.

via W3C:
Thai requirements



source: xkcd 1676





Word	Meaning	Normal	Kasheeda
al-ḥamdu	'the praise'	الحمد	الحمـــــد
Raḥīm	'merciful'	رحيم	رحـــــيم

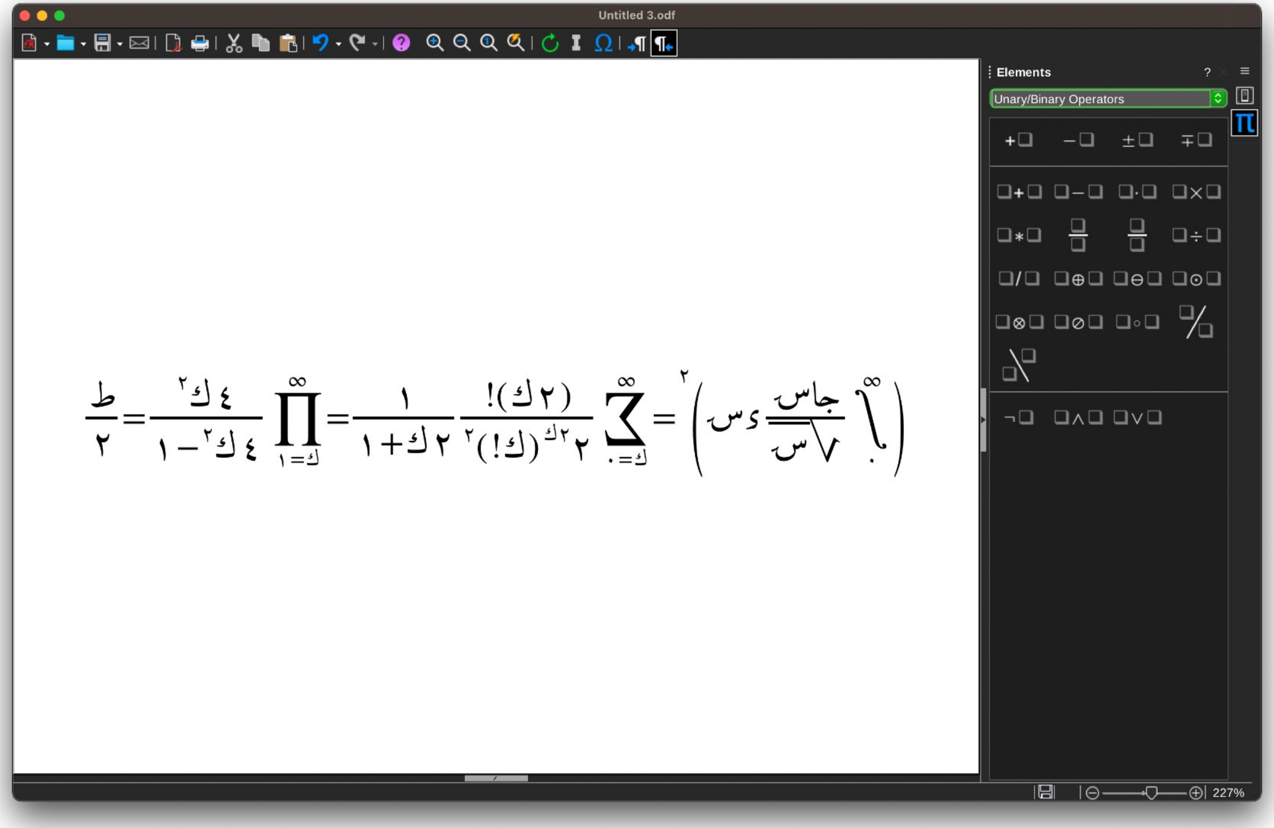
source: EN Wikipedia: Kashida

source: xkcd 1676

Addendum 2.

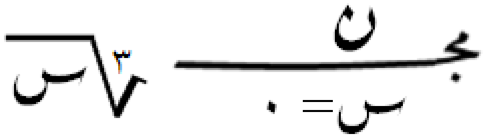
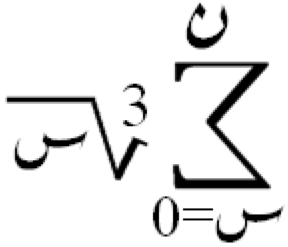
Math

Arabic math notation

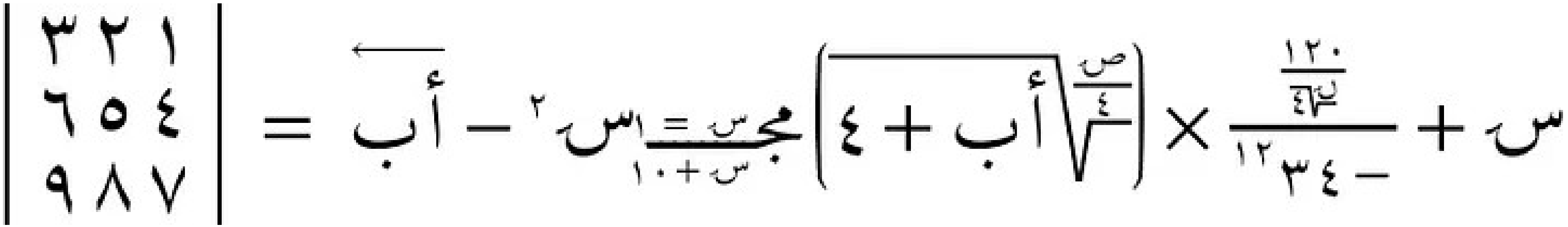


$$\left(\int_{٠}^{∞}\frac{جاس}{\sqrt{س}}\,ءس\right)^{٢}=\sum_{ك=٠}^{∞}\frac{(٢ك)!}{٢^{٢ك}(ك!)^{٢}}\,\frac{١}{٢ك+١}=\prod_{ك=١}^{∞}\frac{٤ك^{٢}}{٤ك^{٢}-١}=\frac{ط}{٢}$$

Arabic math notation

Latin	Arabic	mirrored Latin
$\sum_{x=0}^{n} \sqrt[3]{x}$		

source: Wikipedia: Modern Arabic mathematical notation



source: Khatt.Seen by devghasan