

High Quality Automatic Typesetting

Proposal for a new document model,
typesetting language, and system architecture

Karel Skoupý

Computer Systems Institute
ETH Zürich, Switzerland

High Quality

- the printout should aesthetically please the (discriminating) reader
- result should be comparable to traditional methods
- traditional concern of T_EX users

Automatic Processing

- content + design specification → visual document
- automatization of the whole preparation process
- ability to change the content and/or the design specification
- big gap between abstract specification and formatting control
- more abstract level of control → higher expressivity and productivity
- WYSIWYG works on a very concrete level
- $\text{T}_{\text{E}}\text{X}$ needs a lot of concrete manual tuning for complex layouts

Implementation Limitations of T_EX

- Great difficulty making extensions
 - 20 years old implementation
 - monolithic code
 - obscure dependencies
 - overkill optimizations
 - lack of abstraction
- Poor integrability and reusability
 - no modularization
 - rigid program interface

The $\mathcal{N}\mathcal{T}\mathcal{S}$ Project

- Complete modular redesign of $\text{T}\mathcal{E}\text{X}$
- Reimplementation in Java
- Strict functional identity (compatibility) with $\text{T}\mathcal{E}\text{X}$
- Extendibility and reusability

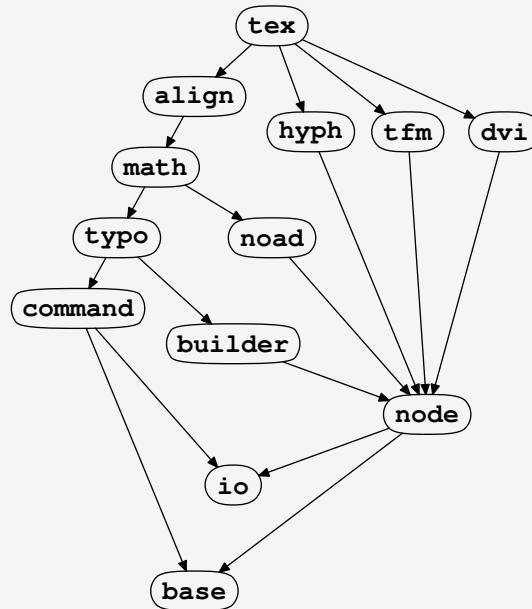


Figure 1 Hierarchy of the $\mathcal{N}\mathcal{T}\mathcal{S}$ packages



Figure 2 \mathcal{N}_{TS} in its Full Speed

Conceptual Limitations

- Rectangular document model
- Static processing semantics
- Low level input/control language
- Monolithic system architecture

Appetizer for the Document Model

The Buddha told Ananda, "All the aspects of everything in the world, such as big and small, inside and outside, amount to the dust before you. Do not say the seeing stretches and shrinks. Consider the example of a square container in which a square of emptiness is seen. I ask you further: is the square emptiness that is seen in the square container a fixed square shape, or is it not fixed as a square shape?
If it is a fixed square shape,

when it is switched to a round container the emptiness would not be round. If it is not a fixed shape, then when it is in the square container, it should not be a square-shaped emptiness. You say you do not know where the meaning lies. The nature of the meaning being thus, how can you speak of its location?
Ananda, if you wished there to be neither squareness nor roundness, you

would only need to remove the container. The essential emptiness has no shape, and so do not say that you would also have to remove the shape from the emptiness.
If, as you suggest, your seeing shrinks and becomes small when you enter a room, then when you look up at the sun shouldn't your seeing be pulled out until it reaches the sun's surface? If walls and eaves can press in and cut off your seeing, then why if you were to drill a small hole, wouldn't there be evidence of the seeing reconnecting? And so that idea is not feasible."

Figure 3 General shape frames in Adobe InDesign

Document Model

- Only rectangular shapes – any graphics is just a box
- Very poor means for specification of non-rectangular text blocks
- Insufficient means for texts along curves

Example of Incorrect Text Wrapping

Various L^AT_EX drawing languages can naturally incorporate texts into drawings because those languages are built on top of L^AT_EX. Some of them support also scaling and rotating of textual objects. The most fancy features are provided by the PStricks package. It is possible to bend a line or even a text block (with proper kerning) along a general curve. However, it is not possible to influence back the text formatting by the curve's shape. $\sum_{i=0}^n a_i$

The information exchange is unidirectional in this case; everything is prepared in advance and passed to DVI specials and the rendering is postponed until printing. In METAPOST, there are two ways to work with text. The first is to use the `infont` operator which makes a picture of a character string using a given PostScript font. However, the glyphs are just put next to each other and without kerning. The second is to call T_EX to perform arbitrary typesetting tasks and the result is then accessible as a fixed picture with a known bounding box. The two ways just described can be even combined making it possible to use individual character glyphs and to let T_EX determine the kerning dimensions (one by one of course). Having collected this information, the METAPOST code can typeset the properly kerned text along a curve. This trick is used in MetaFun and nicely illustrates that people can do just everything with T_EX and METAPOST. Perhaps someone will write METAPOST macros which implement optimal paragraph breaking.

Font sets
no bold

Figure 4 Non-robust L^AT_EX solution for wrapping text

More General Document Model

- introducing paths and graphical facilities (from METAPOST)
- unification of text, font, and graphics objects
- uniform representation of (composite) objects
- natural blending of text and graphics

Example of \blockshape No. 1

The Buddha told Ananda, "All the aspects of everything in the world, such as big and small, inside and outside, amount to the dust before you. Do not say the seeing stretches and shrinks. Consider the example of a square container in which a square of emptiness is seen. I ask you further: is the square emptiness that is seen in the square container a fixed square shape, or is it not fixed as a square shape? If it is a fixed square shape, when it is switched to a round container the emptiness would not be round. If it is not a fixed shape, then when it is in the square container it should not be a square-shaped emptiness. You say you do not know where the meaning lies. The nature of the meaning being thus, how can you speak of its location? Ananda, if you wished there to be neither squareness nor roundness, you would only need to remove the container. The essential emptiness has no shape, and so do not say that you would also have to remove the shape from the emptiness. If, as you suggest, your seeing shrinks and becomes small when you enter a room, then when you look up at the sun shouldn't your seeing be pulled out until it reaches the sun's surface? If walls and eaves can press in and cut off your seeing, then why if you were to drill a small hole, wouldn't there be evidence of the seeing reconnecting? And so that idea is not feasible."

Figure 5 Paragraph shape defined by an orthogonal polygon

Example of \blockshape No. 2

The Buddha told Ananda, "All the aspects of everything in the world, such as big and small, inside and outside, amount to the dust before you. Do not say the seeing stretches and shrinks. Consider the example of a square container in which a square of emptiness is seen. I ask you further: is the square emptiness that is seen in the square container a fixed square shape, or is it not fixed as a square shape? If it is a fixed square shape, when it is switched to a round container the emptiness would not be round. If it is not a fixed shape then when it is in the square container it should not be a square-shaped emptiness. You say you do not know where the meaning lies. The nature of the meaning being thus, how can you speak of its location? Ananda, if you wished there to be neither squareness nor roundness, you would only need to remove the container. The essential emptiness has no shape, and so do not say that you would also have to remove the shape from the emptiness. If, as you suggest, your seeing shrinks and becomes small when you enter a room, then when you look up at the sun shouldn't your seeing be pulled out until it reaches the sun's surface? If walls and eaves can press in and cut off your seeing, then why if you were to drill a small hole, wouldn't there be evidence of the seeing reconnecting? And so that idea is not feasible."

Figure 6 As before but with irregular line heights

Representation and Processing Semantics

- Interdependent input and formatting
- Rigid representation of paragraphs and pages
 - paragraphs are formatted first and never reformatted
 - simple page breaking cannot influence formatting of sub-objects
 - big obstacle for more sophisticated page breaking
 - and for more complex page and column layouts
- broader context optimization limited to paragraphs only

Dynamic Representation and Processing

- separation of input (representation building) and formatting (representation transformation)
- keeping broader dynamic context (whole chapters, documents) allowing global optimization
- information richness of representation
- self-adaptable objects with definable behavior

Input and Control Language

- Primitive macro language with obscure and context dependent syntax and syntactic rules
- Primitive and non-extensible type system
- No provision for modularity
- Unclean specification not separated from the implementation (over-specified)
- T_EX macro-language is powerful for input manipulation but not for object representation (box) manipulation
- incomplete set of primitives (missing `\last*`, `\un*`)

Proposal for Languages

- clean syntax and semantics (with formal specification)
- proper and extendible (definable) type system
- modular: separation of interface and (exchangeable) implementation
- complete: full elegant programmer control, no need for dirty tricks
- regular and basic: providing primitives, not solutions
- open: user constructs as powerful and convenient as primitives
- universal API for different language bindings

- possibly different languages for:
 - input (pre)processing
 - layout specification
 - object representation manipulation

System Architecture

- T_EX architecture is monolithic and not extendible
- T_EX external API non flexible (input, log, output)

Proposed Architecture

- pluggable frontends: T_EX, XML, ...
- pluggable backends: PS, PDF, DVI, plain text content, ...
- pluggable alternative algorithms and policies
- various language bindings: scheme, python, ...
- reusable subsystems (modules)

Conclusion

- T_EX is already (almost) perfect for tasks it was designed for (and many more), no need to improve
- more general document and processing model needed for more challenging tasks
- higher level modular input and control language can increase expressivity and productivity
- open modular architecture can improve flexibility and applicability
- better framework for document/interface design
- it should be easier to express the document design and let the machine to put it into effect

Related Links

- $\mathcal{N}\mathcal{I}\mathcal{S}$ code: <ftp://dante.ctan.org/pub/tex/systems/nts/>
- My papers about $\mathcal{N}\mathcal{I}\mathcal{S}$: <http://www.inf.ethz.ch/~skoupy/papers/>

TOC

- 1 High Quality
- 2 Automatic Processing
- 3 Implementation Limitations of T_EX
- 4 The $\mathcal{N}\mathcal{T}\mathcal{S}$ Project
- 5 Conceptual Limitations
- 6 Appetizer for the Document Model
- 7 Document Model
 - 7.1 Example of Incorrect Text Wrapping
 - 7.2 More General Document Model
 - 7.3 Example of \blockshape No. 1
 - 7.4 Example of \blockshape No. 2
- 8 Representation and Processing Semantics
 - 8.1 Dynamic Representation and Processing
- 9 Input and Control Language
 - 9.1 Proposal for Languages
- 10 System Architecture
 - 10.1 Proposed Architecture
- 11 Conclusion
- 12 Related Links
- 13 TOC