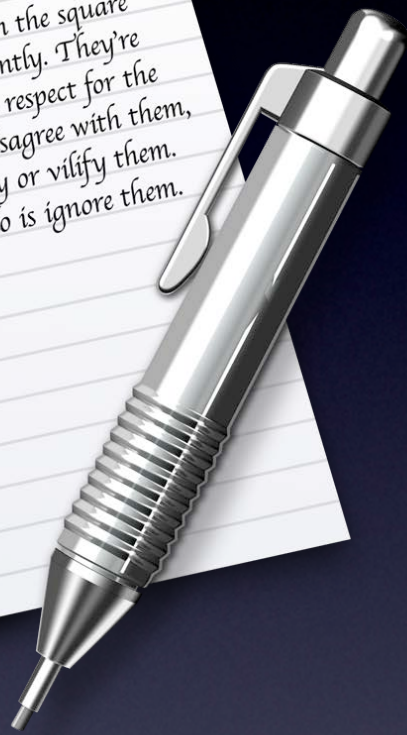


Dear Kate,

Here's to the crazy ones. The misfits. The rebels.
The troublemakers. The round pegs in the square
holes. The ones who see things differently. They're
not fond of rules. And they have no respect for the
status quo. You can praise them, disagree with them,
quote them, disbelieve them, glorify or vilify them.
About the only thing you can't do is ignore them.
Because they change things.

Take Care,
John Appleseed



Programming Documents

Steven R. Bagley

Lectures

Monday 14:00 (C1)

Thursday 11:00 (AMEN-B11)

Friday 10:00 (BSS-A06)

Lectures

Monday 14:00 (C1)

Thursday 11:00 (AMEN-B11)

Friday 10:00 (BSS-A06)

Won't use the Friday lecture every week -- will let you know when we do
Again, let you know when labs start.

Lectures

Monday 14:00 (C1)

Thursday 11:00 (AMEN-B11)

Friday 10:00 (BSS-A06)

Monday 15:00 — 18:00 (CS-B52)

LABS

Won't use the friday lecture every week -- will let you know when we do
Again, let you know when labs start.

Assessment

Coursework (25%)

Exam (75%)

Slides

- Webpage for the module is at:
<http://www.eprg.org/G53DOC>
- Or will be shortly...
- Will put slides there, but you need to take notes as well!
- I will say things that aren't on the slides and might be important

Introduction

- This is not a course in Microsoft Word, or LaTeX etc.
- Although we might study them
- Interested in how we write software that process documents
- And how we model those documents...

process -- everything from display to creation.

Why study Documents?

- Everywhere...
- Complex
- Computer Science in microcosm

Text, colour, images all probably best avoided if you want to keep sane (guess what we are covering)
Also it must look nice too!

What is a Document?

Get them to think about this?

Something written, inscribed, etc.,
which furnishes evidence or
information upon any subject, as a
manuscript, title-deed, tomb-
stone, coin, picture, etc.

OED

Doesn't have to be written, doesn't have to involve text Pictures

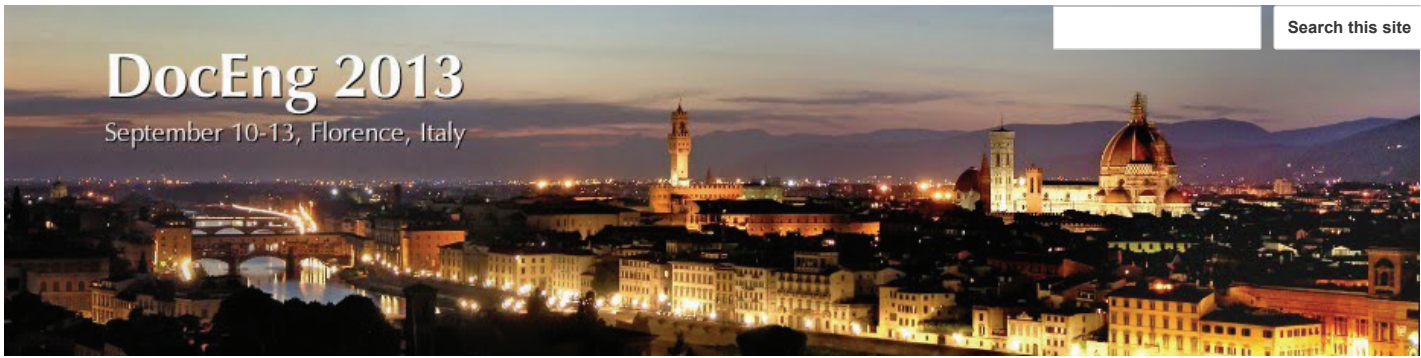
Documents on Computers

- Cover the full gamut from plain text to complex layout
- But also multimedia, video etc. all mixed together
- Crucially, the aim of a document is to communicate information to a human, not the computer...

This is what makes it complex...

What might we want to do with documents?

Get them to think about this?
On a computer



DocEng 2013

September 10-13, Florence, Italy

Search this site

- Home
- Call for Papers
 - Relevant topics
 - Submission procedures
- Committees
 - Conference
 - Programme
 - Steering
- Registration
- Support for students
- Venue
 - Travel and accommodation
- Sitemap

[Call for Papers](#) >

Relevant topics

Topics relevant to the symposium include (but are not limited to):

Modeling and Representation:

- Document representation and standards including interchange standards, markup languages (inc. SGML), style sheets, type representation, multimedia (incl. HTML-5, MPEG, SMIL, SVG), eBook standards (ePub)
- Metadata creation and standards, use of semantic web technologies
- Hypertext/hypermedia, distributed documents, blogs, wikis
- Linking techniques and standards, integration with other digital artifacts

Generation, Manipulation, and Presentation:

- Document authoring tools and systems
- Document presentation (typography, formatting, layout) algorithms and systems (TeX, browsers)
- Automatically generated documents, automated layout and composition, variable data printing
- Adaptive, responsive documents, content customization
- Mobile platforms and documents
- Document transformation and rich-web-client models

Collections, Systems, and Management:

- Collections databases and repositories, storage, indexing, retrieval, versioning, deduplication
- Enterprise content management: models and standards (CMIS), scale and performance, platforms and applications
- Digital libraries and archives, preservation systems
- Document system components: security, APIs (SAX, DOM), versioning, synchronization
- Document systems and workflows

Document Analysis:

- Structure and representation analysis (layout, OCR, visual analysis)
- Linguistic and semantic (content) analysis, categorization, classification

Internationalization:

- Document internationalization, multilingual representations
- Multi-lingual and cross-lingual indexing and search

User Experience:

- Navigation, search
- Usability, accessibility, readability and aesthetics
- Collaborative authoring and editing, curation and annotation
- Workflows, integration and interaction between human and automated processes

Applications:

- Digital humanities, digital preservation and archiving
- eBooks and digital publishing
- and all other application areas

DocEng 2013 topic list -- gives an idea to the breadth of stuff people are doing with documetns

Processing Documents

- Huge variety of things done to documents...
- This course is mainly about how we represent documents in a computer
- And how we can translate between the different representations
- Tend to emphasize documents that can be ‘printed’ in some way

printing also includes eBooks
But will probably look at fonts too

What is a document composed of?

Get them to think about this?

Document Representation

- Can represent documents in many ways...
 - Appearance (how it looks)
 - Structure (logical components)
- Everything in between...

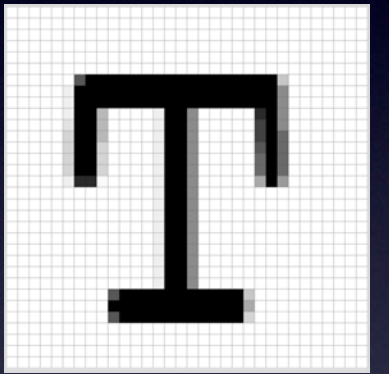
Document Representations

- Can convert between the representations by trading structural information for appearance
- Viewing a document will involve transforming its representation

INFORMATION

Document pipeline — somewhat idealized. Most software processes aren't this clear cut

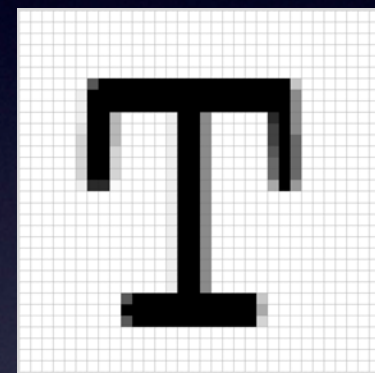
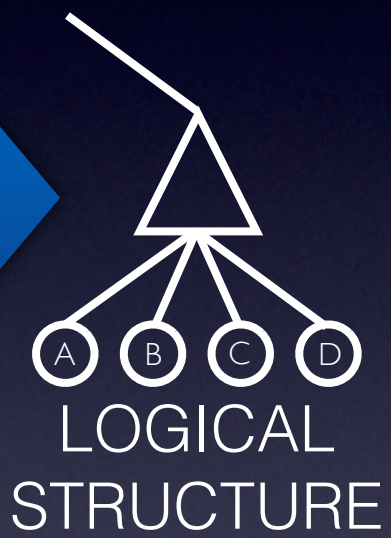
INFORMATION



RASTER

Document pipeline — somewhat idealized. Most software processes aren't this clear cut

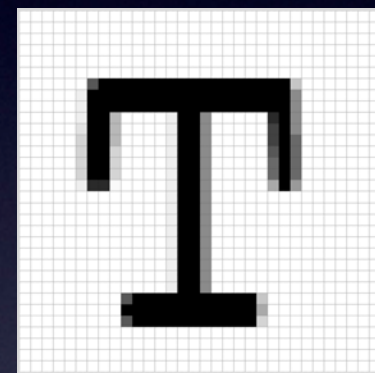
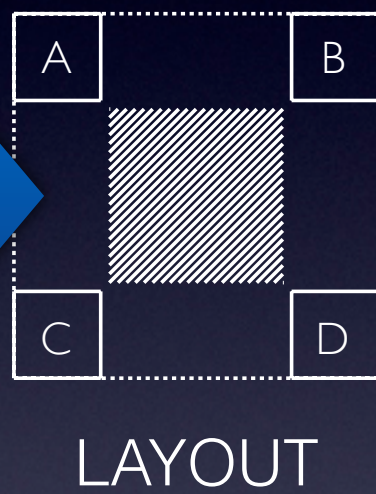
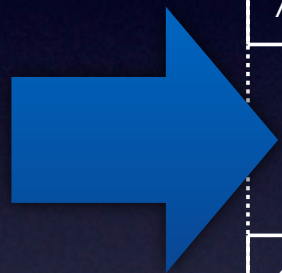
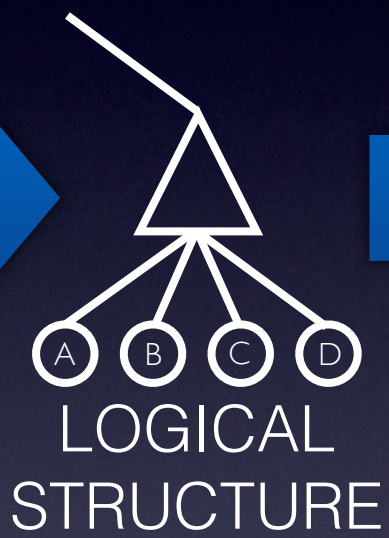
INFORMATION



RASTER

Document pipeline — somewhat idealized. Most software processes aren't this clear cut

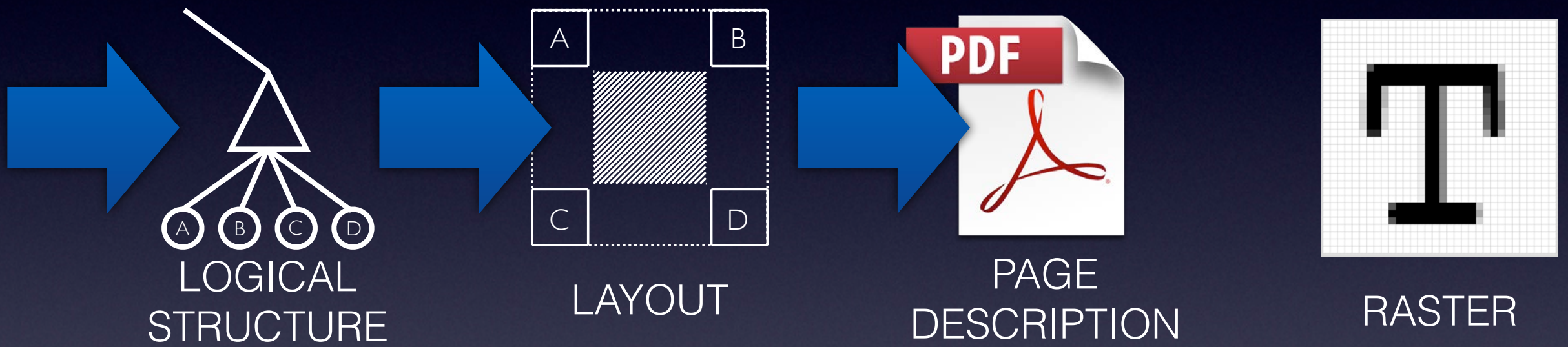
INFORMATION



RASTER

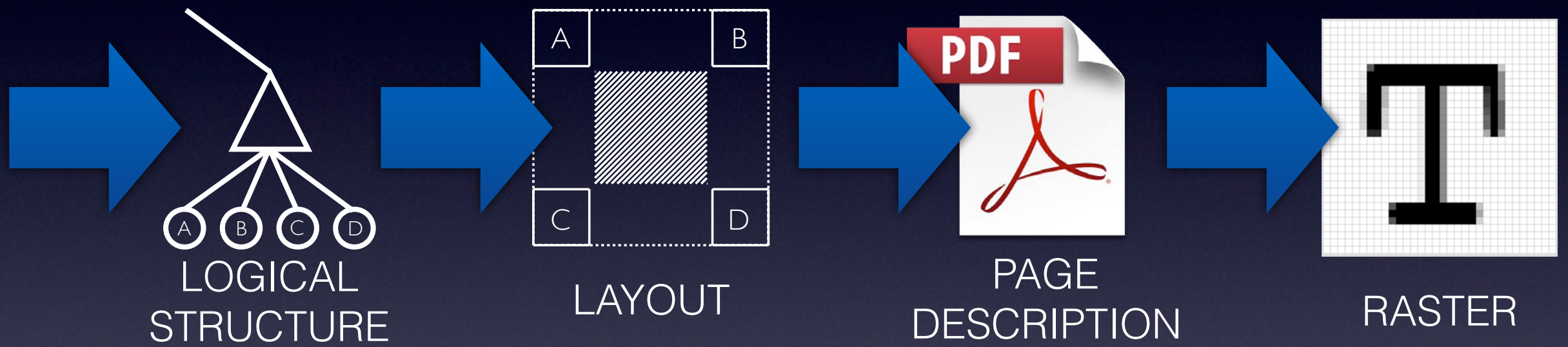
Document pipeline — somewhat idealized. Most software processes aren't this clear cut

INFORMATION



Document pipeline — somewhat idealized. Most software processes aren't this clear cut

INFORMATION



Document pipeline — somewhat idealized. Most software processes aren't this clear cut

Information

- Rather nebulous
- Could be an idea, story, or hard data...

Logical Structure

- Breaks the information into semantic components
- Paragraphs, sections, headings, tables, figures etc.
- Describes how the document is constructed not how it appears
- Examples: DocBook..., FrameMaker etc.

other XML/SGML languages
No pages, styling, etc...

```
<book xml:id="simple_book" xmlns="..." version="5.0">
  <title>Very simple book</title>
  <chapter xml:id="chapter_1">
    <title>Chapter 1</title>
    <para>Hello world!</para>
    <para>I hope that your day is proceeding
<emphasis>splendidly</emphasis>!</para>
  </chapter>
  <chapter xml:id="chapter_2">
    <title>Chapter 2</title>
    <para>Hello again, world!</para>
  </chapter>
</book>
```

Layout

- Starts to describe how the structure appears
- Blocks of text
- Blocks have style properties...
- Layout described in relation to other blocks
- Many ways of describing them (flows, linear constraints, etc)

Examples would include XSL-FO but often these are internal to the software...

```
<?xml version="1.0" encoding="utf-8"?>
<fo:root xmlns:fo="...">
  <fo:layout-master-set>
    <fo:simple-page-master master-name="my-page">
      <fo:region-body margin="1in" />
    </fo:simple-page-master>
  </fo:layout-master-set>
  <fo:page-sequence master-reference="my-page">
    <fo:flow flow-name="xsl-region-body">
      <fo:block>Hello, shyam!</fo:block>
      <fo:block>
        <fo:table>
          ...
        </fo:table>
      </fo:block>
    </fo:flow>
  </fo:page-sequence>
</fo:root>
```

Page Description

- Layout bound to fixed positions on a page (display)
- Blocks of text broken into lines...
- Results is a series of commands
 - Image this glyph at this position
 - Draw a line here...

Note this is the first concept of a page...

0 Tw	(o)Tj	1 0 0 1 471.3	1 0 0 1 116.6	1 0 0 1 292.052	(he)Tj
0 Tc	1 0 0 1 255.32	700.775 Tm	666.775 Tm	666.775 Tm	1 0 0 1 516.963
0 G	700.775 Tm	(and)Tj	(\000)Tj	(help)Tj	666.775 Tm
0 g	(be)Tj	1 0 0 1 490.73	/F0 10 Tf	1 0 0 1 313.671	(pr)Tj
BT	1 0 0 1 268.14	700.775 Tm	1 0 0 1 119.13	666.775 Tm	1 0 0 1 526.833
/F0 10 Tf	700.775 Tm	(e)Tj	666.775 Tm	(y)Tj	666.775 Tm
1 0 0 1 110	(able)Tj	1 0 0 1 495.44	(s)Tj	1 0 0 1 318.901	(ocess)Tj
700.775 Tm	1 0 0 1 288.87	700.775 Tm	1 0 0 1 125.459	666.775 Tm	1 0 0 1 110
(T)Tj	700.775 Tm	(xpr)Tj	666.775 Tm	(ou)Tj	654.775 Tm
1 0 0 1 115.15	(t)Tj	1 0 0 1 510.51	(t)Tj	1 0 0 1 332.48	(of)Tj
700.775 Tm	1 0 0 1 292.01	700.775 Tm	1 0 0 1 128.849	666.775 Tm	1 0 0 1 121.095
(o)Tj	700.775 Tm	(ess)Tj	666.775 Tm	(do)Tj	654.775 Tm
1 0 0 1 123.11	(o)Tj	1 0 0 1 526.28	(he)Tj	1 0 0 1 346.139	(in)Tj
700.775 Tm	1 0 0 1 299.97	700.775 Tm	1 0 0 1 141.548	666.775 Tm	1 0 0 1 129.485
(in)Tj	700.775 Tm	(t)Tj	666.775 Tm	(t)Tj	654.775 Tm
1 0 0 1 131.5	(cap)Tj	1 0 0 1 529.67	(job)Tj	1 0 0 1 349.529	(v)Tj
700.775 Tm	1 0 0 1 315.23	700.775 Tm	1 0 0 1 156.967	666.775 Tm	1 0 0 1 134.715
(v)Tj	700.775 Tm	(hem)Tj	666.775 Tm	(his.)Tj	654.775 Tm
1 0 0 1 136.73	(tur)Tj	1 0 0 1 110	(of)Tj	1 0 0 1 367.088	(ention)Tj
700.775 Tm	1 0 0 1 328.42	688.775 Tm	1 0 0 1 167.846	666.775 Tm	1 0 0 1 165.08
(ent)Tj	700.775 Tm	(in)Tj	666.775 Tm	(The)Tj	654.775 Tm
1 0 0 1 153.1	(e)Tj	1 0 0 1 121.23	(a)Tj	1 0 0 1 385.917	(and)Tj
700.775 Tm	1 0 0 1 335.71	688.775 Tm	1 0 0 1 174.935	666.775 Tm	1 0 0 1 184.315
(pr)Tj	700.775 Tm	(t)Tj	666.775 Tm	(languag)Tj	654.775 Tm
1 0 0 1 162.97	(t)Tj	1 0 0 1 124.62	(pr)Tj	1 0 0 1 421.607	(design)Tj
700.775 Tm	1 0 0 1 339.1	688.775 Tm	1 0 0 1 184.805	666.775 Tm	1 0 0 1 216.05
(og)Tj	700.775 Tm	(he)Tj	666.775 Tm	(e)Tj	654.775 Tm
1 0 0 1 174.13	(he)Tj	1 0 0 1 137.73	(og)Tj	1 0 0 1 428.486	(b)Tj
700.775 Tm	1 0 0 1 352.21	688.775 Tm	1 0 0 1 195.965	666.775 Tm	1 0 0 1 221.27
(r)Tj	700.775 Tm	(pr)Tj	666.775 Tm	(should)Tj	654.775 Tm
1 0 0 1 177.92	(same)Tj	1 0 0 1 147.6	(r)Tj	1 0 0 1 461.145	(y)Tj
700.775 Tm	1 0 0 1 377.57	688.775 Tm	1 0 0 1 199.755	666.775 Tm	1 0 0 1 229.135
(ams,)Tj	700.775 Tm	(og)Tj	666.775 Tm	(f)Tj	654.775 Tm
1 0 0 1 200.99	(kinds)Tj	1 0 0 1 158.76	(amming)Tj	1 0 0 1 464.365	(le)Tj
700.775 Tm	1 0 0 1 404.71	688.775 Tm	1 0 0 1 238.794	666.775 Tm	1 0 0 1 236.805
(y)Tj	700.775 Tm	(r)Tj	666.775 Tm	(acilit)Tj	654.775 Tm
1 0 0 1 206.22	(of)Tj	1 0 0 1 162.55	(languag)Tj	1 0 0 1 485.825	(tting)Tj
700.775 Tm	1 0 0 1 416	688.775 Tm	1 0 0 1 274.484	666.775 Tm	1 0 0 1 259.92
(ou)Tj	700.775 Tm	(am)Tj	666.775 Tm	(at)Tj	654.775 Tm
1 0 0 1 220.21	(abs)Tj	1 0 0 1 178.88	(e)Tj	1 0 0 1 493.995	(y)Tj
700.775 Tm	1 0 0 1 430.62	688.775 Tm	1 0 0 1 281.363	666.775 Tm	1 0 0 1 265.15
(need)Tj	700.775 Tm	(design.)Tj	666.775 Tm	(e)Tj	654.775 Tm
1 0 0 1 244.22	(tr)Tj	1 0 0 1 110	(t)Tj	1 0 0 1 500.874	(ou)Tj
700.775 Tm	1 0 0 1 437.67	666.775 Tm	1 0 0 1 284.503	666.775 Tm	1 0 0 1 278.945
(t)Tj	700.775 Tm	(It)Tj	666.775 Tm	(t)Tj	654.775 Tm
1 0 0 1 247.36	(actions)Tj	/F2 10 Tf	(o)Tj	1 0 0 1 504.264	(encode)Tj
700.775 Tm				666.775 Tm	1 0 0 1 312.66

Page Description

- Not necessarily any containment...
- Order of commands doesn't matter if final output is OK
- This can have problems with accessibility since at this level it's difficult to know what things are

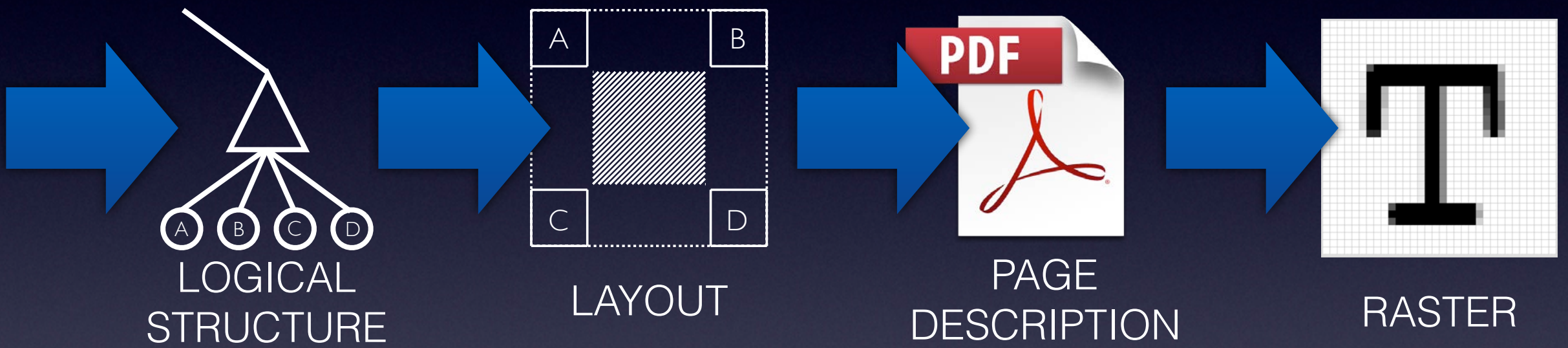
Raster

- Converts the page description into pixels (rasterization)
- Font Rendering, Line drawing...
- Might be only 1-bit display, so will need to halftone (dither) to create greyscale

Idealized pipeline

- This is the idealized pipeline
- Most software blurs the boundaries between them, or skips certain sections altogether

INFORMATION



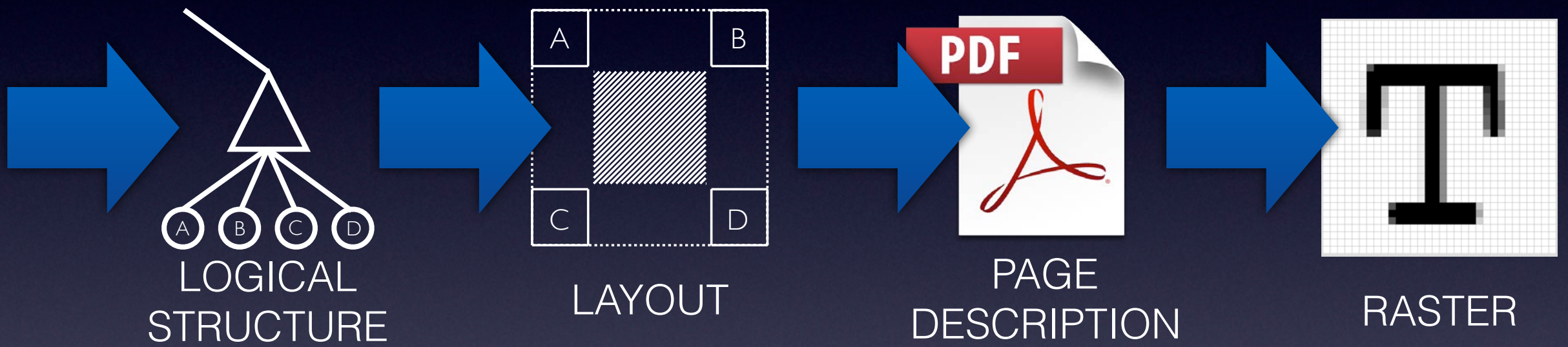
Document Processing

- Think about how you'd generate a table of contents...
- Need to know both what are headings (structural)
- But also what page they are on (PDL)

Untransforming

- Easy to go from structure to appearance
- Much harder to go from appearance to structure
- Need to know unpick the meaning from generated code

INFORMATION



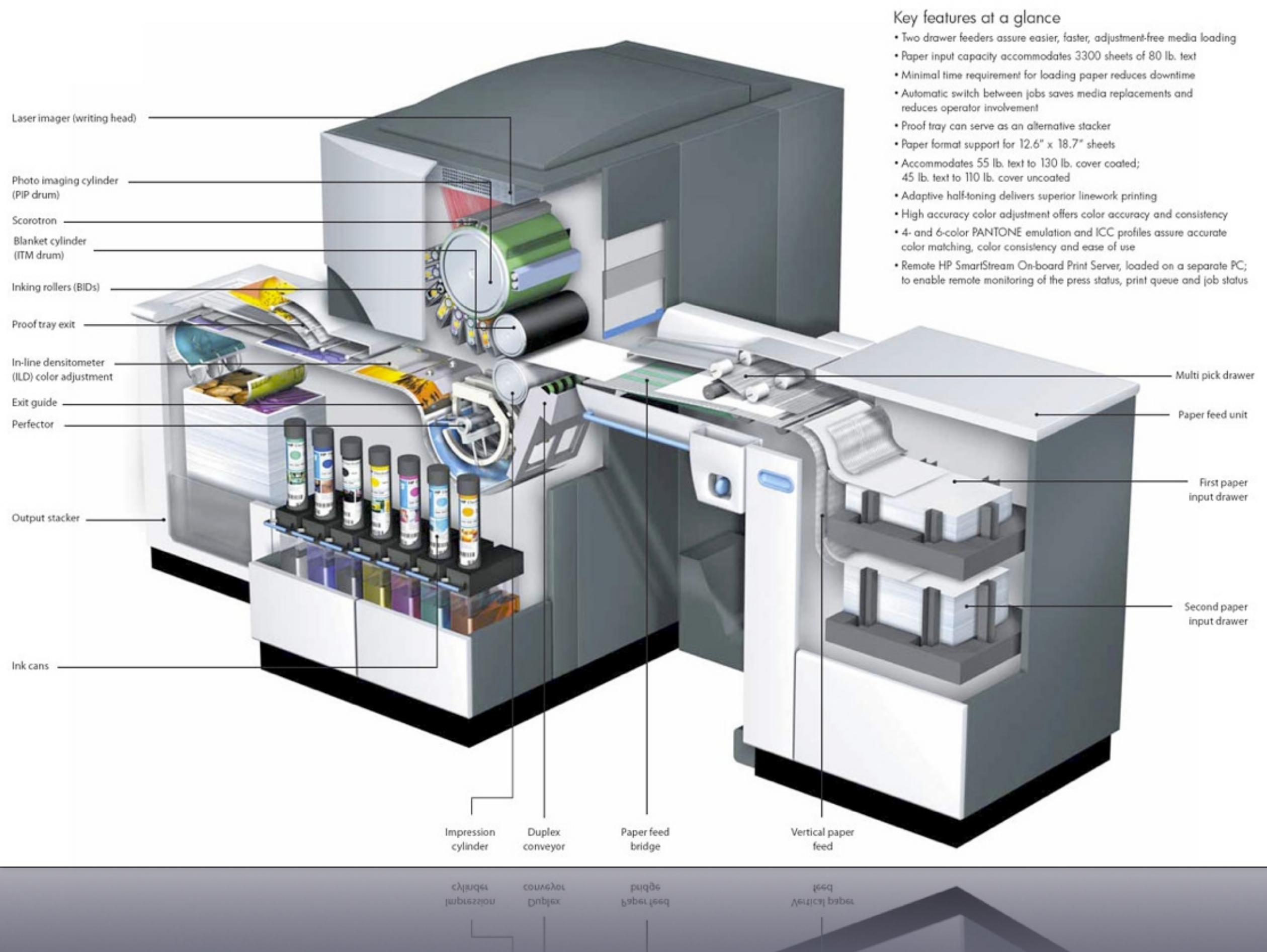
What makes something a heading?

VDP and eBooks

- In the past decade, the document landscape has shifted significantly
- Variable Data Printing and eBooks require us to 'look again' at how we process and represent documents

Variable Data Printing

- Variable Data Printing has been made possible by the advent of digital printing presses
- Rather than using a fixed plate containing the document
- Generated on the fly by a laser



HP Indigo printer...

Variable Data Printing

- Means every page can be different and still keep the press running at high speed (up to 240ppm)
- Great for personalizing documents (junk mail)
- But need to generate these documents as the print is printing

Variable Data Printing

- Think about the numbers
- A4 page is 210mm x 297mm
- At 1219dpi, that's 137MB per page per colour
- At four colours per page, that's ~548MB of data that needs to be fed to the printer every half second

in full colour, its 120ppm, or a page every half second

Variable Data Printing

- Personalized print often works by binding in data from a database into a template
- Means we need to run through each stage of the pipeline to generate the raster
- Better have fast code...
- Or find a better way to represent it

eBooks

- eBooks have a different but related problem
- One document but designed to be viewed on a number of different devices
- All with different display sizes
- And various spec machines

VDP and eBooks

- Both VDP and eBooks are requiring us to think again about how we represent and transform documents
- The problem of Documents on computers is not solved...

The rest of the course

- Look at how we can represent the document in a computer (Data structures)
- And the algorithms to transform them through the different stages
- The big problem is that documents have to look nice and that's hard...

Topics...

- Text
 - Algorithms for processing
 - Formatting, Justification and Hyphenation
- Page Description Models
- Fonts
- Logical Structure

In no particular order

Topics...

- Common Document Formats and Document Software
- Automated Layout
- VDP
- eBooks
- Multimedia Documents

In no particular order