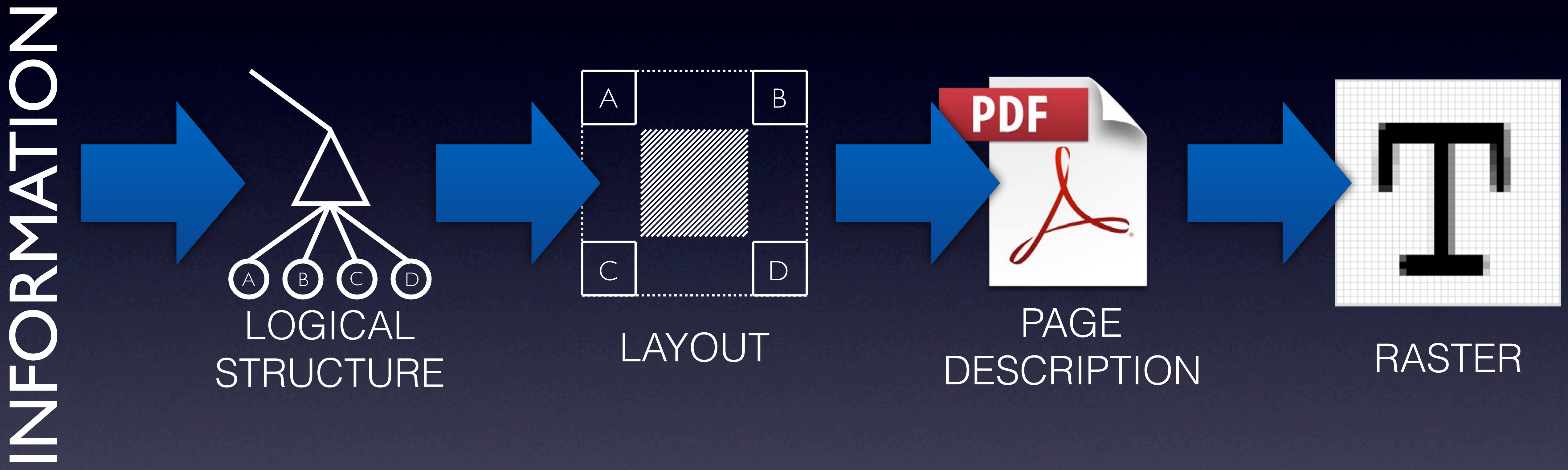
Electronic Books

Steven R. Bagley



INFORMATION

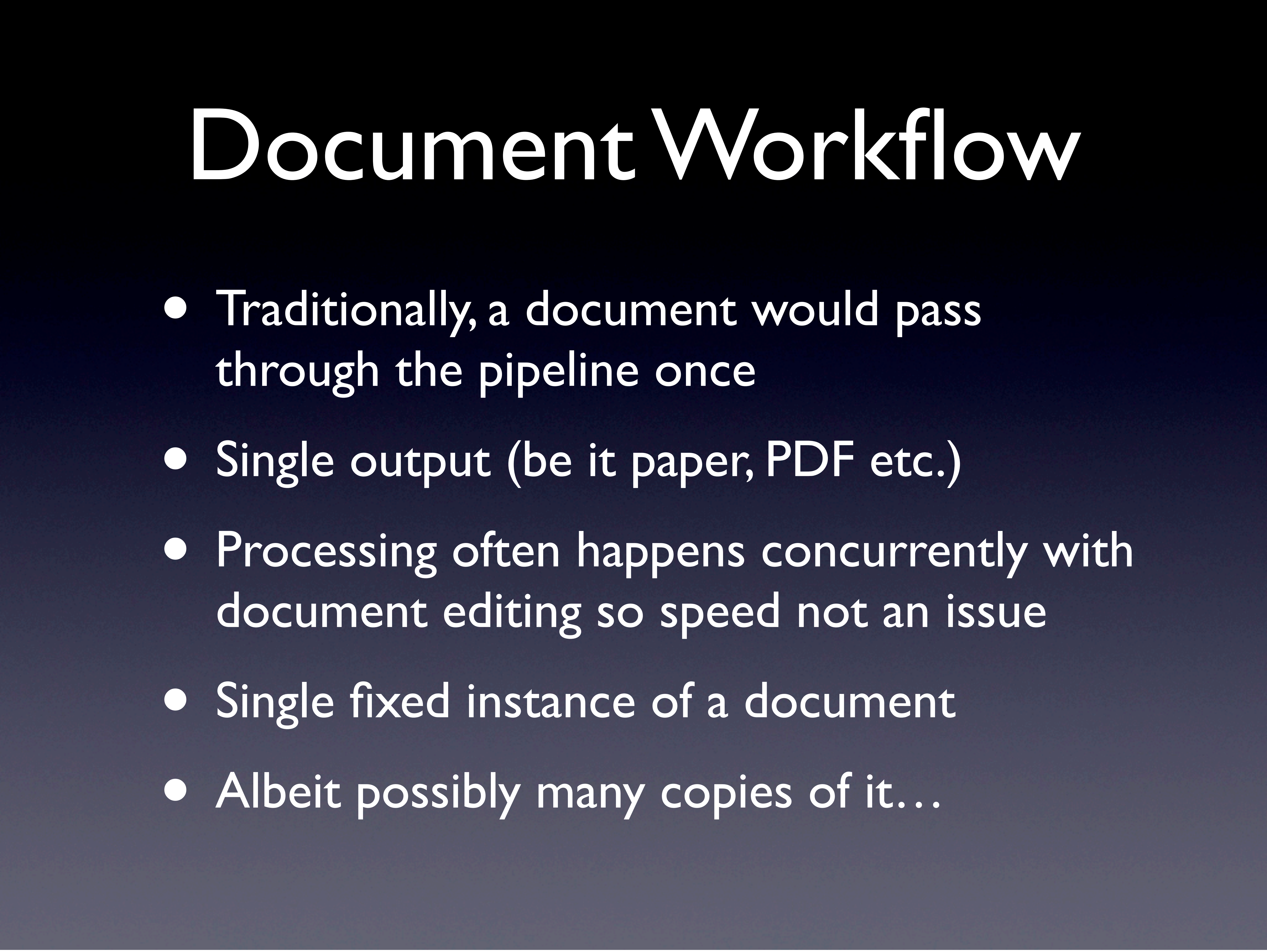
LOGICAL STRUCTURE

LAYOUT

PAGE DESCRIPTION

RASTER

Started off looking at this, and so far everything we’ve looked at fits neatly into this pipeline
Idealized document pipeline, in reality the result is different.



Document Workflow

- Traditionally, a document would pass through the pipeline once
- Single output (be it paper, PDF etc.)
- Processing often happens concurrently with document editing so speed not an issue
- Single fixed instance of a document
- Albeit possibly many copies of it…

PRoviding it is faster than the user thinks…

Digital Documents

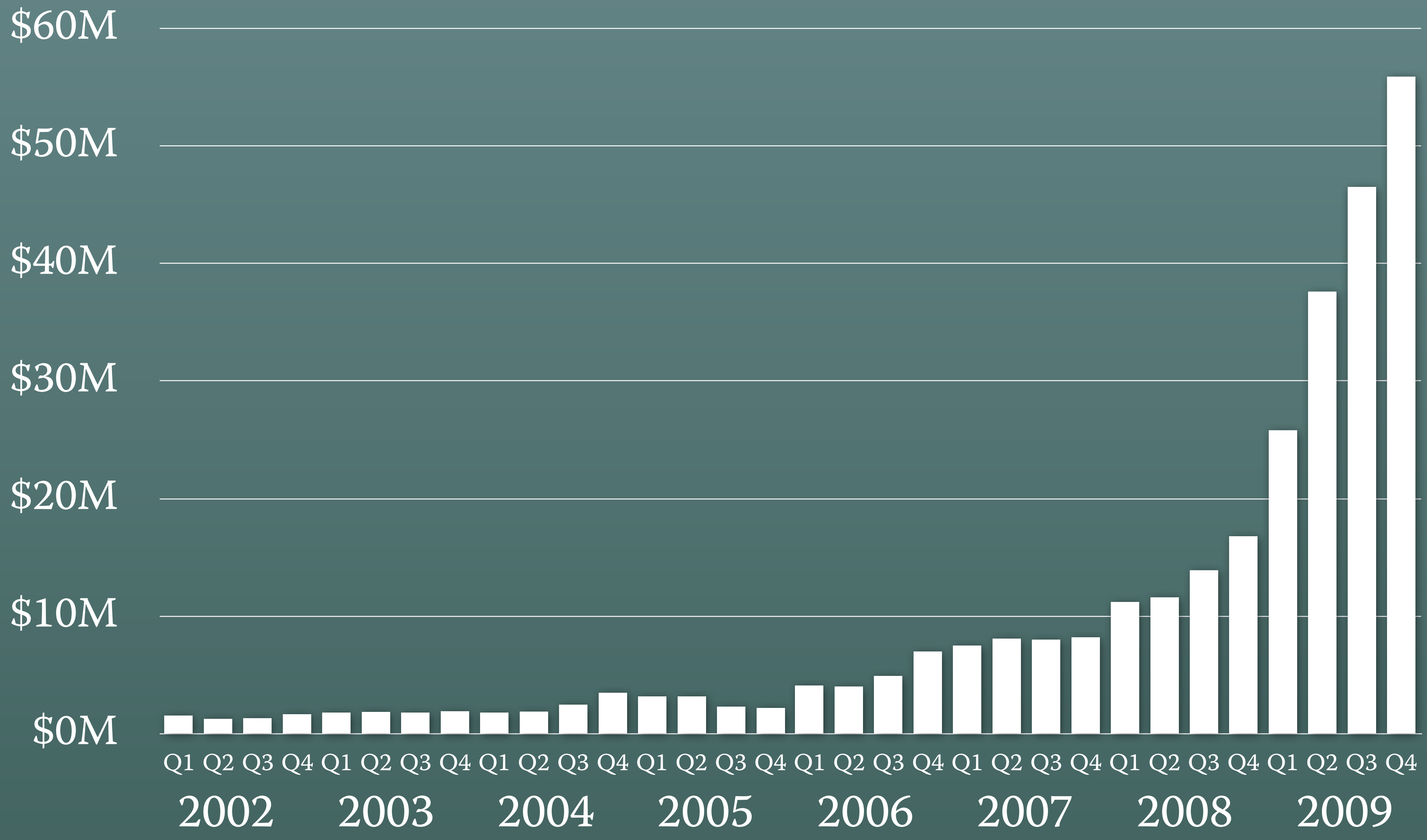
- The last 15 years has seen a change from fixed-form documents
- Personalized Print
  - Changing the Content of the document to suit the reader
- Electronic Books
  - Changing the display of the document

eBooks

- Explosion in popularity
- 2009 saw Amazon sell more eBooks than physical books on Christmas Day
- In 2010, Amazon was selling more than 180 eBooks for every 100 hardback books
- Now selling more EBooks than books…
- Exponentially increasing sales…

not including free ebooks

US Trade Wholesale Electronic Book Sales (USD)



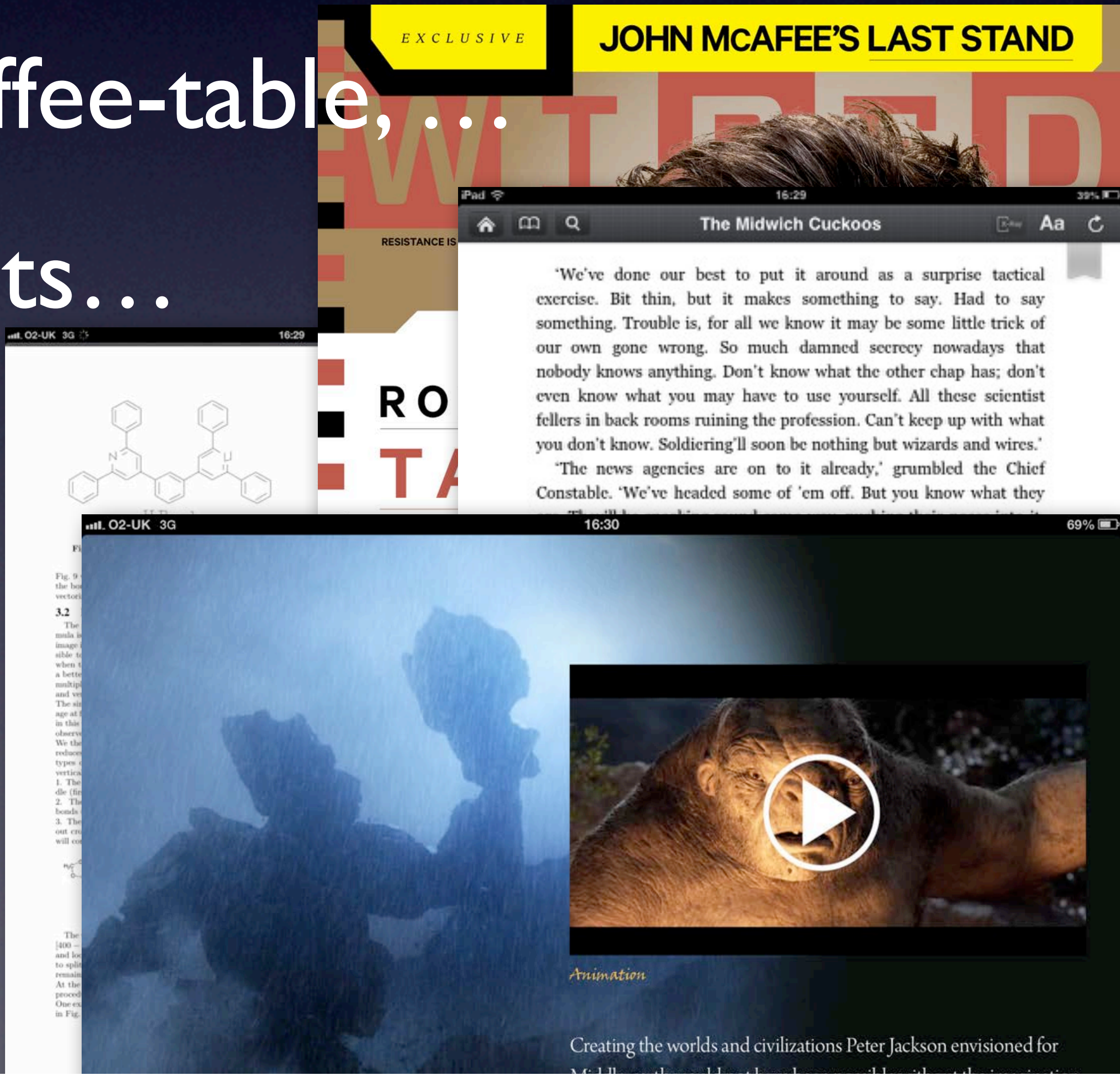
Source: International Digital Publishing Forum (IDPF) / Association of American Publishers (AAP)

> We’re now seeing the transition we’ve been expecting. After five years, ebooks is a multi-billion dollar category for us and growing fast — up approximately 70 percent last year. In contrast, our physical book sales experienced the lowest December growth rate in our 17 years as a book seller, up just 5 percent.”
>
> *Jeff Bezos, Amazon.com, 2013*

The company didn’t share actual ebook numbers, but North American media revenues were $2.9 billion for the quarter and $9.19 billion for the year — up 13 percent and 15 percent, respectively, over 2011. International media revenues were $3.6 billion for the quarter and $10.75 billion for the year.

eBooks

- Different classes of Books
  - Novels, Technical, Coffee-table, …
  - Different requirements…
- Newspapers
- Magazines



'We've done our best to put it around as a surprise tactical exercise. Bit thin, but it makes something to say. Had to say something. Trouble is, for all we know it may be some little trick of our own gone wrong. So much damned secrecy nowadays that nobody knows anything. Don't know what the other chap has; don't even know what you may have to use yourself. All these scientist fellers in back rooms ruining the profession. Can't keep up with what you don't know. Soldiering'll soon be nothing but wizards and wires.'

'The news agencies are on to it already,' grumbled the Chief Constable. 'We've headed some of 'em off. But you know what they

Creating the worlds and civilizations Peter Jackson envisioned for

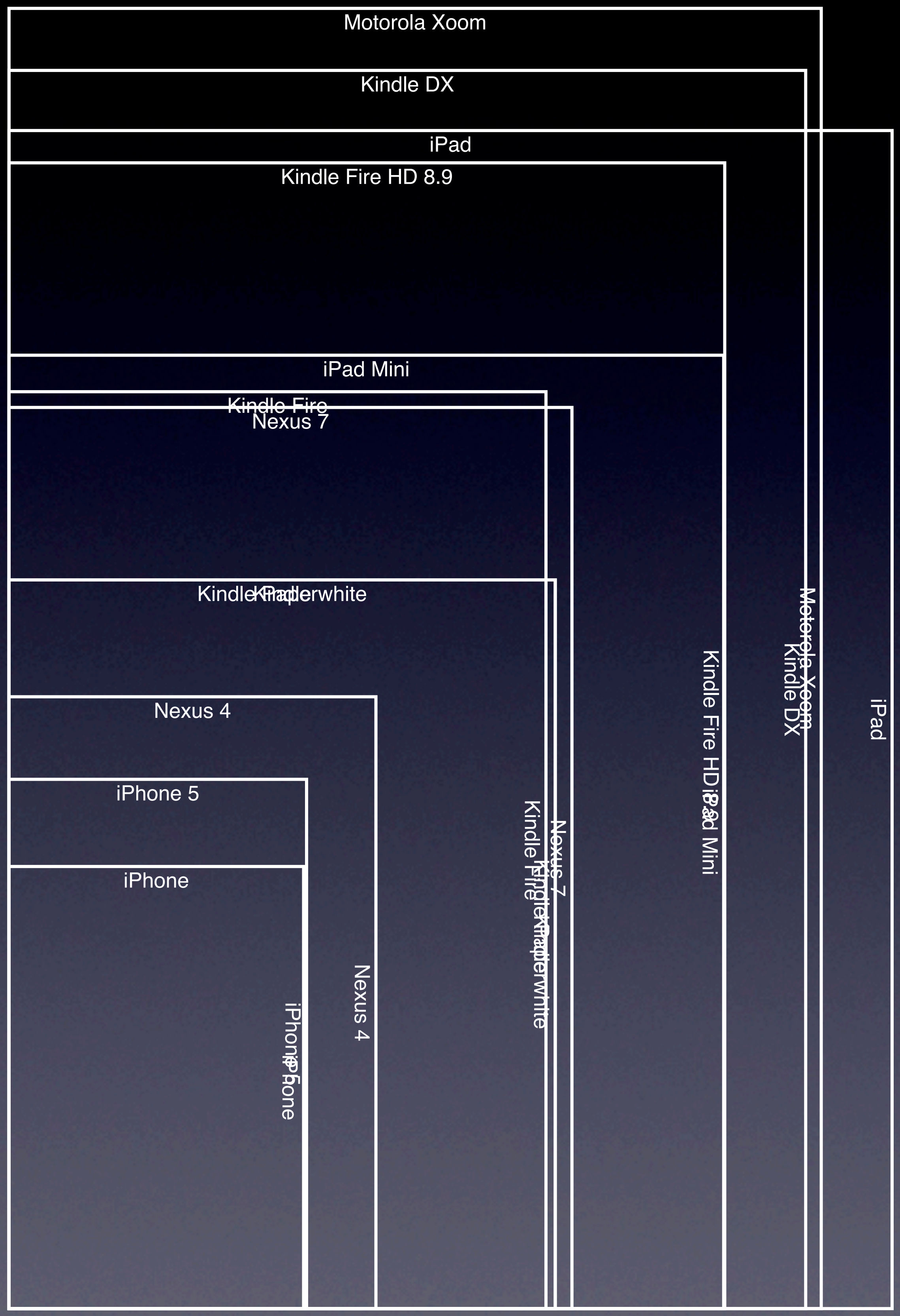
Wired, Popular Science on iPad

Hardware

- eBooks consumed on many different devices
  - Computers: Desktop, Laptops, ~~NetBooks~~
  - iPads and other tablets, dedicated eBook Readers, mobile phones
- Radically different hardware specs (CPU power, screen size, resolution)

eBook Displays

- Variety of sizes
- High-resolution
- LCD Display (Tablets, Mobile Phones)
- eInk Displays (readers)
  - Original Barnes and Noble *Nook* combined both



A variety of phone/tablet/ebookReader screen sizes aligned with each other

eBook Displays

- Variety of sizes
- High-resolution
- LCD Display (Tablets, Mobile Phones)
- eInk Displays (readers)
  - Original Barnes and Noble *Nook* combined both

eBook Displays

- Variety of sizes
- High-resolution
- LCD Display (Tablets, Mobile Phones)
- eInk Displays (readers)
  - Original Barnes and Noble *Nook* combined both

PRIDE & PREJUDICE

1 / 230

CHAPTER 1

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do you not want to know who has taken it?" cried his wife impatiently.

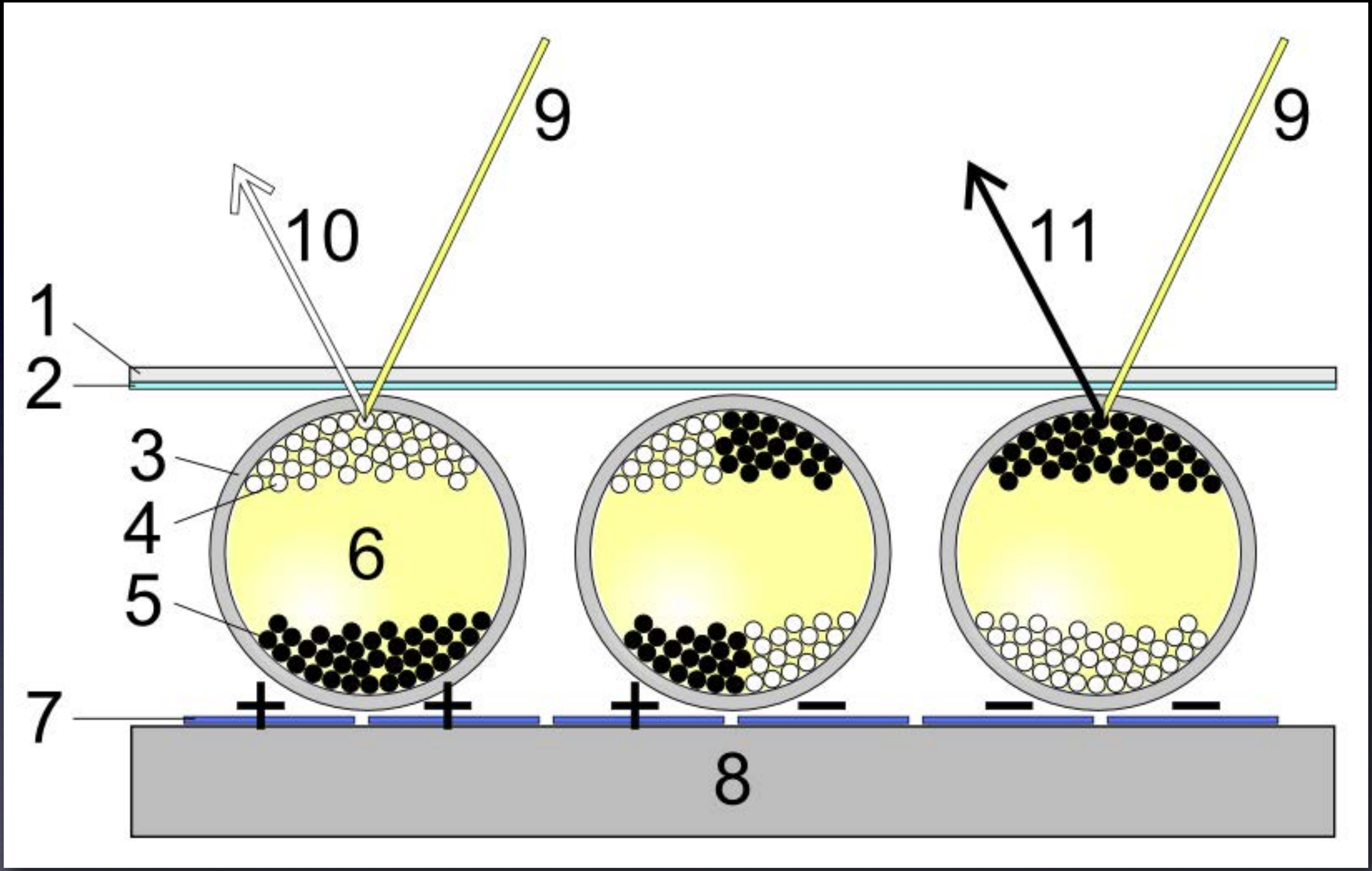
"You want to tell me, and I have no objection to hearing it."

LCD Displays

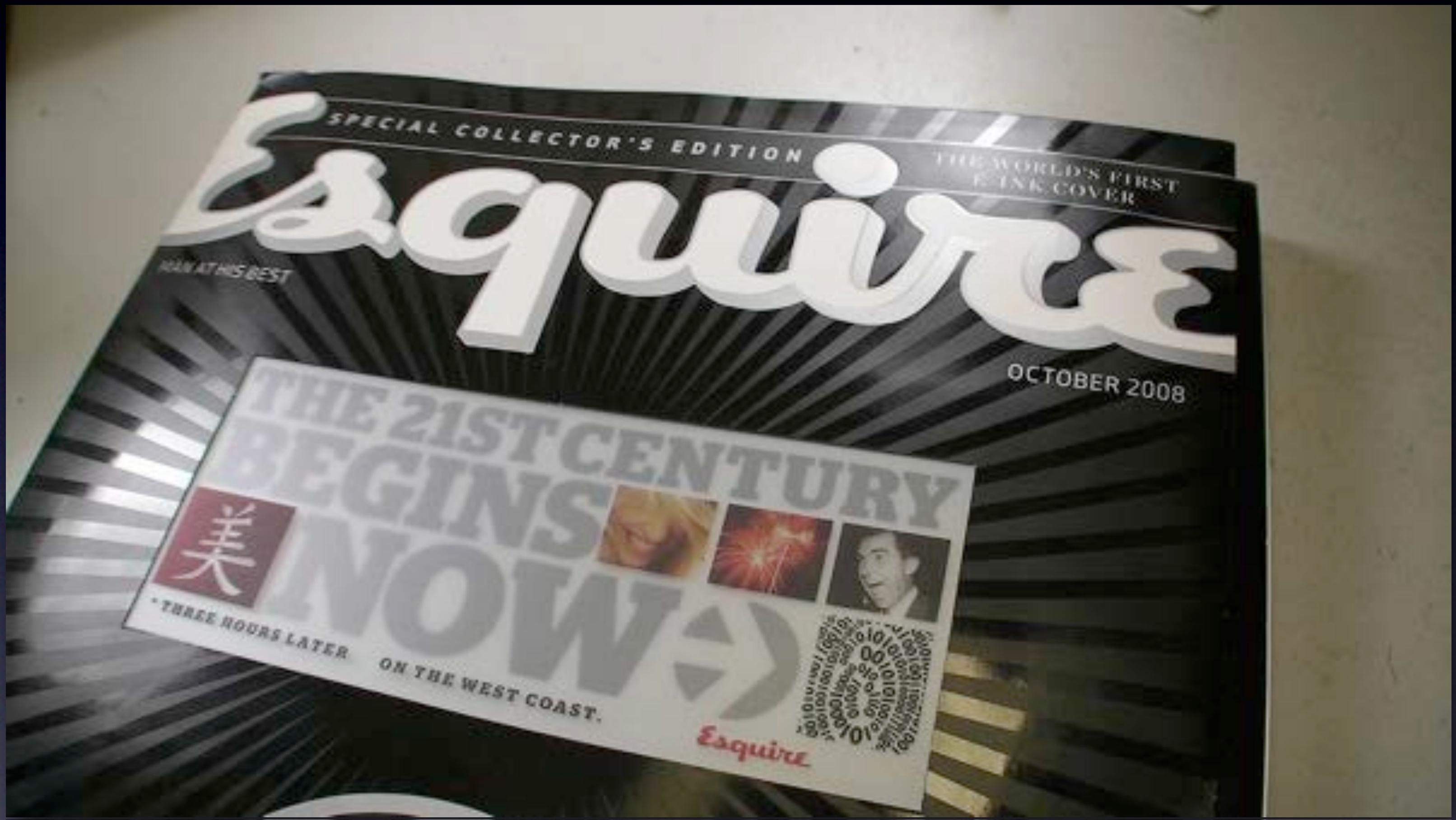
- Tablets, Mobile Phones
- Identical to those found in computers
- High resolutions (e.g. iPad is 2048x1536)
- Continuous power, high refresh rate
- Good for multimedia

eInk Displays

- Typical in eReaders
- Reflective, look like paper
- Slow to refresh
- Low power — once set, permanent
- Resolution higher than on screen (~150dpi)



eInk Display

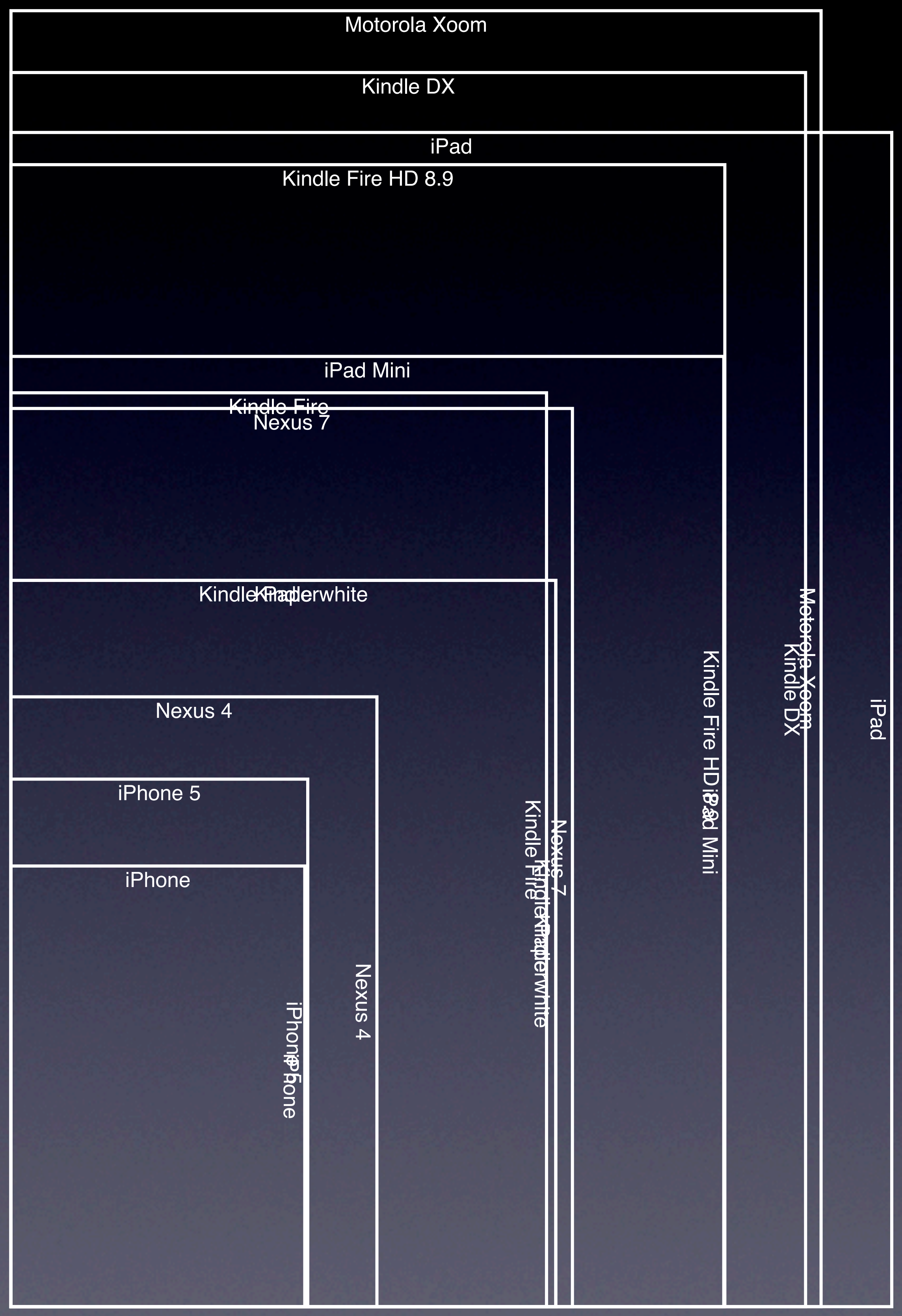

SPECIAL COLLECTOR'S EDITION
THE WORLD'S FIRST E-INK COVER
Esquire
MAN AT HIS BEST
OCTOBER 2008
THE 21ST CENTURY BEGINS NOW
* THREE HOURS LATER ON THE WEST COAST.
Esquire

CPU power

- CPU power of devices varies dramatically
- iPads and Tablets: 1GHz ARM
- Mobile Phones: 500MHz — 1GHz
- eBook Readers: 532MHz
- Efficient algorithms are a must (especially as more pixels are needed on screen)

Hardware Summary

- Variety of display sizes and shapes
    - Documents need to be adaptable to all
    - Display technology alters capabilities of device
- CPU generally low-power and battery-powered
- Software needs to be efficiently written



Viewing format is now the reader’s choice

eBooks

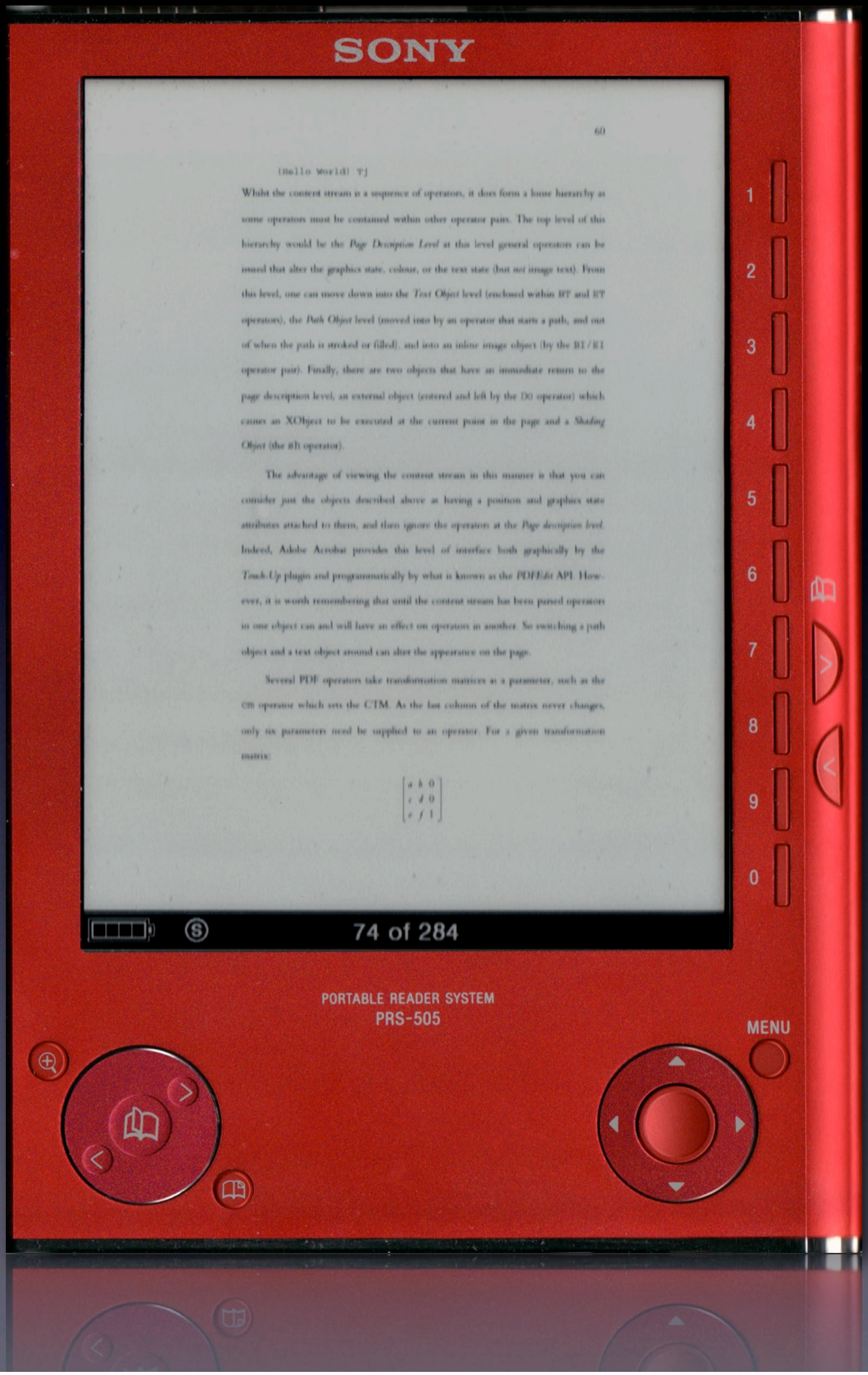
- Viewing format is now the reader’s choice, not the designer’s
- Devices let the reader change the point size, and even fonts used
- Document format must be amenable to change

Current Formats

- Current document formats have tended to mimic paper
- Fixed size at creation
- Fixed position of content
- Seen this with PDF and PostScript
- Nightmare in the eBook world

PDF on eBook readers

- PDF almost always doesn’t quite fit…
- Can either zoom or crop it to fit
- Some software attempts to reflow, but this usually goes horribly wrong…



60

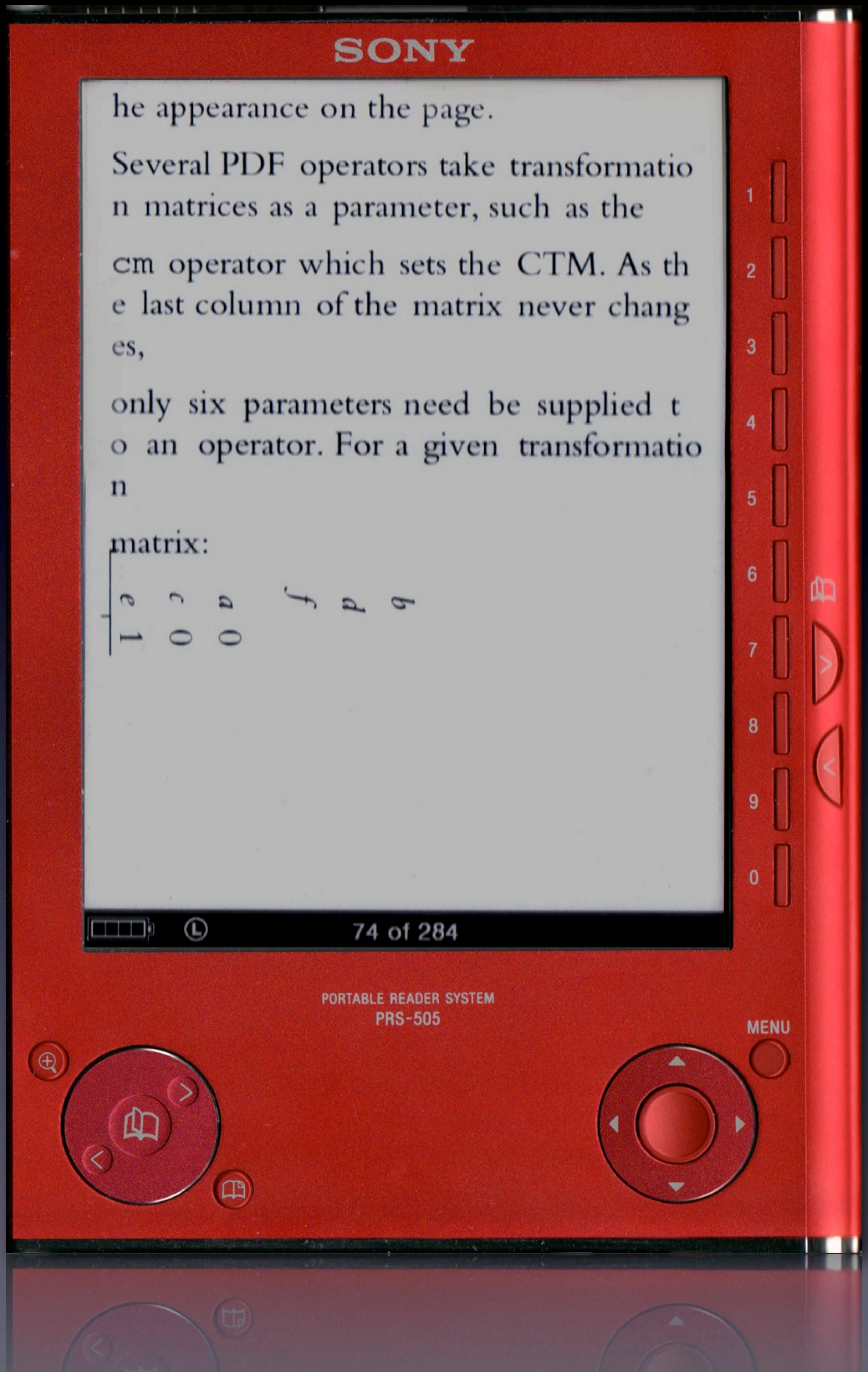
```
(Hello World) Tj
```

Whilst the content stream is a sequence of operators, it does form a loose hierarchy as some operators must be contained within other operator pairs. The top level of this hierarchy would be the *Page Description Level* at this level general operators can be issued that alter the graphics state, colour, or the text state (but *not* image text). From this level, one can move down into the *Text Object* level (enclosed within BT and ET operators), the *Path Object* level (moved into by an operator that starts a path, and out of when the path is stroked or filled), and into an inline image object (by the BI/EI operator pair). Finally, there are two objects that have an immediate return to the page description level, an external object (entered and left by the Do operator) which causes an XObject to be executed at the current point in the page and a *Shading Object* (the sh operator).

The advantage of viewing the content stream in this manner is that you can consider just the objects described above as having a position and graphics state attributes attached to them, and then ignore the operators at the *Page description level*. Indeed, Adobe Acrobat provides this level of interface both graphically by the *Touch-Up* plugin and programmatically by what is known as the *PDFEdit* API. However, it is worth remembering that until the content stream has been parsed operators in one object can and will have an effect on operators in another. So switching a path object and a text object around can alter the appearance on the page.

Several PDF operators take transformation matrices as a parameter, such as the cm operator which sets the CTM. As the last column of the matrix never changes, only six parameters need be supplied to an operator. For a given transformation matrix:

$$\begin{bmatrix} a & b & 0 \\ c & d & 0 \\ e & f & 1 \end{bmatrix}$$



he appearance on the page.

Several PDF operators take transformatio
n matrices as a parameter, such as the

cm operator which sets the CTM. As th
e last column of the matrix never chang
es,

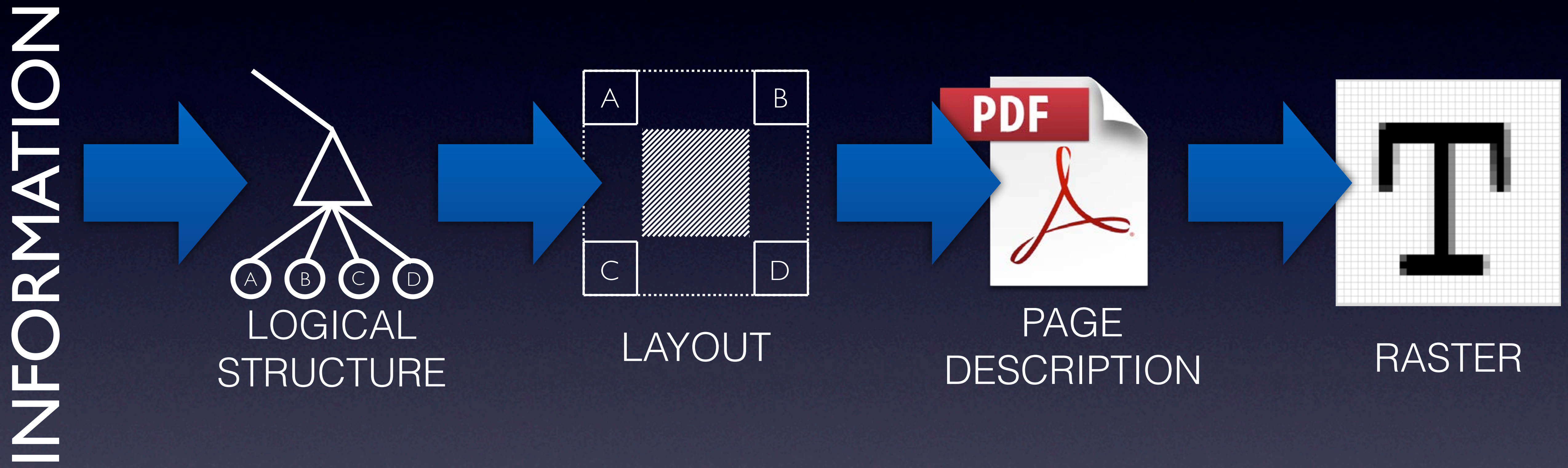
only six parameters need be supplied t
o an operator. For a given transformatio
n

matrix:

e c a f d b
1 0 0

eBook Document Formats

- Final appearance not known when document created
- Must be able to allow the document to be flowed based on the device viewed
- And its configuration
- But also let this be done quickly…



Conceptually, Need to store the document nearer the left-hand edge of the pipeline than the right-side
And finish the transformation on the device…

Software

- Several formats currently used for eBooks
- PDF offers high-quality typesetting but a fixed-layout
- EPUB provides a reflowable alternative – but with limited layout
- Amazon Kindle has its own format (MobiPocket) similar to ePub

Going to look at EPUB2 today… Will look at EPUB3 in more detail in next lecture

EPUB

- Built upon a collection of existing standards XHTML, SVG, Daisy etc.
- Defined in several standards see `http://www.idpf.org/specs.htm`
- Standards form layers
- All contained within a ZIP file
- Worth considering the standard in detail

EPUB

- Multiple versions
- Currently, most books are described using EPUB 2 (or Kindle mobi)
- Limitations to what can be expressed
- EPUB 3 aims to address these limitations

EPUB 3

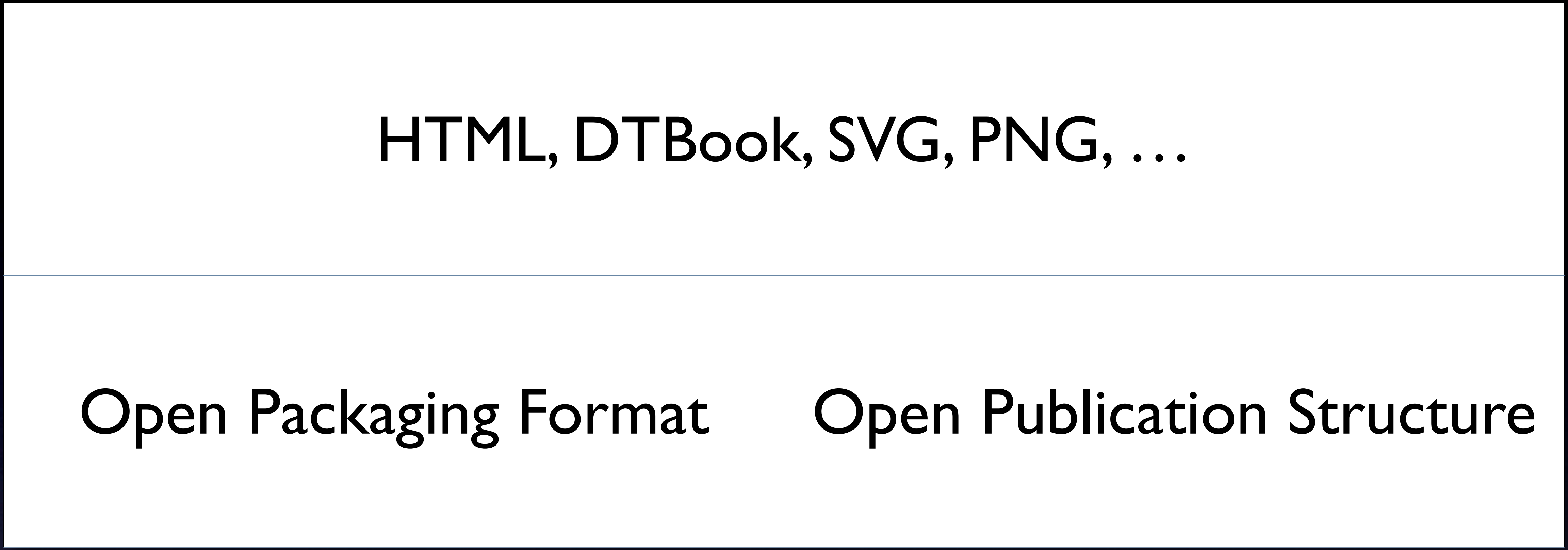
- Support for:
  - Right-to-left and vertical text
  - Mathematics (MathML)
  - Multimedia
  - Interactivity and Scripting
  - Fixed layout

EPUB

- Open-Standard
- But multiple encryption platforms…
- Files generally transfer across platforms
- But formatting issues can arise…

Demo Alice in Wonderland epub on Sony eReader and Apple iPad

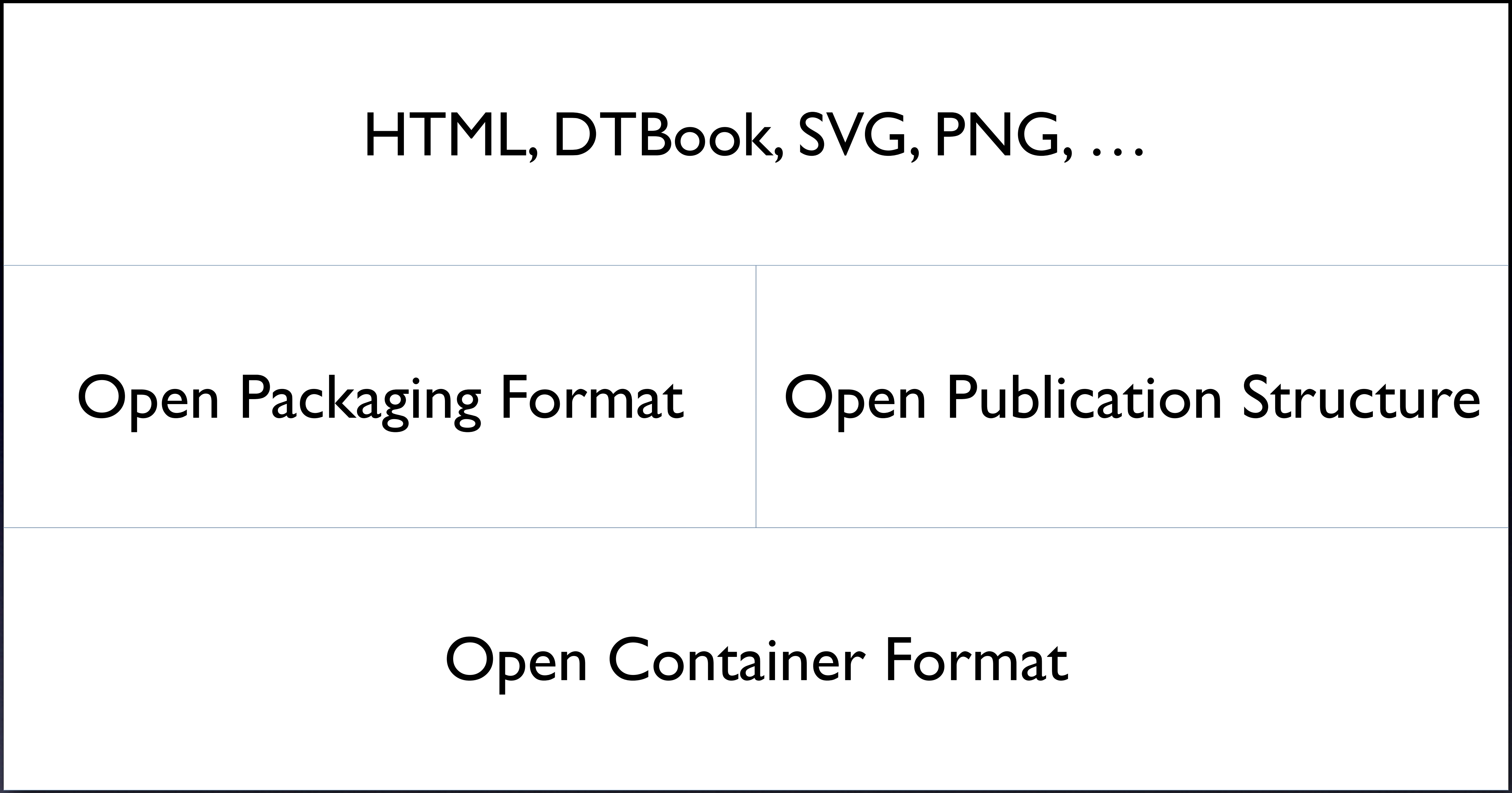
HTML, DTBook, SVG, PNG, …



HTML, DTBook, SVG, PNG, …

Open Packaging Format

Open Publication Structure

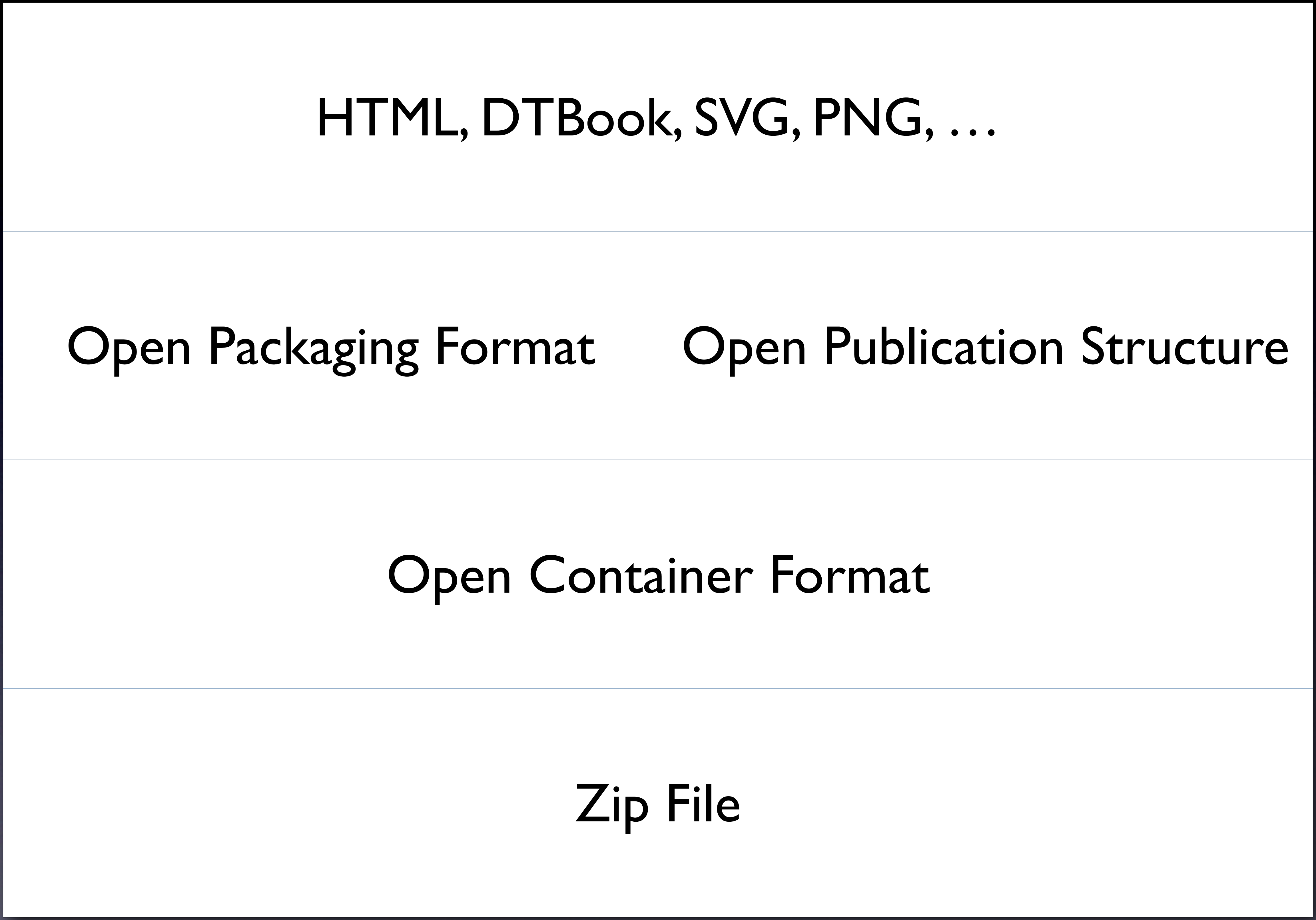


HTML, DTBook, SVG, PNG, …

Open Packaging Format

Open Publication Structure

Open Container Format



HTML, DTBook, SVG, PNG, …

Open Packaging Format

Open Publication Structure

Open Container Format

Zip File

ZIP

- EPUBs are generally distributed as a ZIP file with the `.epub` extension
- Various rules on how ZIP is created (particularly regarding encryption)
- Must contain a `mimetype` file with the value `application/epub+zip`
- The `mimetype` must be first file in ZIP

Open Container Format

- Contains one document, but may contain multiple renditions, e.g. an embedded PDF
- Abstract Container which is mapped onto a file system or ZIP file
- Common root directory for all files

Filenames must be UTF8, restrictions on characters allowed etc

Open Container Format

- Files referenced by relative paths
- Rules defining filenames
- OCF requires a `META-INF` directory and the `mimetype` file in the root

Filenames must be UTF8, restrictions on characters allowed etc

META-INF

File	Description
`container.xml`	(Required) Identifies the MIME type and path to the root file of the OPF/OPS version of the publication plus optional alternate renditions
`metadata.xml`	Optional, contains container-level metadata
`manifest.xml`	

`container.xml`
`encryption.xml`
`rights.xml`
	Used for Digital Rights/Encryption

Container

```
<?xml version="1.0"?>
<container version="1.0" xmlns="urn:oasis:names:tc:opendocument:xmlns:container">
  <rootfiles>
    <rootfile full-path="OEBPS/Great Expectations.opf"
     media-type="application/oebps-package+xml" />
  </rootfiles>
</container>
```

Must be exactly one file of this type

OP[FS]

- Open Publication Structure defines how content is described
- Open Packaging Format packages the content into a publication
  e.g. specifies the order of the individual content files

OP[FS]

- Publication must have one Open Packaging file that ends with `.opf`
- Every file used must be specified in the OPF
- Specifies a list of Core File Types
- Additional, file types can be used providing a fallback core type is provided

Core Media Types

Mime Type	Description
`application/xhtml+xml`	Used for OPS Content Documents
`application/x-dtbook+xml`	
`text/css`	Used for OPS CSS-subset style sheets
`image/svg+xml`	Used for vector graphics
`image/gif`	Used for raster graphics
`image/jpeg`	
`image/png`	
`application/x-dtbncx+xml`	The NCX
`application/xml`	User for Out-Of-Line XML islands
`text/x-oebl-document`	Deprecated
`text/x-oebl-css`	

Open Packaging Format

- Contains:
  - Metadata — using Dublin Core
  - Manifest — must list every file used by the publication
  - Spine — i.e. the ordering of content
  - Optional, Guide or Tours

Go off and look at the Alice eBook

OPS Content

- Often XHTML, OPS specifies limitations on the XHTML allowed
- Can mix SVG (or other XML islands) inside the XHTML (at least according to the spec)

Show problem with the mouse’s tail in Alice

Have a walk through the Alice epub

OPS Content

- Can use CSS (again with limitations)
- Use of the CSS position property values to achieve absolute positioning (i.e. `absolute` and `fixed`) is strongly discouraged.

Show problem with the mouse’s tail in Alice

Have a walk through the Alice epub