University of Southampton

Web Formats

COMP3227 Web Architecture & Hypertext Technologies

Dr Heather Packer – hp3@ecs.soton.ac.uk

University of Southampton

Web Formats

HTML is the main Web format

  – Many other formats in use on the Web

  – Many other formats use Web standards

NOTE: This lecture goes into a lot of detail, but for illustrative purposes only. You should be broadly familiar with the range of formats, what they're for and (roughly) how they work

Standard Generalised Markup Language

- Standard Generalized Markup Language (SGML)

- ISO standard 1986

- Language for defining Markup Languages

- Elements <..> ...</..>

- Graph Structure

- Abbreviation of expressions

Family of Languages

SGML

DocBook   HTML   XML

SVG

XHTML   MathML
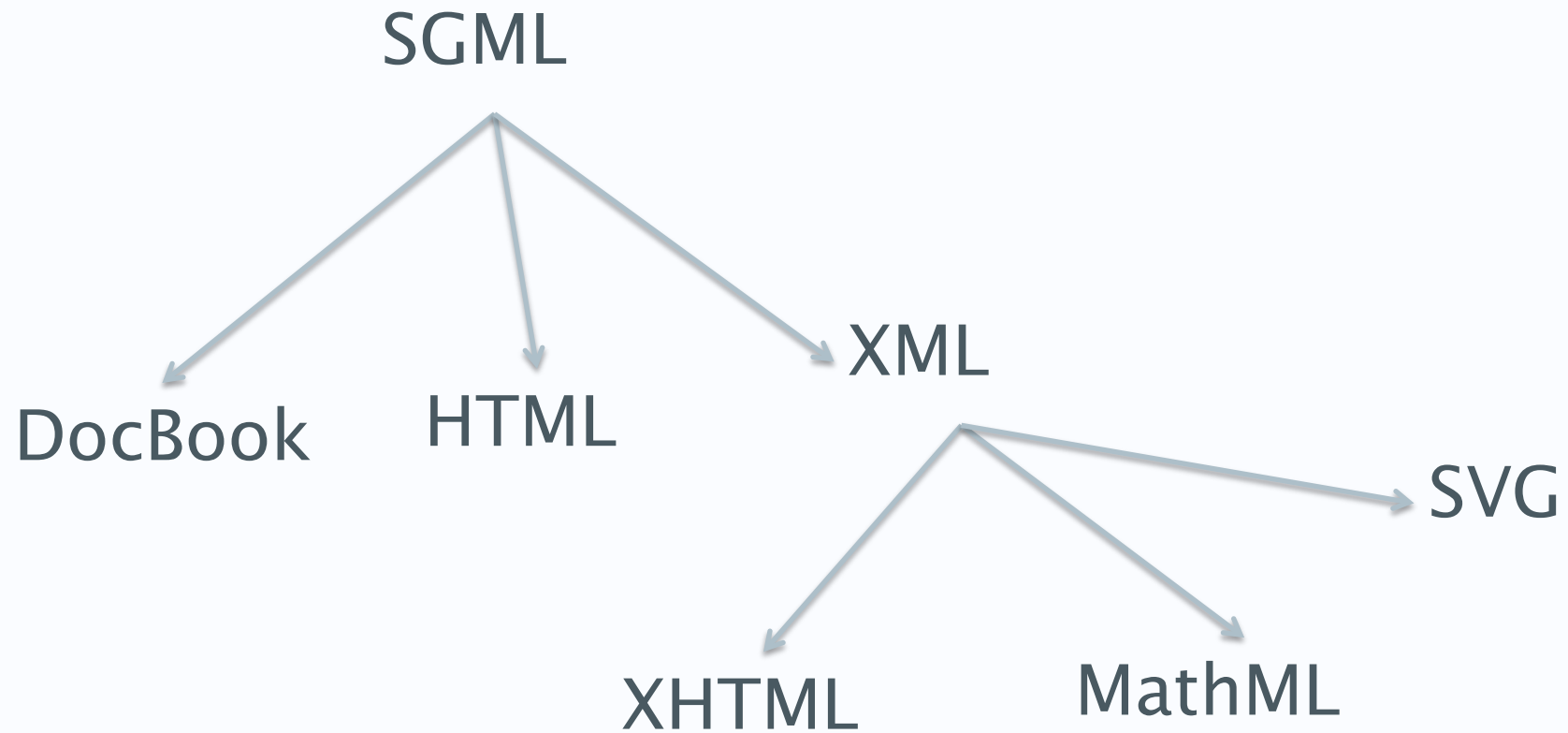
eXentsible Markup Language

The eXtensible Markup Language

A **general purpose** markup language

– A W3C-defined subset of the Standard Generalized Markup Language

A markup language for defining domain-specific markup languages

Used as the basis for a number of Web formats:

– Scalable Vector Graphics

– Resource Description Framework

– Synchronised Multimedia Integration Language

– Simple Object Access Protocol

– eXtensible Stylesheet Language Transformations

– (but not HTML5)

XML example

```
<?xml version="1.0"?>
<!DOCTYPE booklist SYSTEM "books.dtd">
<booklist>
 <books>
  <item cat="S">
   <title>I, Robot</title>
   <author>Asimov, Isaac</author>
   <price>5.95</price>
   <quantity>3</quantity>
  </item>
 <item cat="C">
   <title>Persuasion</title>
   <author>Austen, Jane</author>
   <price>6.95</price>
   <quantity>2</quantity>
  </item>
 </books>
</booklist>
```

XML declaration
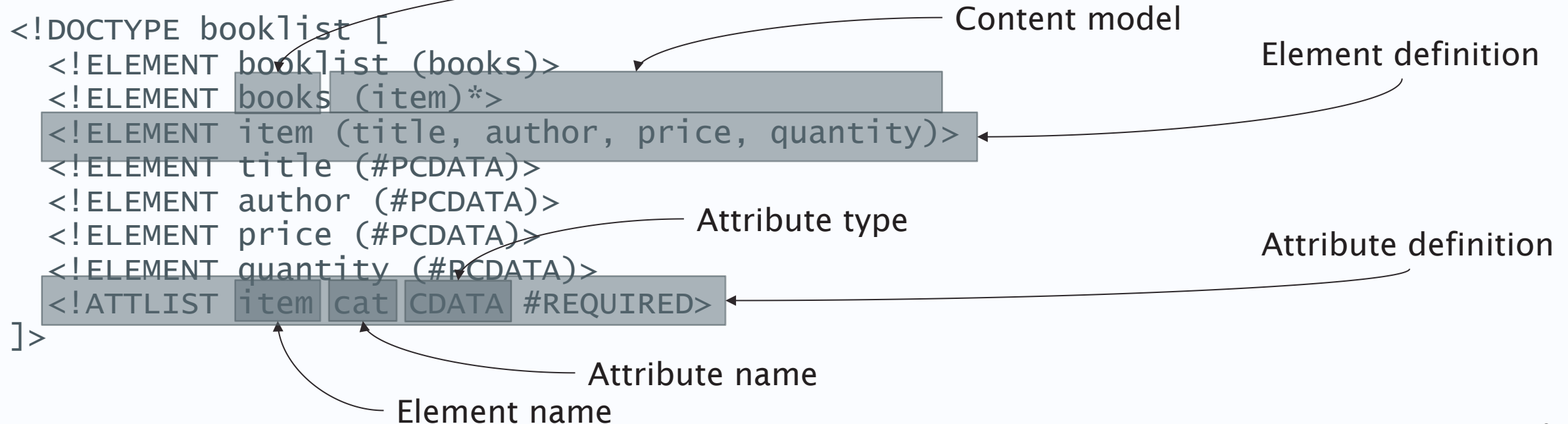Tells a document processor that this is XML

Reference to a Document Type Definition
Tells a document processor how to parse this document

Document Type Declaration (doctype)
Tells a document processor what type of document this is

Document Type Definition (DTD)

A formal definition of the grammar for an XML document type

- What elements and attributes exist

- What elements can exist inside other elements (the content model)

- Referenced by the document type declaration

Element name

Content model

Element definition

```
<!DOCTYPE booklist [
  <!ELEMENT booklist (books)>
  <!ELEMENT books  (item)*>
  <!ELEMENT item (title, author, price, quantity)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT author (#PCDATA)>
  <!ELEMENT price (#PCDATA)>
  <!ELEMENT quantity (#PCDATA)>
  <!ATTLIST item cat CDATA #REQUIRED>
]>
```

Element definition

Attribute type

Attribute definition

Attribute name

Element name

9

Well-Formedness versus Validity

An XML document is *well-formed* if it obeys the syntax rules in the XML spec:

- Single root element

- Elements are correctly nested (no overlapping)

- Tag names contain only legal characters

- Start and end tag names have matching capitalisation

- (to name but a few of the rules)

Well-Formedness versus Validity

An XML document is *valid* if:

- It contains a reference to a DTD

- It only contains elements and attributes that are defined in that DTD

- Its use of those elements and attributes follows the grammar rules in the DTD

- All valid XML documents are well-formed

- Not all well-formed XML documents are valid

Other Schema Languages

Document Type Definitions have expressive limitations

- Cannot specify the range of values taken by attributes
- Cannot specify the range of non-markup element content

Two main competitors:

- XML Schema
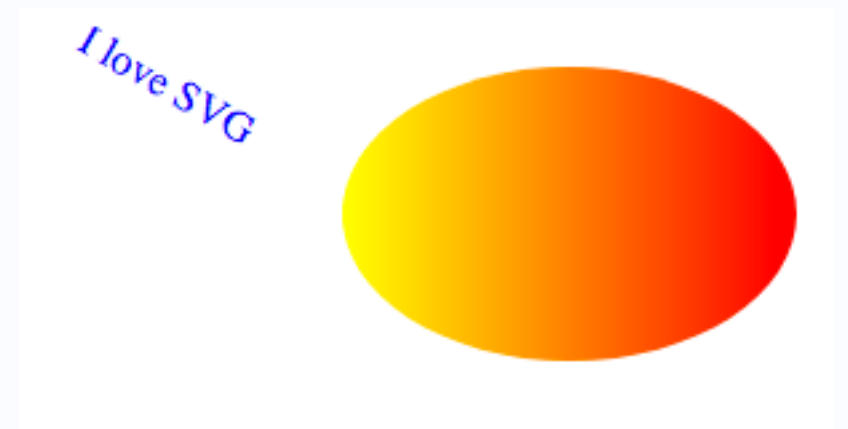- RELAX NG

Scalable Vector Graphs

Scalable Vector Graphics

XML-based language for describing 2D graphics

- – Resolution independent

- – Support for Javascript event handlers

- – Support for manipulation via the Document Object Model (DOM)

- – Uses CSS for styling and animation

- – Integrates with HTML5

SVG Example

```
<svg height="150" width="400" xmlns:xlink="http://www.w3.org/1999/xlink">
  <defs>
    <linearGradient id="grad1" x1="0%" y1="0%" x2="100%" y2="0%">
      <stop offset="0%" style="stop-color:rgb(255,255,0);stop-opacity:1" />
      <stop offset="100%" style="stop-color:rgb(255,0,0);stop-opacity:1" />
    </linearGradient>
  </defs>
  <ellipse cx="200" cy="70" rx="85" ry="55" fill="url(#grad1)" />
  <text x="0" y="15" fill="blue" transform="rotate(30 20,40)">I love
    <a xlink:href="http://www.w3.org/SVG/" target="_blank">SVG</a></text>
</svg>
```

MathML

MathML

XML-based language for expressing mathematical expressions

– Integrates with HTML5

Two sub-languages:

– Presentation-oriented (for display)

– Semantics-oriented

Presentational MathML

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup><mi>a</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup><mi>b</mi><mn>2</mn></msup>
    <mo>=</mo>
    <msup><mi>c</mi><mn>2</mn></msup>
  </mrow>
</math>
```

$$a^2 + b^2 = c^2$$

Semantic MathML

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <apply>
    <eq/>
    <apply>
      <plus/>
      <apply>
        <power/><ci>a</ci><cn>2</cn>
      </apply>
      <apply>
        <power/><ci>b</ci><cn>2</cn>
      </apply>
    </apply>
    <apply>
      <power/><ci>c</ci><cn>2</cn>
    </apply>
  </apply>
</math>
```

$$a^2 + b^2 = c^2$$

Web Data

Structured and Linked Data on the Web

The Resource Description Framework

- – Subject of a later lecture on this module
- – Covered (in considerable depth) in COMP6215 Semantic Web Technologies next semester

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:ns0="http://data.ordnancesurvey.co.uk/ontology/spatialrelations/"
  xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:oo="http://purl.org/openorg/"
  xmlns:ns1="http://id.southampton.ac.uk/ns/"
  xmlns:ns2="http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#"
  xmlns:ns3="http://purl.org/NET/c4dm/event.owl#"
  xmlns:ns4="http://purl.org/NET/c4dm/timeline.owl#">

  <rdf:Description rdf:about="http://id.southampton.ac.uk/room/35-1005">
    <rdf:type rdf:resource="http://vocab.deri.ie/rooms#Room"/>
    <rdf:type rdf:resource="http://id.southampton.ac.uk/ns/SyllabusLocation"/>
    <rdf:type rdf:resource="http://id.southampton.ac.uk/ns/CentrallyBookableSyllabusLocation"/>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Lecture Theatre</rdfs:label>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">35 / 1005</rdfs:label>
    <ns0:within rdf:resource="http://id.southampton.ac.uk/building/35"/>
    <ns0:within rdf:resource="http://id.southampton.ac.uk/floor/35-1"/>
    <geo:lat rdf:datatype="http://www.w3.org/2001/XMLSchema#float">50.9338573</geo:lat>
    <geo:long rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-1.394543395</geo:long>
```
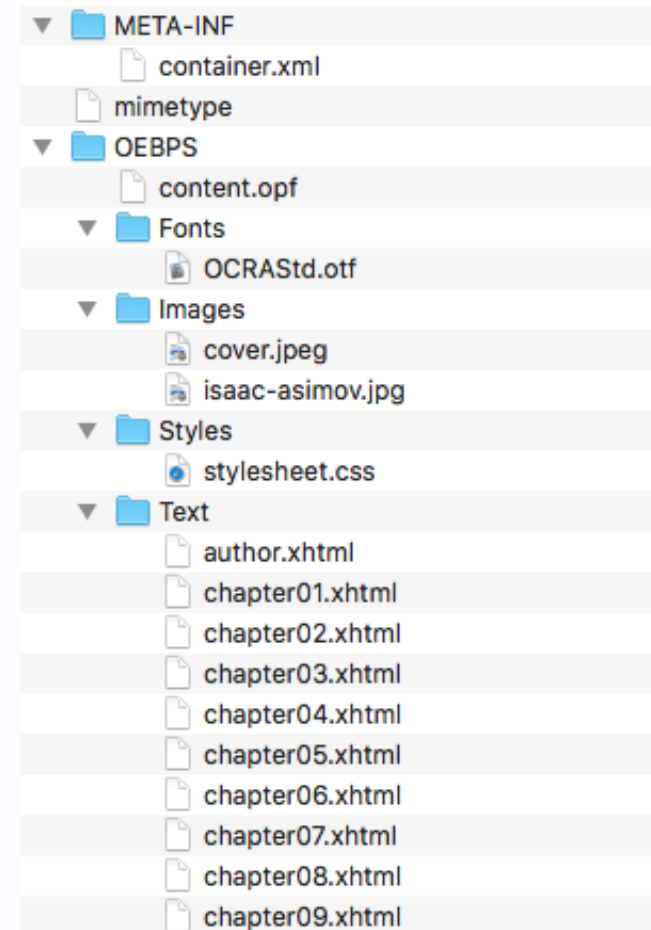
EPub

ePub Format

Open vendor-neutral standard for e-books defined by IDPF (now part of W3C)

ZIP of XML/HTML files

- META-INF/container.xml

- OEBPS/content.opf

- Kindle (.azw), Mobipocket, Apple iBooks

Use of HTML allows resizable and reflowable content – essential for adapting to a wide variety of readers

Other common ebook formats take similar approach (ZIP of XML/HTML files)

▼ 📁 META-INF
   📄 container.xml
📄 mimetype
▼ 📁 OEBPS
   📄 content.opf
   ▼ 📁 Fonts
      📄 OCRAStd.otf
   ▼ 📁 Images
      🖼 cover.jpeg
      🖼 isaac-asimov.jpg
   ▼ 📁 Styles
      📄 stylesheet.css
   ▼ 📁 Text
      📄 author.xhtml
      📄 chapter01.xhtml
      📄 chapter02.xhtml
      📄 chapter03.xhtml
      📄 chapter04.xhtml
      📄 chapter05.xhtml
      📄 chapter06.xhtml
      📄 chapter07.xhtml
      📄 chapter08.xhtml
      📄 chapter09.xhtml

META-INF/container.xml

Points to OPF package which describes the other components of the document

```
<?xml version="1.0" encoding="UTF-8"?>
<container version="1.0"
  xmlns="urn:oasis:names:tc:opendocument:xmlns:container">
  <rootfiles>
    <rootfile full-path="OEBPS/content.opf"
              media-type="application/oebps-package+xml"/>
  </rootfiles>
</container>
```

OEBPS/content.opf

Three key components:

– Metadata about document

```
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/"
          xmlns:dcterms="http://purl.org/dc/terms/"
          xmlns:opf="http://www.idpf.org/2007/opf"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:identifier id="uuid_id" opf:scheme="uuid">
    df3d24ec-aa53-4a72-9075-e97b5b7bc26f</dc:identifier>
  <dc:title>The Stars, Like Dust</dc:title>
  <dc:creator opf:file-as="Asimov, Isaac" opf:role="aut">Isaac Asimov</dc:creator>
  <dc:language>en</dc:language>
</metadata>
```

OEBPS/content.opf

Three key components:

– Metadata about document

– Manifest listing files that comprise document

```
<manifest>
  <item href="Images/cover.jpeg" id="cover"
        media-type="image/jpeg"/>
  <item href="Styles/stylesheet.css" id="css"
        media-type="text/css"/>
  <item href="Text/cover.xhtml" id="cover.xhtml"
        media-type="application/xhtml+xml"/>
  <item href="Text/chapter01.xhtml" id="chapter01.xhtml"
        media-type="application/xhtml+xml"/>
  ...
</manifest>
```

OEBPS/content.opf

Three key components:

– Metadata about document

– Manifest listing files that comprise document

– Spine listing table of contents

```
<spine toc="ncx">
  <itemref idref="cover.xhtml"/>
  <itemref idref="title.xhtml"/>
  <itemref idref="chapter01.xhtml"/>
  <itemref idref="chapter02.xhtml"/>
  <itemref idref="chapter03.xhtml"/>
  ...
</spine>
```

OEBPS/Text/chapter01.xhtml

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org/TR/xhtml11/DTD/
xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <link href="../Styles/stylesheet.css" rel="stylesheet"
        type="text/css"/>
  </head>
  <body>
    <h1>ONE: The Bedroom Murmured</h1>
    <p>The bedroom murmured to itself gently. It was almost below the limits of hearing
-an irregular little sound, yet quite unmistakable, and quite deadly.</p>
    <p>But it wasn't that which awakened Biron Farrill and dragged him out of a heavy,
unrefreshing slumber. He turned his head restlessly from side to side in a futile
struggle against the periodic burr-r-r on the end table.</p>
    <p>He put out a clumsy hand without opening his eyes and closed contact.</p>
    <p>"Hello," he mumbled.</p>
```
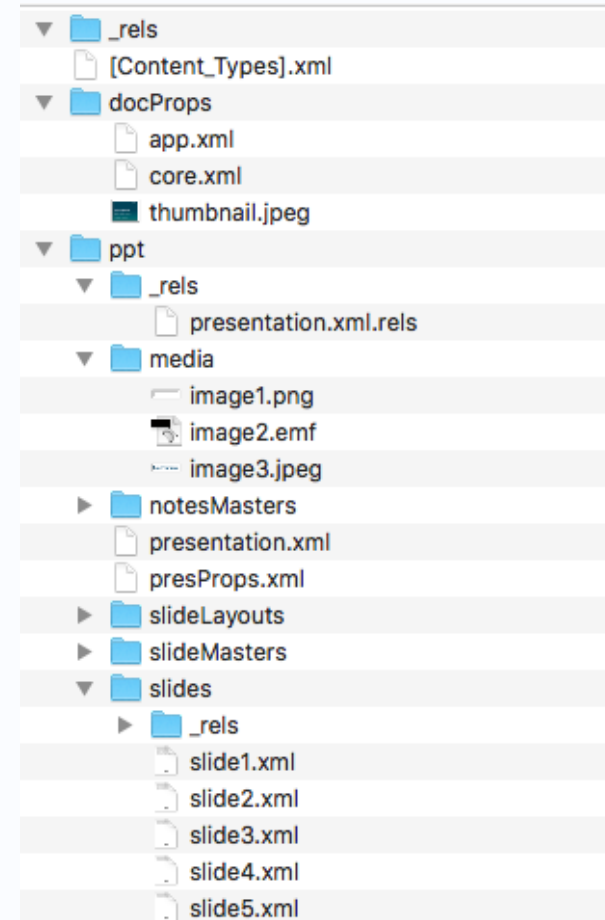
Open Office XML

Open Office XML

Microsoft-originated XML-based format

- – Standardised by Ecma and ISO/IEC
- – Replaced pre-2007 proprietary format

ZIP file of directory hierarchy containing XML

- – docprops/ contains metadata
- – ppt/slides contains slides
- – ppt/media contains images
- – _rels translates file names into XML attribute values

```
▼ 📁 _rels
  📄 [Content_Types].xml
▼ 📁 docProps
  📄 app.xml
  📄 core.xml
  🖼 thumbnail.jpeg
▼ 📁 ppt
  ▼ 📁 _rels
    📄 presentation.xml.rels
  ▼ 📁 media
    🖼 image1.png
    🖼 image2.emf
    🖼 image3.jpeg
  ▶ 📁 notesMasters
  📄 presentation.xml
  📄 presProps.xml
  ▶ 📁 slideLayouts
  ▶ 📁 slideMasters
  ▼ 📁 slides
    ▶ 📁 _rels
    📄 slide1.xml
    📄 slide2.xml
    📄 slide3.xml
    📄 slide4.xml
    📄 slide5.xml
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<p:sld xmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main"
    xmlns:r="http://schemas.openxmlformats.org/officeDocument/2006/relationships"
    xmlns:p="http://schemas.openxmlformats.org/presentationml/2006/main">
<p:cSld><p:spTree>
<p:nvGrpSpPr><p:cNvPr id="1" name=""/><p:cNvGrpSpPr/><p:nvPr/></p:nvGrpSpPr>
<p:grpSpPr><a:xfrm><a:off x="0" y="0"/><a:ext cx="0" cy="0"/><a:chOff x="0" y="0"/><a:chExt cx="0"
cy="0"/></a:xfrm></p:grpSpPr>
<p:sp><p:nvSpPr><p:cNvPr id="2" name="Title 1"/><p:cNvSpPr><a:spLocks noGrp="1"/></p:cNvSpPr>
<p:nvPr><p:ph type="ctrTitle"/></p:nvPr></p:nvSpPr><p:spPr/>
<p:txBody><a:bodyPr/><a:lstStyle/><a:p><a:r><a:rPr lang="en-GB" dirty="0"/><a:t>Web Formats</
a:t></a:r></a:p></p:txBody></p:sp><p:sp><p:nvSpPr><p:cNvPr id="3" name="Subtitle 2"/
><p:cNvSpPr><a:spLocks noGrp="1"/></p:cNvSpPr><p:nvPr><p:ph type="subTitle" idx="1"/></
p:nvPr></p:nvSpPr><p:spPr/><p:txBody><a:bodyPr/><a:lstStyle/>
<a:p><a:r><a:rPr lang="en-GB" dirty="0"/><a:t>COMP3220 Web Infrastructure</a:t></a:r></a:p>
</p:txBody></p:sp><p:sp><p:nvSpPr><p:cNvPr id="4" name="Text Placeholder 3"/
><p:cNvSpPr><a:spLocks noGrp="1"/></p:cNvSpPr><p:nvPr><p:ph type="body" sz="quarter" idx="13"/
></p:nvPr></p:nvSpPr><p:spPr/><p:txBody><a:bodyPr/><a:lstStyle/><a:p><a:r><a:rPr lang="en-GB"
dirty="0"/><a:t>Dr Nicholas Gibbins </a:t></a:r><a:r><a:rPr lang="mr-IN" dirty="0"/><a:t>–</a:t></
a:r><a:r><a:rPr lang="en-GB" dirty="0"/><a:t> </a:t></a:r><a:r><a:rPr lang="en-GB" dirty="0" err="1"/
><a:t>nmg@ecs.soton.ac.uk</a:t></a:r><a:endParaRPr lang="en-GB" dirty="0"/></a:p></p:txBody></
p:sp></p:spTree><p:extLst><p:ext uri="{BB962C8B-B14F-4D97-AF65-F5344CB8AC3E}">
...
```

Portable Document Format

Portable Document Format

Not "of the Web", but important for the Web

- 8.5bn HTML documents in Google

- 2.3bn PDF documents in Google

Structured for rendering of pre-formatted documents

- Set characters from fonts at position

- Draw lines (etc) at position

- No structure to text: no paragraphs, headings, lists, etc

Often used as official format of record

- Searchable – unlike scanned documents

PDF History

Derived from Adobe's earlier PostScript language

– Subset of PostScript's page description language
(but not a programming language like PostScript)

Other features

– Font embedding in documents

– Structured object storage, with data compression

– Access control and DRM

– Extensible metadata

– Fillable forms, annotations

– Links!

Sample PDF

```
%PDF-1.0
1 0 obj
<<
/Type /Catalog
/Pages 3 0 R
/Outlines 2 0 R
>>
endobj
2 0 obj
<</Type /Outlines /Count 0>>
endobj
3 0 obj
<<
/Type /Pages
/Count 1
/Kids [4 0 R]
>>
endobj
```

Root Object

Outlines Object (TOC)

Page List

```
4 0 obj
<<
/Type /Page
/Parent 3 0 R
/Resources <<
/Font << /F1 7 0 R >>
/ProcSet 6 0 R >>
/MediaBox [0 0 612 792]
/Contents 5 0 R
>>
endobj
5 0 obj
<< /Length 44 >>
stream
BT /F1 24 Tf
100 100 Td (Hello World) Tj ET
endstream
endobj
```

First Page

Drawing commands for first page

BeginText, use font F1 at size 24, move to (100,100), draw the text "Hello World", EndText

Sample PDF

```
6 0 obj
[/PDF /Text]
endobj
7 0 obj
<<
/Type /Font
/Subtype /Type1
/Name /F1
/BaseFont /Helvetica
>>
endobj
```

Definitions for first page

Fonts for first page

Number of objects, ID of root

```
xref
08
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000322 00000 n
0000000415 00000 n
0000000445 00000 n
trailer
<<
/Size 8
/Root 1 0 R
>>
startxref 553
%%EOF
```

Index

Next:
Web APIs