Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 2. FORMATS FOR ELECTRONIC DOCUMENTS AND IMAGES

LESSON 1. TYPES OF MARK-UP: INTRODUCTION

NOTE

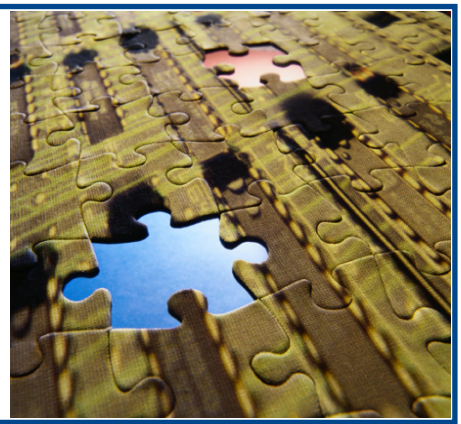
Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Objectives

At the end of this lesson, you will able to:

- understand the **purpose of mark-up**, and
- distinguish between **different kinds of mark-up**.



Why we need Mark-up

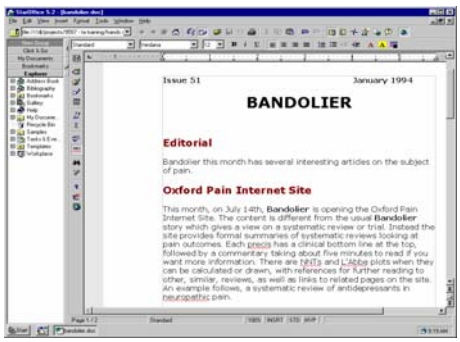
Electronic text documents are stored in files on our computer disks. We can read electronic documents using software applications, such as word processors or desktop publishing systems, that assist us in creating, managing and sharing them with other people.

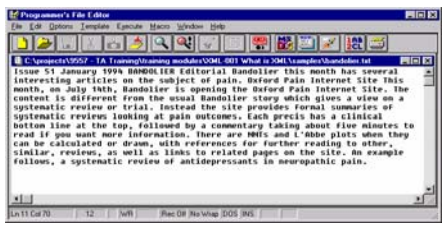
We often exchange electronic documents over computer networks, either **networks** internal to an organization or the Internet, either as **web pages** or as attachments to **e-mail messages**.

Often we **print** electronic documents in order to read them, and so this needs to be taken into account when creating them.

Why we need Mark-up

These two electronic documents contain the same text.
The one on the left is easy to read (and to edit) because it is laid out with a title, sections and headings, while the one on the right is not.





This is because the document on the right has no **mark-up** to instruct the software to display the document with an easy to understand layout.

Why we need Mark-up

Mark-up originally referred to the hand-written notations that a designer would add to typewritten text.

These notations contained instructions to a typesetter about **how to lay out the copy** and what **typeface** to use.

Why we need Mark-up

Today, almost every electronic document that we use contains two types of information:

- the text **content** of the document itself, and
- a set of **codes** that provides information on how to display or interpret the text.

These additional codes that are contained in the electronic file are the **mark-up**.

Mark-up is everything in a document that is not content.

Types of Mark-up

There are three types of mark-up codes that can be used in an electronic document:

Procedural mark-up consists of codes that contain information on how a specific application should process the document.

Presentational mark-up consists of codes that describe how the document should be presented or laid out, either on a computer screen or on a printed page.

Descriptive mark-up consists of codes that describe the logical structure and semantics of a document, usually in a way that can be interpreted by many different software applications.

Now, let's have a look at the different characteristics of each kind of mark-up...

Procedural Mark-up

Most electronic publishing systems today, such as word processing software and desktop publishing software, use **procedural mark-up**.

Procedural mark-up refers to the special control characters that are inserted into electronic text files prior to their submission and subsequent interpretation by output devices.

"Choose option one **or** two."

" Choose option one \fB or \fR two."

Print the following characters in Times Bold

Revert to the default style – Times Roman

Different codes are attached to section headings, paragraphs of body text, references and even individual characters and words so that each is set in an appropriate type style, size and line spacing.

On the left you have two examples of commands used to determine font style.

Procedural Mark-up

Procedural mark-up usually takes the form of **formatting codes** that are **mixed in with the text** of the document.

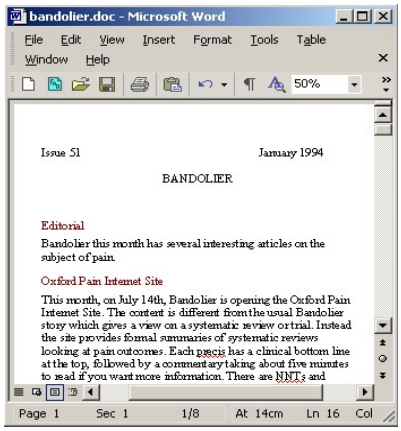
Can you identify, in the following example, which is the text content of the document?

```
{\pgdsc0\pgdscuse195\pgwsxn11905\pghsxn16837\marglsxn1800\margrsxn1800\margtsxn1440\m
\paperh16837\paperw11905\margl1800\margr1800\margt1440\margb1440\sectd\shknone\pgwsxn
\pard\plain \s1\f2{\b Issue 51}\tab \tab \tab \tab \tab \tab {\b January 1994}
\par
\par \pard\plain \s1\f2\fs48\b\qc BANDOLIER
\par
\par \pard\plain \s1\f2
\par \pard\plain \s1\cf1\f2\fs32\b Editorial
\par \pard\plain \s1\f2
```

Type the text in the box.

Then, click on View Answer.

Procedural Mark-up



Generally speaking, procedural mark-up formats are designed (and owned) by vendors of **specific software products**, and the best application to process documents in that format is the one that the mark-up was designed for.

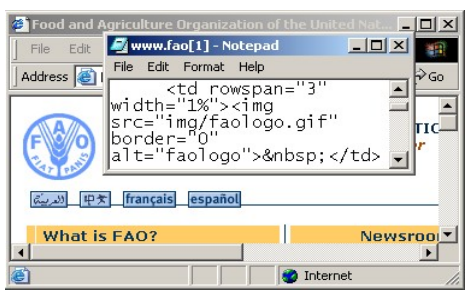
One of the most popular procedural formats is **Microsoft Word**.

Procedural mark-up codes apply to a single way of presenting the information, such as a printed page, and provide no capability to define appearance for other media, such as CD-ROM and Internet.

Presentational Mark-up

Presentational mark-up codes apply to different ways of presenting the information.

Presentational mark-up describes graphics, layout and page control features, either on a computer screen or on a printed page.
One of the most widely-used forms of presentational mark-up is HTML (Hyper Text Mark-up Language).



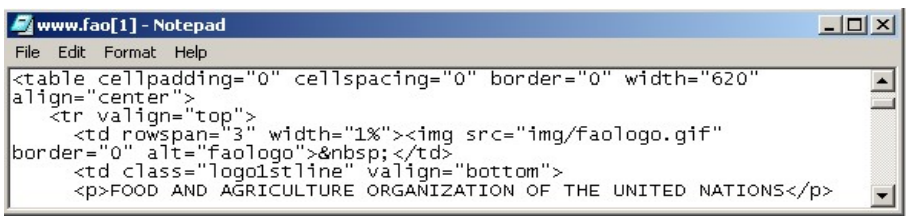
HTML is used to mark-up pages for presentation in a **web browser**.

In this example, the HTML source describes the position of the FAO logo on the web page.
Unlike many procedural mark-up languages, HTML is an open standard, (not a proprietary format owned by a single software vendor), published by the World Wide Web Consortium.

Presentational Mark-up

The HTML mark-up provides a standard way of specifying how the document will be presented in a web browser; when you select "**Source**" from the "View" menu in Internet Explorer, you can see the HTML description of the web page displayed.

HTML mark-up is in **angle brackets < >** and specifies headers, paragraphs, bold text, lists, tables, etc. Exactly how each of these elements is displayed depends on the browser used to view the document.



HTML mark-up codes are 'clear text' that can be read by almost any text processing software and are easily distinguished from the text content of the document.

Descriptive Mark-up

HTML marks up how the document content is presented, not the type, structure or meaning of the content: if we want to capture that information we need to use **descriptive mark-up**.

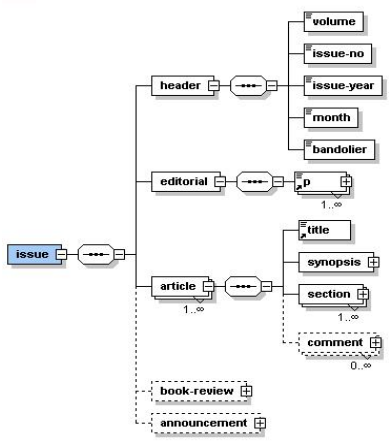
Rather than containing codes that describe the layout or presentation of the document, descriptive mark-up contains codes that define a **logical**, usually **hierarchical structure**.

```
- <issue>
  - <header>
      <volume>6</volume>
      <issue-no>5</issue-no>
      <issue-year>99</issue-year>
      <month>July</month>
      <bandolier>65</bandolier>
    </header>
  - <editorial id="b65-1" clinical-code="123">
      <p>Bandolier this month has several
    interesting articles on the subject of pain.</p>
    </editorial>
  - <article id="b65-2" clinical-code="456">
      <title>Oxford Pain Internet Site</title>
    - <synopsis>
```

The illustration shows a document where elements are marked up as issue-number, volume, editorial, article, etc. These are all **logical elements** in the document structure, rather than instructions about how those elements should be presented or processed.

Since no directions about formatting are included, the **interpretation of the mark-up tags** occurs entirely **within the processing system**.

Descriptive Mark-up



Our example uses **XML**: the Extensible Mark-up Language.

XML is the most prevalent form of descriptive mark-up in use today, and is a standard of the World Wide Web Consortium (www.w3.org).

XML describes only the logical structure of the document: the figure illustrates the type of hierarchical structure that can be defined using XML.

The presentational style can be applied by referencing a stylesheet that is held in a separate file from the document and specifies how each logical element in the document should be displayed.

XML

Extensible Markup Language (XML) is a meta-language. This means you can use it to define your own document structures and mark-up codes.

XML is a simple, very flexible text format derived from an earlier standard called SGML.

SGML was originally designed to meet the challenges of large-scale electronic publishing.

But XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere, particularly for electronic commerce.

XML

XML allows people and organizations to create their own mark-up languages specifically adapted to their needs and to the type of information produced.

Although everyone could create vocabularies for their own applications, in practice we usually prefer to **share our documents with other people** who have a common understanding of the descriptive mark-up in them.

The set of names used to tag the elements in an XML application is often referred to as an **XML Vocabulary**.

Experts have already created specific vocabularies for **applications**, such as mathematics or vector graphics.

They have also created vocabularies for market-specific information types such as equities research or aircraft maintenance.

More about XML vocabularies

XML

XML vocabularies have been created and agreed upon by organizations that want to **share information** in specific vertical industries (such as publishing, electronics, financial services, aerospace, etc).
Examples include the Docbook standard for technical publishers, the Business Reporting Markup Language (BRML) and the AECMA series of XML standards for the aerospace industry (http://www.aecma.org).

XML standards for business and e-commerce are being developed in the ebXML initiative (www.ebxml.org) and the Universal Business Language (UBL).

XML vocabularies have also been agreed upon for specific types of application.
For example, the next generation of HTML has been defined using an XML vocabulary (xhtml).
Other examples are the Mathematical Markup Language (MathML), the Scalable Vector Graphics language (SVG) and the Chemical Mark-up Language (CML).

XML

Literally thousands of XML vocabularies have been defined.

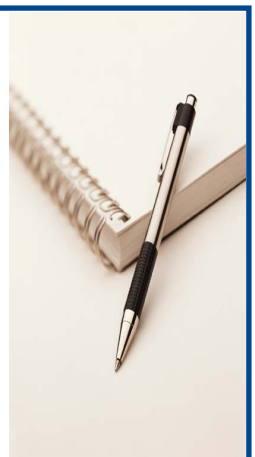
Some of the most important application vocabularies come from the World Wide Web Consortium, and an increasing number of vertical market vocabularies are being agreed upon using the standards process of OASIS – the Organisation for the Advancement of Structured Information Standards (www.oasis-open.org).

The figure shows a page from Robin Cover, which lists many of the vocabularies that have been defined since 1998.

You can access this list at: xml.coverpages.org

Summary

- **Mark-up** is everything in a document that is **not content**.
- **Procedural mark-up** are codes that contain information on how a specific application should process the document (example of procedural mark-up formats: Microsoft Word).
- **Presentational mark-up** are codes that describe how the document should be presented or laid out, either on a computer screen or on a printed page (example of presentational mark-up language: HTML).
- **Descriptive mark-up** are codes that describe the logical structure and semantics of a document, usually in a way that can be interpreted by many different software applications (example of descriptive mark-up meta-language: XML).
- **XML** is a meta-language that allows you to define your own document structures and mark-up languages.



Exercises

The following four exercises will allow you to test your understanding of the concepts covered in the lesson and provide you with feedback.

Good luck!

Exercise 1

In an electronic document, procedural mark-up is:

- ○ the text content of the document
- ○ a set of formatting codes
- ○ the description of the logical structure of a document

Click on your answer

Exercise 2

Which of the following is an example of descriptive mark-up?

```
<!DOCTYPE issue SYSTEM 'bandolier.dtd' []>
<?xml-stylesheet href="bandolier.xsl"
type="text/xsl"?>
<issue>
  <header>
    <volume>6</volume>
    <issue-no>5</issue-no>
    <issue-year>99</issue-year>
    <month>July</month>
    <bandolier>65</bandolier>
  </header>

  <editorial id="b65-1" clinical-code="123">
      <p>Bandolier this month has several
interesting articles on the subject
```

```
<body>
<h1>BANDOLIER</h1>
<h2>Editorial</h2>
<p>Bandolier this month has several interesting
articles on the subject
of pain.</p>
<h2>Oxford Pain Internet Site</h2>
<p>This month, on July 14th, <b>Bandolier</b>
is opening the Oxford Pain
Internet Site. The content is different from
the usual <b>Bandolier</b>
story which gives a view on a systematic review
or trial. Instead the
site provides formal summaries of systematic
reviews looking at pain
```

Click on your answer

Exercise 3

What are the main differences between XML and HTML?

XML

HTML

focuses on how the data looks

focuses on what the data is

was designed to describe data

was designed to display data

Click each option, drag it and drop it in the corresponding box.

*When you have finished, click on the **Confirm** button.*

Exercise 4

What does it mean that XML is a meta-language?

- ○ It provides standard ways of displaying a document in a web browser
- ○ It is information about the text of a document, rather then the text itself.
- ○ It allows the creation of personalized mark-up languages.

Click on the answer of your choice

If you want to know more...

World Wide Web Consortium (www.w3.org). Open information standards for the Web, including HTML and XML

OASIS – the Organisation for the Advancement of Structured Information Standards (www.oasis-open.org). Applications of open standards, including Docbook and UBL, the Universal Business Language.

ebXML (www.ebxml.org) - Electronic Business using eXtensible Markup Language

The Cover Pages (http://xml.coverpages.org) information about XML standards and vocabularies.