Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 2. ELECTRONIC DOCUMENTS AND FORMAT

LESSON 1. ELECTRONIC DOCUMENTS AND 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.

Learning Objectives

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

- define electronic documents;
- identify major electronic document formats;
- distinguish between different kinds of mark-up; and
- choose between various electronic document formats.

Content and electronic content

Imagine a historical museum consisting of objects such as portraits, old photos, documents, books, and many others.

These constitute the **objects of the collection** of that particular museum.

An object, for example a book or a picture, that is captured and converted into digital form becomes **digital content** and is potentially part of a digital library collection.

Content and electronic content

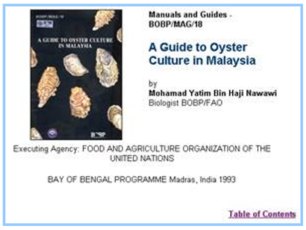
In the digital world, libraries are no longer confined to text based documentary resources; They have expanded to include **all types of content**, like old manuscripts, newspapers, images of all types, sound recordings, movies, data sets and other such material.

Today a wide variety of content is **created electronically** (e.g. mails, teaching materials, presentations, photographs, video clippings). In the same way, analog originals are converted into digital format.

Electronic documents

An electronic document is a digital representation of ideas or creative or intellectual works which are logically complete and can exist on their own as an independent unit of work. For example:

- text based documents such as books and journals, or
- multimedia objects, which include text, images and/or other representations such as audio and video.



Manuals and Guides - BOBP/MAG/18

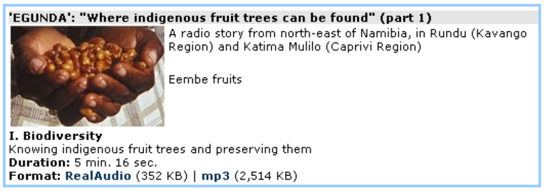
A Guide to Oyster Culture in Malaysia

by
Mohamad Yatim Bin Haji Nawawi
Biologist BOBP/FAO

Executing Agency: FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

BAY OF BENGAL PROGRAMME Madras, India 1993

Table of Contents



'EGUNDA': "Where indigenous fruit trees can be found" (part 1)

A radio story from north-east of Namibia, in Rundu (Kavango Region) and Katima Mulilo (Caprivi Region)

Eembe fruits

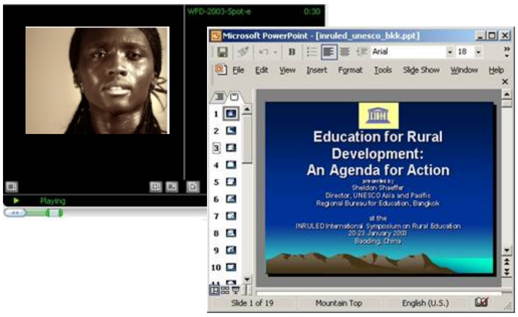
I. Biodiversity
Knowing indigenous fruit trees and preserving them
Duration: 5 min. 16 sec.
Format: RealAudio (352 KB) | mp3 (2,514 KB)

An electronic document could consist of a unit of work or it could be the complete work itself. Digital technologies enable us to increase the granularity and layers of electronic content. An entire book is an electronic document; a chapter of a book is also an electronic document. A picture embedded in a book is an electronic document.

Genre of Documents

It is important to understand the concept of genre of documents.

A genre refers to the medium of expression. Poems, a piece of prose, a novel, and dramas are all different mediums for expressing ideas. A skit, different dance forms, paintings, sculptures are also different genres of expressing ideas.



Education for Rural Development:
An Agenda for Action
presented by:
Sheldon Shaeffer
Director, UNESCO Asia and Pacific
Regional Bureau for Education, Bangkok

at the
INRULED International Symposium on Rural Education
20-23 January 2003
Baoding, China

Theses, journals and books are other examples of **genres of documents**.

Digital mediums, due to their versatility of expression including features such as multi media, hyper linking, animations, etc.., have given rise to new document genres. For example, Power Point presentations have become a new and popular genre of documents.

Document formats

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

We often **print** electronic documents in order to read them; This needs to be taken into account when creating them.

The use, usability and value of electronic documents depend on several factors in addition to the intrinsic value of the content.

Document **format** for text and images is one important factor.

Document formats

Document formats may be broadly grouped into the following categories:

Screenshot of a doc. opened with an open source word processor

Text-based formats: In these formats, content is largely textual. Text-based documents use a variety of mark-up codes to store, process and render documents.

Image formats: These are digital images of text pages, photographs, illustrations, artwork, and other graphical material

Audio and video formats: These are formats used for capturing, storing, processing and rendering audio and video presentations.

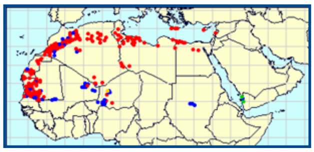
In this module, our focus is mainly on the creation of digital library collections whose content is predominantly textual and image documents.

Image formats

Image formats are mainly grouped into two types: Bitmap (or raster) and Vector formats.

Photographs or scanned images are usually saved in TIFF format; You can then them in other formats (e.g. GIF, JPG, PNG) which use compression techniques to reduce file size.

These images are called **Bitmaps**, as they are sampled and mapped as a grid of dots or picture elements represented in **binary code** (**bits**).



Vector based images come in the form of points and lines arranged on a grid.

They are directly created by using the computer. They can represent **cartoon-like drawings**, but are inappropriate for photo-realistic images.

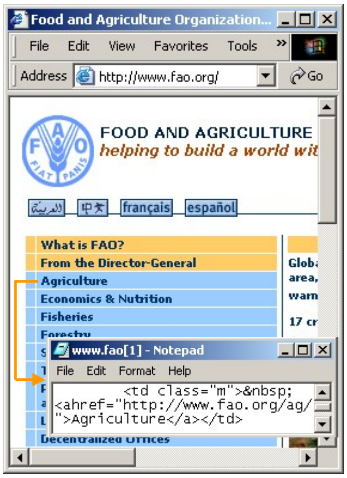
Text Formats and Mark-up

For text formats, the kind of **mark-up codes** used is a very important aspect.

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



```
        <td class="m">&nbsp;
<ahref="http://www.fao.org/ag/
">Agriculture</a></td>
```

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

- the text **content** of the document itself, and

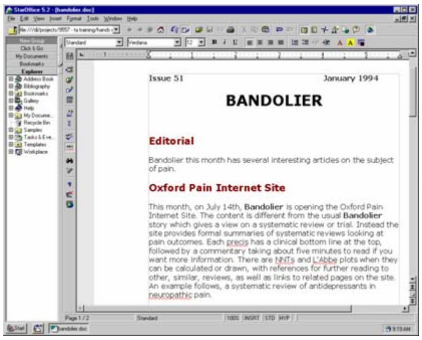
- a set of **codes** that provides information on how to display or interpret the text.

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.

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.



Issue 51 January 1994

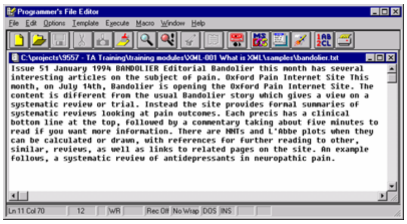
BANDOLIER

Editorial

Bandolier this month has several interesting articles on the subject of pain.

Oxford Pain Internet Site

This month, on July 14th, **Bandolier** is opening the Oxford Pain Internet Site. The content is different from the usual **Bandolier** story which gives a view on a systematic review or trial. Instead the site provides formal summaries of systematic reviews looking at pain outcomes. Each precis has a clinical bottom line at the top, followed by a commentary taking about five minutes to read if you want more information. There are NNTs and L'Abbe plots when they can be calculated or drawn, with references for further reading to other, similar, reviews, as well as links to related pages on the site. An example follows, a systematic review of antidepressants in neuropathic pain.


Programmer's File Editor
File Edit Options Template Execute Macro Window Help
C:\projects\9557 - TA Training\training modules\XML-001 What is XML\samples\bandolier.txt
Ln 11 Col 70
12
WR
Rec Off
No Wrap
DOS
INS

```
Issue 51 January 1994 BANDOLIER Editorial Bandolier this month has several
interesting articles on the subject of pain. Oxford Pain Internet Site This
month, on July 14th, Bandolier is opening the Oxford Pain Internet Site. The
content is different from the usual Bandolier story which gives a view on a
systematic review or trial. Instead the site provides formal summaries of
systematic reviews looking at pain outcomes. Each precis has a clinical
bottom line at the top, followed by a commentary taking about five minutes to
read if you want more information. There are NNTs and L'Abbe plots when they
can be calculated or drawn, with references for further reading to other,
similar, reviews, as well as links to related pages on the site. An example
follows, a systematic review of antidepressants in neuropathic pain.
```

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

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.

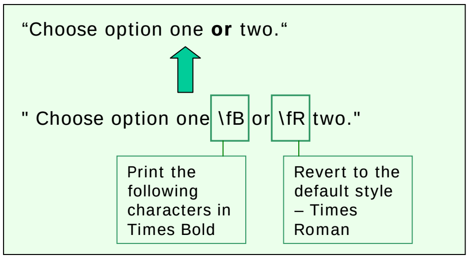
Descriptive mark-up consists of codes that describe the logical structure and semantics of a document, usually in a way that can be interpreted by many different software applications.

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

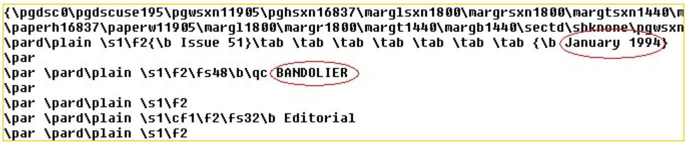
Different codes are attached to section headings, paragraphs of body text, references and even individual characters and words so that each is set in an appropriate type style, size and line spacing.

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?

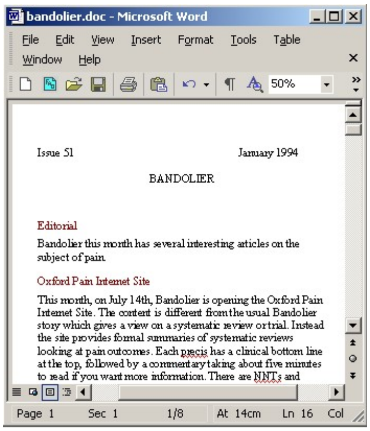


```
{\pgdsc0\pgdscuse195\pgwsxn11905\pghsxn16837\marglsxn1800\margrsxn1800\margtsxn1440\m
\paperh16837\paperw11905\margl1800\margr1800\margt1440\margb1440\sectd\sbknone\pgwsxn
\pard\plain \s1\f2{\b Issue 51}\tab \tab \tab \tab \tab \tab \tab {\b January 1994}
\par
\par \pard\plain \s1\f2\fs48\b\qc BANDOLIER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


bandolier.doc - Microsoft Word
File Edit View Insert Format Tools Table
Window Help
50%
Page 1 Sec 1 1/8 At 14cm Ln 16 Col

Issue 51 January 1994

BANDOLIER

Editorial

Bandolier this month has several interesting articles on the subject of pain.

Oxford Pain Internet Site

This month, on July 14th, Bandolier is opening the Oxford Pain Internet Site. The content is different from the usual Bandolier story which gives a view on a systematic review or trial. Instead the site provides formal summaries of systematic reviews looking at pain outcomes. Each precis has a clinical bottom line at the top, followed by a commentary taking about five minutes to read if you want more information. There are NNTs and

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

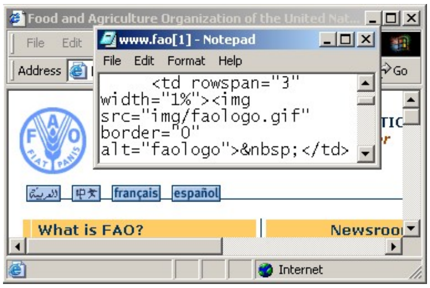
Non-proprietary word processors are also available, e.g. from the Open Office suite.

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).



```
    <td rowspan="3"
width="1%"><img
src="img/faologo.gif"
border="0"
alt="faologo">&nbsp;</td>
```

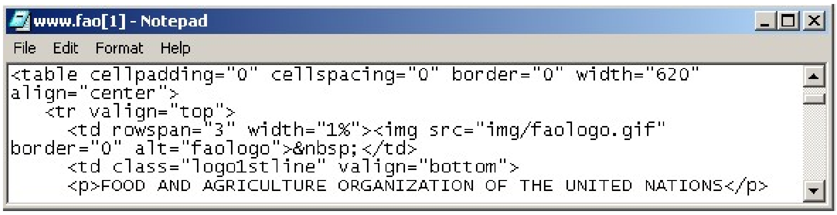
HTML is used to mark-up pages for presentation in a **web browser**.

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

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.



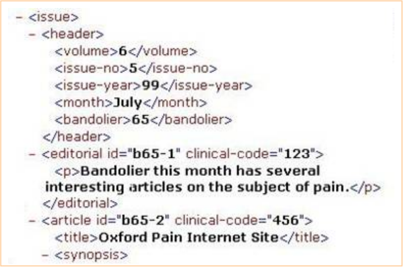
```
<table cellpadding="0" cellspacing="0" border="0" width="620"
align="center">
   <tr valign="top">
      <td rowspan="3" width="1%"><img src="img/faologo.gif"
border="0" alt="faologo">&nbsp;</td>
      <td class="logo1stline" valign="bottom">
      <p>FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS</p>
```

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**.

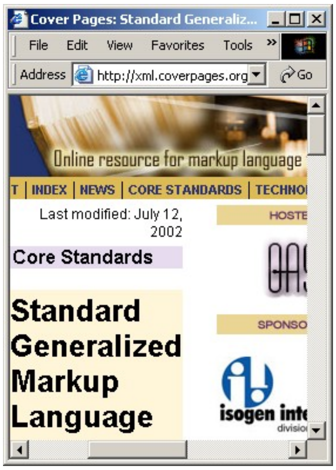


```
- <issue>
  - <header>
      <volume>6</volume>
      <issue-no>5</issue-no>
      <issue-year>99</issue-year>
      <month>July</month>
      <bandolier>65</bandolier>
    </header>
  - <editorial id="b65-1" clinical-code="123">
      <p>Bandolier this month has several
     interesting articles 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


Cover Pages: Standard Generaliz...
File
Edit
View
Favorites
Tools
Address
http://xml.coverpages.org
Go
Online resource for markup language
T | INDEX | NEWS | CORE STANDARDS | TECHNO
HOSTE
SPONSO
isogen inte

Last modified: July 12, 2002

Core Standards

Standard Generalized Markup Language

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

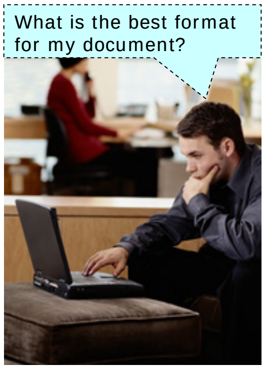
It is the most prevalent form of descriptive mark-up in use today, and is a standard of the World Wide Web Consortium.

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.

Choosing among different formats



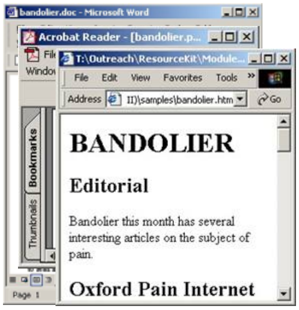
Christian, a member of the Food Security department in his organization, has to write a research document on desertification.
The document will then be distributed to other members of the Department, and to other Departments in the organization.

How should he choose the best format for the document? Should it be an HTML page, a Word document or something else?

Choosing among different formats

The same document can have different **renditions**, that is different formats, each one with the related mark-up codes.

Different renditions of a document can be useful when the document is used in **several scenarios**. For example:


bandolier.doc - Microsoft Word
Acrobat Reader - [bandolier.p...
T:\Outreach\ResourceKit\Module...
File Edit View Favorites Tools
Address II)\samples\bandolier.htm
Go
Bookmarks
Thumbnails
Page 1

BANDOLIER

Editorial

Bandolier this month has several interesting articles on the subject of pain.

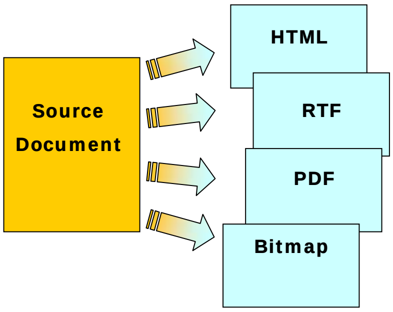
Oxford Pain Internet

- a rendition in a **word processing** format, such as Microsoft Word, is useful when **creating or editing** the document,
- an **HTML** rendition is useful when viewing it on the **Web**, and
- a **page rendition** as a bitmap graphic or PDF format may be useful when a **read-only** page layout view is required.

View and print the document format comparison table:

Table of formats

Choosing among different formats



Source Document

HTML

RTF

PDF

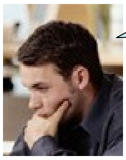
Bitmap

When different renditions are used for a document, it is important to keep a single **source document**, so that updates and changes are made in that document, before it is transformed into different formats.

But, what should the format be for the source document?

Choosing among different formats

Which of the following formats would you recommend to Christian?



I have to create a printable document that can be displayed on the Web. Which format should I choose for the source document?

- Microsoft Word format
- HTML format
- Bitmap format

Please click on the answer of your choice

Choosing among different formats

Document formats have many characteristics that should be considered when deciding on the suitability of any format. For example:

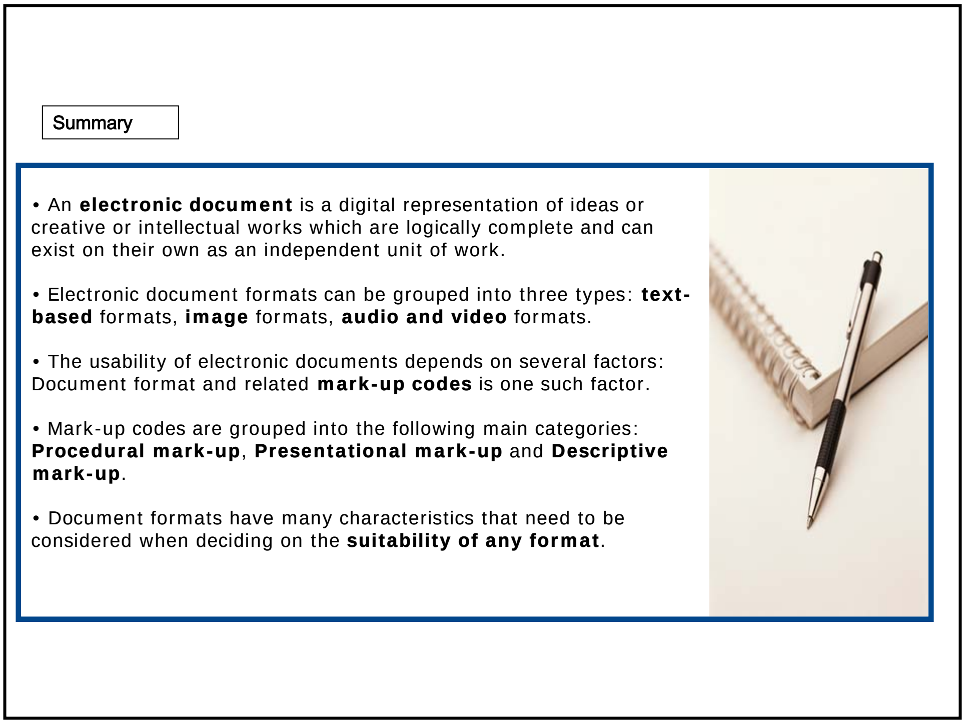
- file size,
- ease of creation,
- convenience of managing them, including archiving, accessing and display.

The tools and technologies required to handle different formats are also important considerations, as well as whether they are proprietary or open source formats.

In the following lessons, you will explore other widely used formats and their characteristics.

Selecting Electronic Document Formats

(http://www.ifla.org/VI/5/op/udtop11/udtop11.htm)



Summary

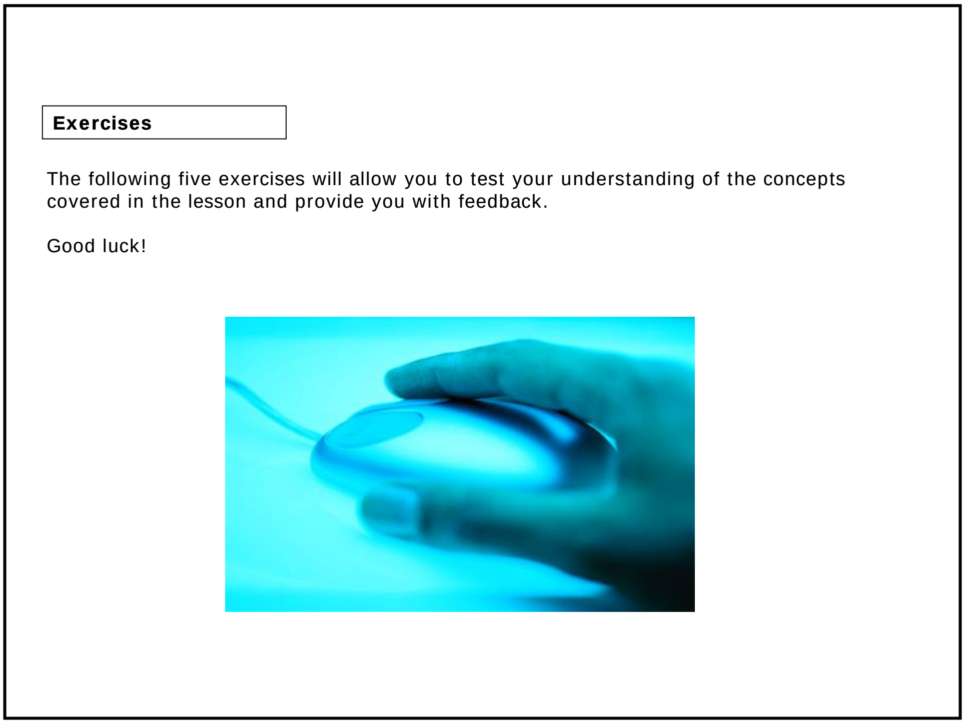
- An **electronic document** is a digital representation of ideas or creative or intellectual works which are logically complete and can exist on their own as an independent unit of work.

- Electronic document formats can be grouped into three types: **text-based** formats, **image** formats, **audio and video** formats.

- The usability of electronic documents depends on several factors: Document format and related **mark-up codes** is one such factor.

- Mark-up codes are grouped into the following main categories: **Procedural mark-up**, **Presentational mark-up** and **Descriptive mark-up**.

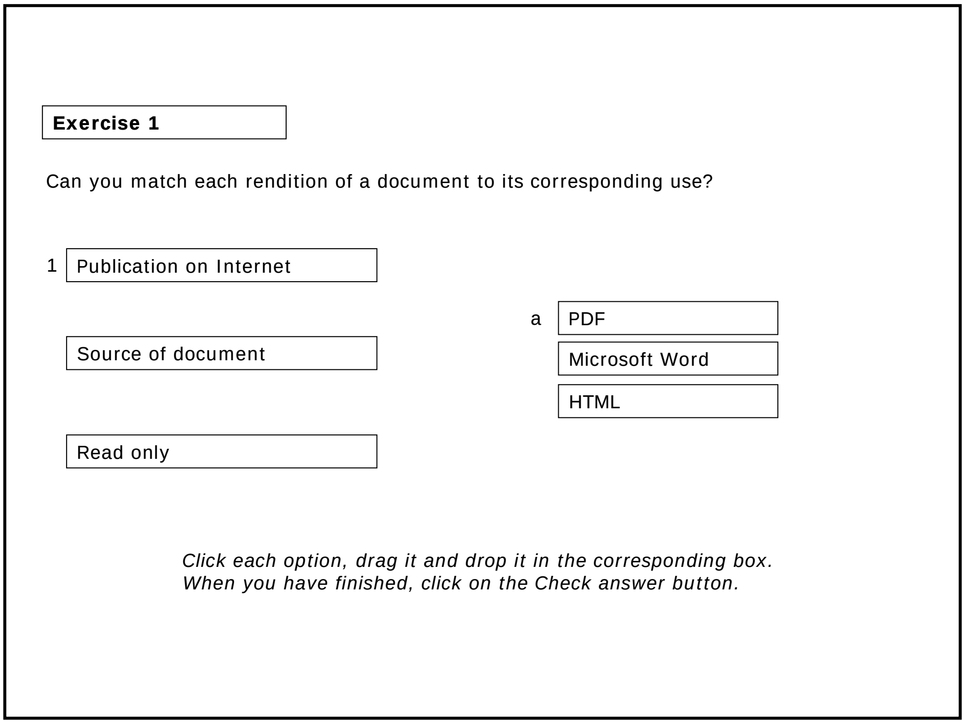
- Document formats have many characteristics that need to be considered when deciding on the **suitability of any format**.



Exercises

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

Good luck!



Exercise 1

Can you match each rendition of a document to its corresponding use?

1 Publication on Internet

Source of document

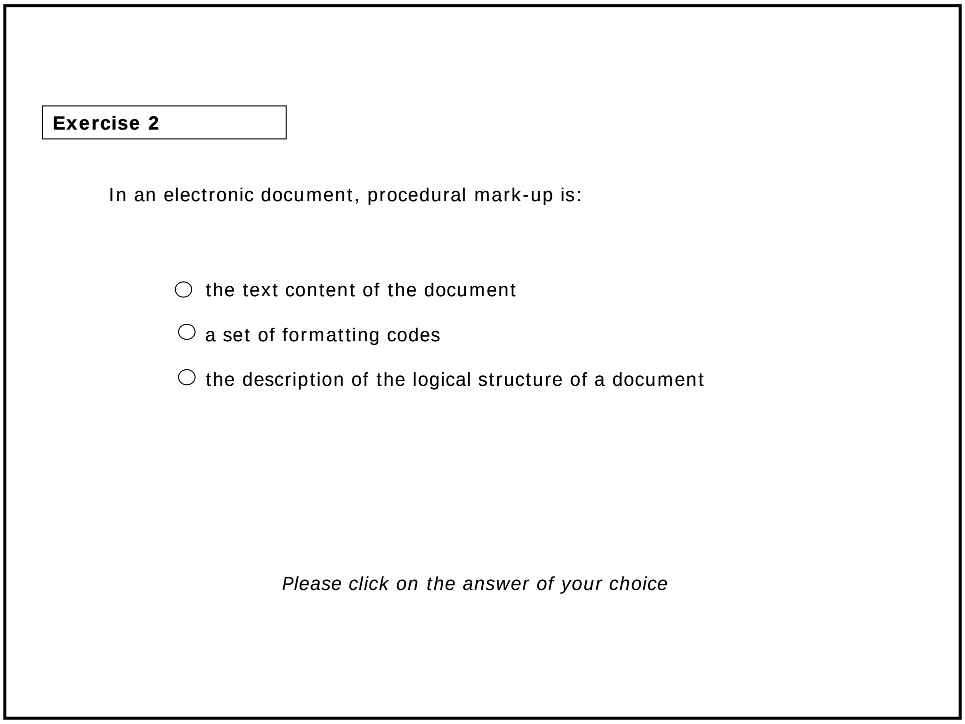
Read only

a PDF

Microsoft Word

HTML

Click each option, drag it and drop it in the corresponding box.
When you have finished, click on the Check answer button.



Exercise 2

In an electronic document, procedural mark-up is:

○ the text content of the document

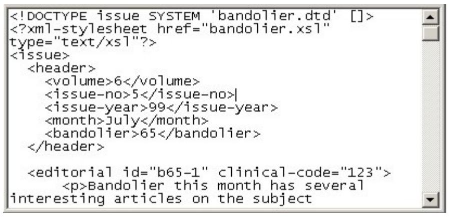
○ a set of formatting codes

○ the description of the logical structure of a document

Please click on the answer of your choice

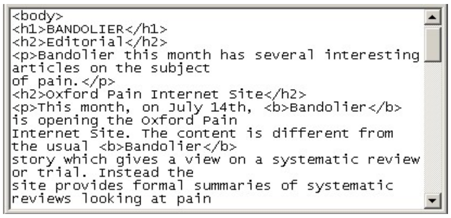
Exercise 3

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
```

Please click on the answer of your choice

Exercise 4

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 Check answer button.

Exercise 5

Why is XML called 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.

Please click on the answer of your choice

If you want to know more...

Online Resources:

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

OpenOffice.org - OpenOffice is an open source (free) suite of software available in various languages which includes a word processor, spreadsheet, presentation and drawing software with PDF capabilities:(http://www.openoffice.org/)

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

ebXML - an open XML-based infrastructure enabling the interchange of electronic business information globally: (http://www.ebxml.org)

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

Selecting Electronic Document Formats, by Gary Cleveland, IFLA UDT Programme (August 1999); A document examining the characteristics of some of the most popular electronic text and image formats and providing a rough guide for their selection. It includes a table presenting a summary of electronic document format characteristics : (http://www.ifla.org/VI/5/op/udtop11/udtop11.htm)