Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 4. CREATION AND MANAGEMENT OF DIGITAL DOCUMENTS

LESSON 5. HANDLING BORN-DIGITAL DOCUMENTS

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 be able to:

- distinguish the different **steps for the production** of **born-digital documents**; and
- identify the **requirements** for, and the **options** you have for structuring a workflow for producing born-digital documents to be added to the digital library

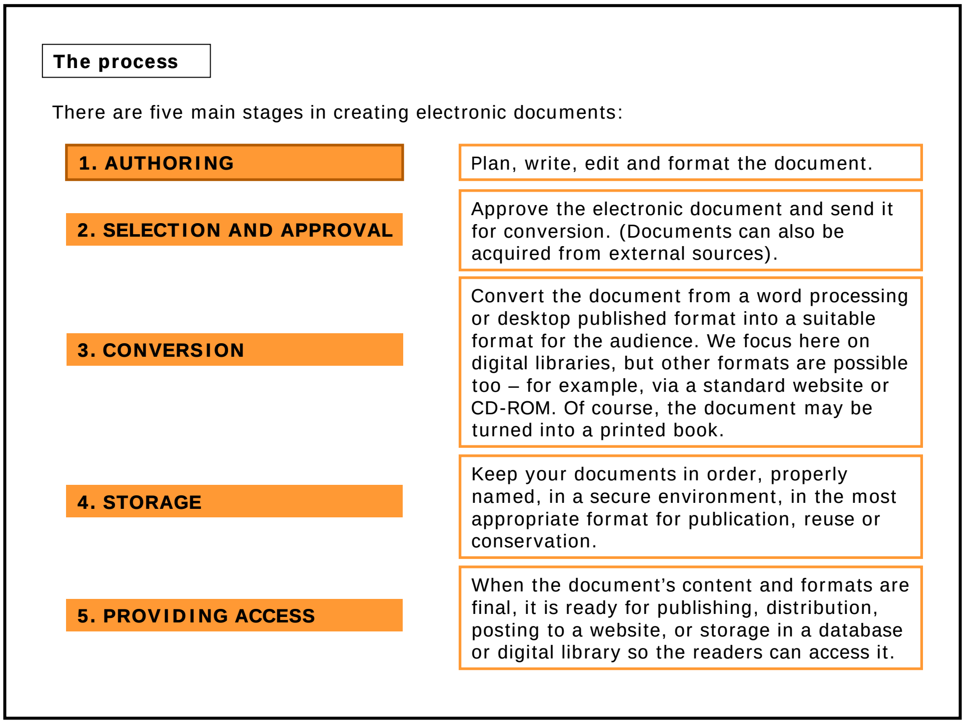
Introduction

Ms Lee, Publications Committee member

Ms. Lee works for the **Publications Committee** of her organization.

The three research teams in the organization are quite active and produce several reports and research articles each year.

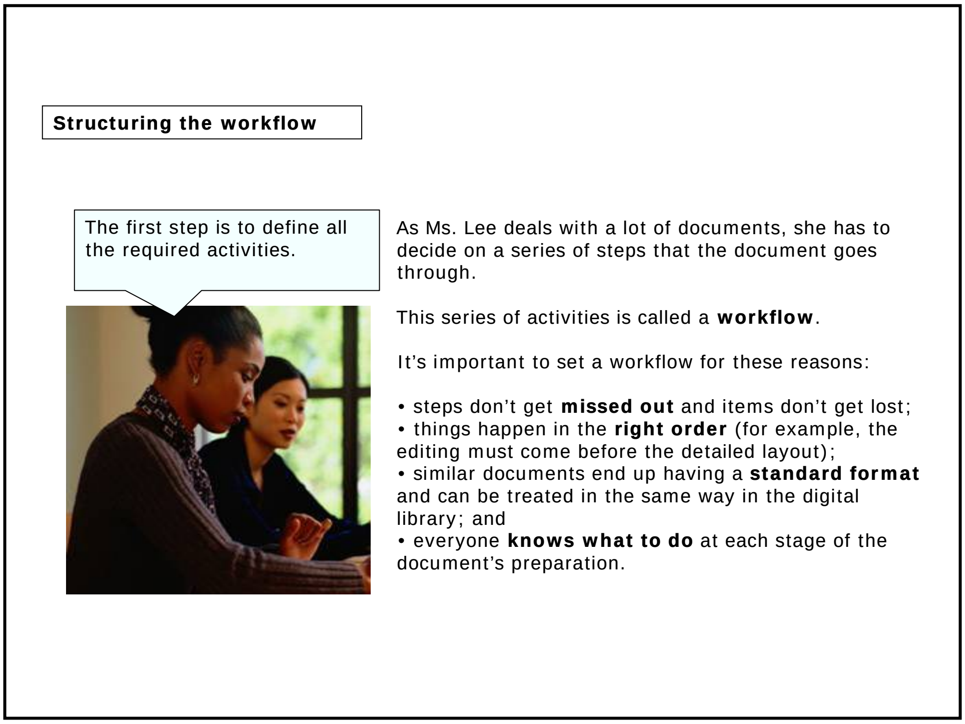
The Publications Committee has decided to build a digital library to make the documents created by these teams accessible.



The process

There are five main stages in creating electronic documents:

1. AUTHORING	Plan, write, edit and format the document.
2. SELECTION AND APPROVAL	Approve the electronic document and send it for conversion. (Documents can also be acquired from external sources).
3. CONVERSION	Convert the document from a word processing or desktop published format into a suitable format for the audience. We focus here on digital libraries, but other formats are possible too – for example, via a standard website or CD-ROM. Of course, the document may be turned into a printed book.
4. STORAGE	Keep your documents in order, properly named, in a secure environment, in the most appropriate format for publication, reuse or conservation.
5. PROVIDING ACCESS	When the document's content and formats are final, it is ready for publishing, distribution, posting to a website, or storage in a database or digital library so the readers can access it.



Structuring the workflow

The first step is to define all the required activities.

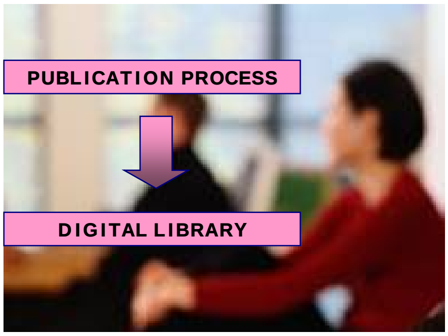
As Ms. Lee deals with a lot of documents, she has to decide on a series of steps that the document goes through.

This series of activities is called a **workflow**.

It's important to set a workflow for these reasons:

- steps don't get **missed out** and items don't get lost;
- things happen in the **right order** (for example, the editing must come before the detailed layout);
- similar documents end up having a **standard format** and can be treated in the same way in the digital library; and
- everyone **knows what to do** at each stage of the document's preparation.

Structuring the workflow



PUBLICATION PROCESS

DIGITAL LIBRARY

Ms Lee has to consider the needs of the **digital library** when designing the publications workflow.

A digital library normally (but not always) comes **at the end** of the publishing process. It contains documents that are completed, published and available for readers.

The task of Ms Lee is to plan and coordinate the **preparation of electronic documents** to be included in the digital library.

Structuring the workflow

Preparing documents for a digital library imposes certain **requirements**.

Defining them **early on** makes all stages in the publishing process easier and more efficient.

We should define some requirements for our documents. For example, each document must have appropriate **metadata**, like the filename, title and author…

Structuring the workflow

Some things to think about when defining a workflow:

DOCUMENT STANDARDS → How are documents structured and formatted? You may have to develop standards for **templates and styles**, **formats** and **metadata**. Make sure that standards are consistently applied.

SOFTWARE → Which software best helps you to apply the standards? Some programs can do their job right away (e.g., authors may use Microsoft Word just because it is widely used). You may have to **customize** or build other tools yourself so they fit your requirements.

KEY ROLES → Who does what? Authors, publications officers, information systems officers, librarians and web administrators are among the key roles your staff will play in the workflow.

Once you have established standards, tools and goals, it is possible to identify tasks and procedures and assign them to the roles needed to implement the workflow.

Checklist for structuring a workflow

Templates and styles

I want to make sure that the researchers are able to provide useful information in a readily accessible format!

1. AUTHORING

2. SELECTION AND APPROVAL

3. CONVERSION

4. STORAGE

5. PROVIDING ACCESS

Authoring includes planning, writing, editing and formatting the document.

Templates can help in the authoring process. They can help:

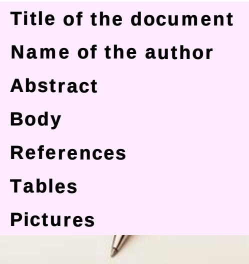
- **structure the document** in a logical, consistent way;

- give the document a consistent appearance through **styles**; and

- **assign metadata** to the document.

You may be able to require authors to **submit manuscripts** using a certain template and styles. You may have to teach them how to use the styles and templates. This is a lot of work, but it is a good investment!

Templates and styles



Title of the document

Name of the author

Abstract

Body

References

Tables

Pictures

This is the structure of a journal article.

Different types of documents have different structures. For example:

- A **book** contains *chapters*, which in turn contain *headings*, *subheadings*, *paragraphs*, *tables*, *figures* and *captions*.

- A **journal article** usually has the following sections: *title*, *author*, *abstract*, *body* (with headings and subheadings), *tables* (with a title, row and column headings, and body), *figures* (with captions and labels), and *references*.

- A **newspaper article** usually has the following: *headline*, *dateline*, *byline*, *text*, *pictures*.

Each element in the document contains a certain type of information, and it is related in a logical, hierarchical way to the other elements. It should also have its own formatting style.

Templates and styles

Assigning metadata

Metadata include information on the title, author, subject, filename, etc. of a document.

They are needed so the digital library can identify, manage and sort documents.

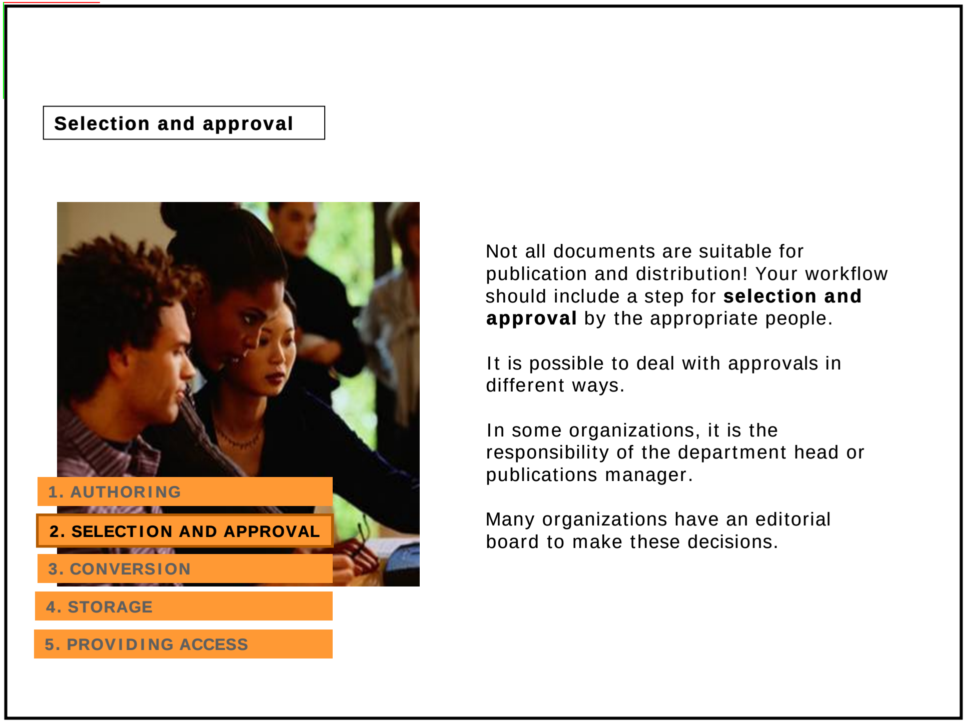
Different types of metadata are appropriate for different types of documents. For example:

- **Memos** may be given the metadata *To*, *From*, *Date* and *Subject*.

- **Journal articles** may be given the metadata *Author(s)*, *Date*, *Title*, *Journal Title*, *Journal Issue*, *Page Numbers* and *Language*. They may also have metadata on *Subject* and *Abstract*.

- **Books** may be given the metadata *Author(s), Date*, *Title*, *Publisher*, *Publication City*, *Number of Pages* and *Language*. They may also have metadata on *Subject*, *Abstract*, *Table of Contents*, *Series Name*, *Series Editor*, and so on.

How to assign metadata



Selection and approval

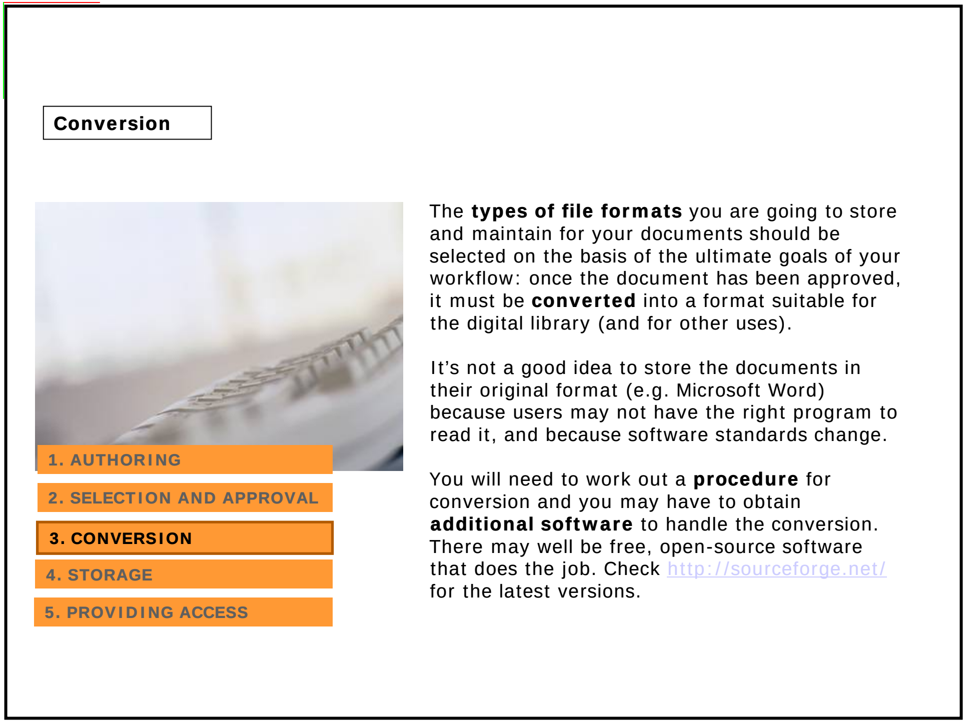
1. AUTHORING
2. **SELECTION AND APPROVAL**
3. CONVERSION
4. STORAGE
5. PROVIDING ACCESS

Not all documents are suitable for publication and distribution! Your workflow should include a step for **selection and approval** by the appropriate people.

It is possible to deal with approvals in different ways.

In some organizations, it is the responsibility of the department head or publications manager.

Many organizations have an editorial board to make these decisions.



Conversion

1. AUTHORING
2. SELECTION AND APPROVAL
3. **CONVERSION**
4. STORAGE
5. PROVIDING ACCESS

The **types of file formats** you are going to store and maintain for your documents should be selected on the basis of the ultimate goals of your workflow: once the document has been approved, it must be **converted** into a format suitable for the digital library (and for other uses).

It's not a good idea to store the documents in their original format (e.g. Microsoft Word) because users may not have the right program to read it, and because software standards change.

You will need to work out a **procedure** for conversion and you may have to obtain **additional software** to handle the conversion. There may well be free, open-source software that does the job. Check http://sourceforge.net/ for the latest versions.

Conversion

Choose the file format on the basis of the ultimate goals of your workflow.

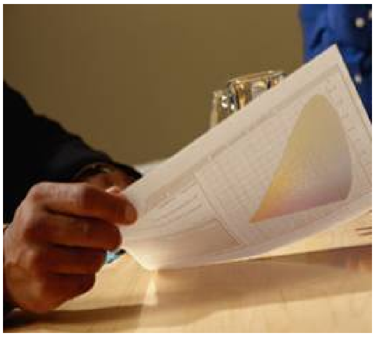
If your goal is to…	You should…	Suggested format	
		Text documents	Graphics
Preserve content, look and feel of documents	Choose a software-independent format to ensure the document can be retrieved exactly as intended over time, and regardless of the software used to create it.	PDF, XML	BMP, TIFF, GIF, JPG, PNG, EPS
Reuse the documents or their components	Select a format that: - allows you to specify the level of detail you want to capture (the “granularity”), and - gives you most flexibility in transforming the document into other formats.	XML	BMP, TIFF, PNG, EPS
Providing access to documents	Select a format that enables your end users to access the content. This means using available software such as web-browsers (Internet Explorer, Netscape) and free plug-ins (Adobe Reader).	PDF, HTML, XML	GIF, JPG, PNG

The table below reviews the suitability of different text and image file formats for the three purposes:

Table of storage formats for documents and images

Conversion

You may want to convert the document into **several different formats**, depending on their use.



For example:

• You may want to enable users to access a **compressed** (but low-quality) PDF file, but keep a **higher quality** (but larger) PDF in case you need to republish the printed document or for preservation purposes.

• You may want to provide users with the option of downloading **several smaller** PDF files of different chapters, or a **single larger** file containing a whole book. This option may be particularly useful in developing countries, where internet access is relatively costly and is interrupted frequently.

Storage

1. AUTHORING
2. SELECTION AND APPROVAL
3. CONVERSION
4. STORAGE
5. PROVIDING ACCESS

Storage means keeping your documents in order, properly named, and in a secure environment.

- You need to **name** files in a logical consistent way, so you, others, and the digital library software can find them.
- Store your files on a **secure** computer – where they cannot be accidentally deleted or tampered with.
- Keep them in a separate set of **directories** so they can be accessed easily and as a group.
- Keep **backup** copies in case disaster happens – a hard disk crash, virus attack, computer theft, or fire. CDs or DVDs can be used for this purpose.

In a larger organization you may conduct all the various stages of the workflow on a file server accessible through a LAN which is backed-up on a regular basis and accessible to all people involved.

Storage

Use a **standard system** for naming files:

- to keep track of **versions** and **translations** of a document and its components (such as graphics) as it undergoes the authoring, editing, publication and translation process; and
- to **store and access** the document once it is in the digital library.

Once you have a suitable filenaming convention, use it to **rename** existing files. Make sure that new files are given suitable names following your rules.

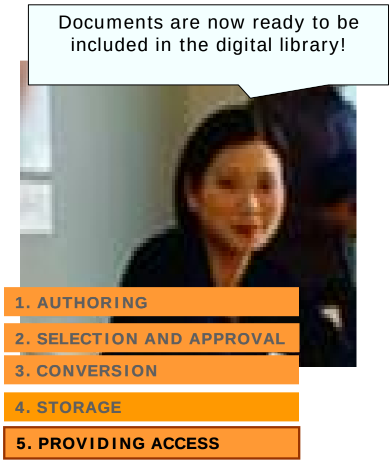
Once files have been given suitable names, avoid renaming them again, as this can break hyperlinks and cause files to be lost.

For your digital library, you may need to rename **pictures** to retain their association with text files. For example, in Greenstone, the picture of a document cover must have the same filename as the document text's HTML file for Greenstone to display it correctly.

SOME RULES FOR FILENAMES

Avoid **accented** characters (such as *à* and *é*)
Avoid **punctuation** and certain other characters (*; “ @ % /,* etc)
Use underscores instead of **spaces** (*my_file.pdf* instead of *my file.pdf*)
Use all **lowercase** letters (*another_file.pdf*, not *Another_File.pdf*)
Avoid very **long** filenames.
For **dates**, use the format yyyymmdd (so 13 April 2004 is 20040413). This is unambiguous, and means that dates are automatically in the right chronological order.

Providing access to documents


Documents are now ready to be included in the digital library!

1. AUTHORING
2. SELECTION AND APPROVAL
3. CONVERSION
4. STORAGE
5. PROVIDING ACCESS

Once you have all the files in the correct format, and you have developed all the metadata you need, you are ready to include documents in the digital library.

After adding documents and metadata to the digital library, the Web administrator will update the online version of the digital library, or the CD-ROM.

How to perform these tasks will depend on the specific digital library software you have chosen.

Maintaining the workflow

The table below shows an example of how the workflow might work in practice.

Stage	Possible activities and roles
AUTHORING	**Publications Committee** plans publication. **Authors** write draft using templates and provide metadata. **Editor** edits draft. **Designer** prepares graphics and formats output.
SELECTION AND APPROVAL	**Publications Committee** approves publication.
CONVERSION	**Editor** converts publication to PDF format. **Librarian** checks metadata and prepares metadata file.
STORAGE	**Editor/ Librarian** stores and prepares the electronic documents for addition to the digital library.
PROVIDING ACCESS	**Web/ Digital library administrator** updates online version of digital library or updates the CD-ROM.

Now let’s give you some tips on how to maintain it...

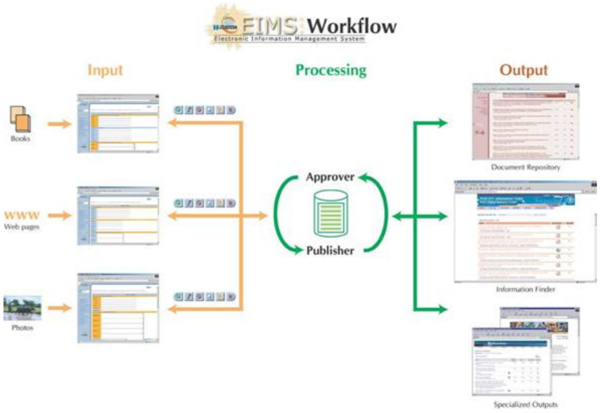
Maintaining the workflow

Document and workflow management systems can be used for providing access to documents to end users and thus cover the whole document workflow. How much of it is supported by the systems depends on the specific requirements and needs of your group and organization.

Level of complexity	How to maintain the workflow
The workflow is **simple**: - 1-2 authors - 1 editor/approver - 1 producer/Webmaster	Provide written **guidelines and policies** on templates, formats, conversion options and file naming conventions. Make sure they are always up-to-date and circulate changes.
The workflow is **more complex**: - Multiple authors inside and outside the organization - Multiple levels of approval (e.g. for content, expenses, translation) - Parallel or subsequent processing procedures for different output formats	You should consider adopting a **workflow management system** to keep track of the status of documents through metadata (e.g. owner, language, review stage, approval stage, etc.) and to assign roles and rights to team members (e.g. author, approver, producer). If you also need to control versions and access to documents, you should consider a **content management system** with workflow management capabilities.

Example of a workflow

The **FAO Document Repository** is a large system to store and disseminate FAO's documents and publications in electronic formats.



An electronic publishing workflow, managed through FAO's **Electronic Information Management System**, controls the flow of documents into the Repository.

This Electronic Information Management System tracks the cycle of a publication throughout the stages of creation, translation, conversion and publishing.

Guidelines and procedures

Here you can download and print the documents provided in this lesson.

You may use them as tools for your job.

Checklist for structuring workflow

Templates and styles

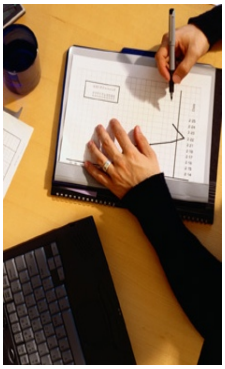
How to assign metadata

Table of storage formats for documents and images

The workflow we have seen is related to new documents produced by your organization. Existing documents may also need some processing before they can be put into a digital library.

Here are some things you may need to do for them:

Handling existing electronic copy



Summary

- When planning a workflow for producing and processing future documents, think of your **goals**, document and conversion **standards**, **tools** and organizational **roles**.
- **Structured templates** that use **styles** facilitate conversion, storage and access to documents.
- **Metadata** are cataloguing information about a document: the author, title, content, etc. They can be included in the document properties, in the document text, stored in a separate file, or added through an onscreen form.
- **Preservation, reuse** and **access** set the priorities for deciding on which formats should be used to create and store documents.
- **File naming conventions** ensure that files are named consistently and can be identified and located easily.

Exercises

The following five exercises will allow you to test your understanding of the concepts described up to now.

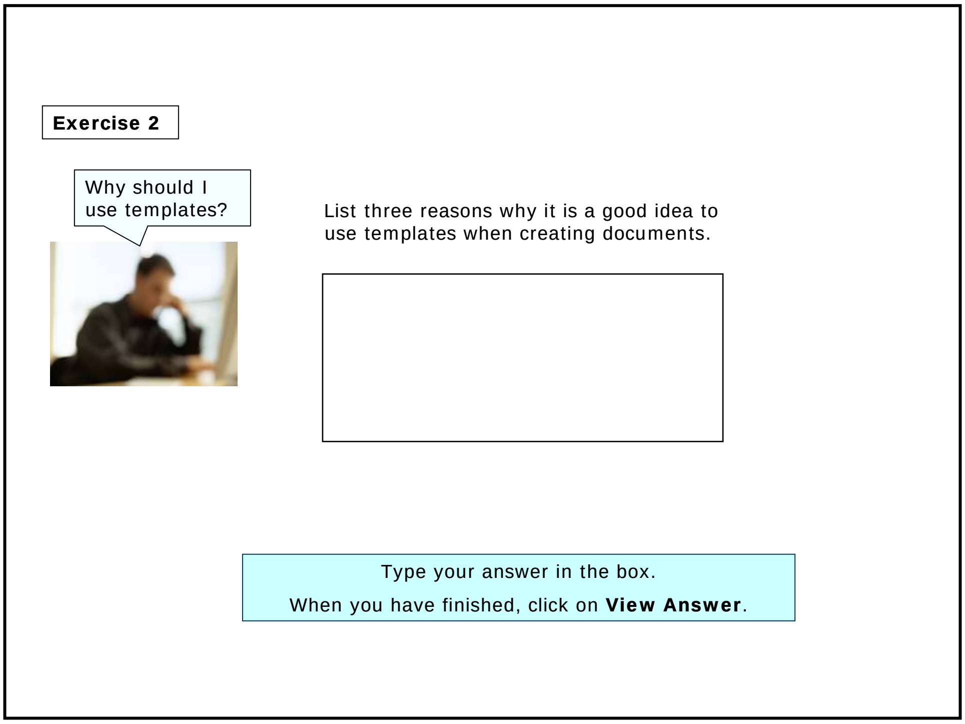
Good luck!

Exercise 1

Put the five steps in the publishing process into the correct order.

1	a
STORAGE	
AUTHORING	
CONVERSION	
PROVIDING ACCESS	
SELECTION AND APPROVAL	

Click on each option and drag it into the corresponding box.
When you have finished, click on the Confirm button.



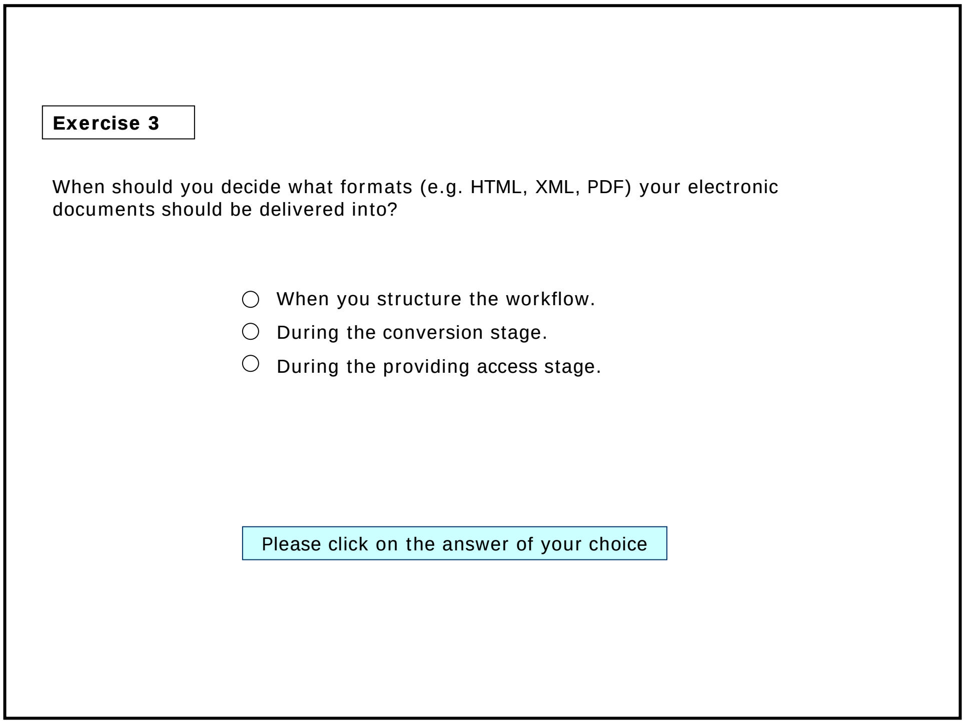
Exercise 2

Why should I use templates?

List three reasons why it is a good idea to use templates when creating documents.

Type your answer in the box.

When you have finished, click on **View Answer**.



Exercise 3

When should you decide what formats (e.g. HTML, XML, PDF) your electronic documents should be delivered into?

○ When you structure the workflow.

○ During the conversion stage.

○ During the providing access stage.

Please click on the answer of your choice

Exercise 4

You want to make sure you have the most flexibility in transforming the document into other formats.

Which of the following formats would you select?

- □ PDF
- □ XML
- □ HTML
- □ GIF
- □ TIFF

Please select the options of your choice (2 or more) and press “Check Answer”

Exercise 5

Read this file name:

How to set up_standard, guidelines-3 / 2 / 2003.doc

How could you rewrite it in an easily understood and compatible way?

- ○ how_to_set_standard_guidelines_20030203.doc
- ○ how to set standard guidelines_20030203.doc
- ○ guidelines_feb032003.doc

Please click on the answer of your choice

If you want to know more...

Online Resources:

Ins and Outs of Word Templates:
(http://www.pcworld.com/howto/article/0,aid,64164,00.asp)

Tips for Understanding Styles in Word:
(http://www.microsoft.com/office/using/column14.asp)

Free software for handling conversion: (http://sourceforge.net/)

For more information on content management systems:

The difference between document and content management, by Bob Boiko/Metatorial Services Inc., (http://metatorial.com/Papers/dm_v_cm.asp)

Your clients need a Content Management System, by Martin Burns
(http://www.evolt.org/article/MartinB/20/5127/)

How to evaluate a content management system, by James Robertson
(http://www.steptwo.com.au/papers/kmc_evaluate/)

Additional Reading:

Andrew Hampson et al. Digitisation of exam papers. The Electronic Library, 17,4; Aug. 1999; 239-46. Discusses complete workflow, project planning and management for digitizing and providing intranet access to exam papers.