Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 4. WORKFLOWS

LESSON 3. CREATION AND PROCESSING OF ELECTRONIC FILES

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 be able to:

- understand the **usefulness of a workflow** for creating, processing and delivering documents on different media;
- distinguish the different **steps of electronic production and management** of documents; and
- identify the **requirements and options** you have in structuring your workflow.

Introduction

Ms. Lee is in charge of publications in her organization.

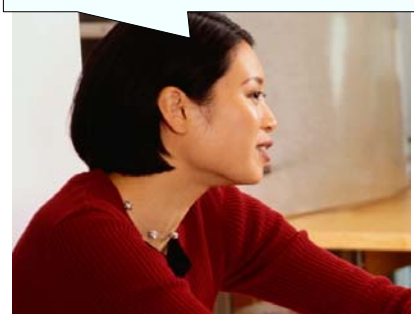
She publishes reports and results from three different research teams.

The teams are quite active and produce several reports and research papers each year.

She has to collect and publish the reports and documents on her organization's website, as well as in hard copy and for e-mail distribution.

Introduction

The current process is mainly meant for printing: this involves a lot of work when we have to convert documents into formats that are more suitable for the Internet or e-mail.



Ms. Lee noticed that, as publication of electronic format documents increases, the process she follows for creating and delivering documents is becoming obsolete.

In fact, it is mainly for delivering documents in print and unlikely to favour electronic dissemination.

What Ms. Lee needs is a new process **designed from the start** to disseminate documents **through both electronic and printed media**.

The process

The process for creating documents **to be disseminated through both electronic and print media** goes through five main stages:

Click on each stage to see the description

1. AUTHORING

Documents are planned, authored and edited in a format that facilitates conversion for electronic and print media.

2. SELECTION AND APPROVAL

Documents are approved and sent for conversion. They can also be acquired from external sources.

3. CONVERSION

Documents are converted into the formats appropriate for delivery on the media you have selected to best reach your audience: a website, a CD-ROM or a print book.

The process

4. STORAGE

More of a concept than an activity, storage means keeping 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 content and formats are final, documents are published, distributed, posted to a website or stored in a database for the intended audience to access them.

Structuring the workflow

OK, the phases of the process are quite clear. Now, we must define all the steps. Moreover: how will we coordinate the work and the people involved in it?

Before starting the process, you should think about structuring the **workflow**.

A workflow can be defined as number of tasks performed in sequence or in parallel by two or more members of a workgroup to reach a common goal.

A workflow can be simple or complex depending on your organization's needs and the type of audience you are targeting.

There are some questions you should ask yourself to **identify the goal** of your electronic document workflow. Let's look at them…

Structuring the workflow

Answering the following questions helps you identify the objectives that your workflow should be supporting.

•What is the **final output** you want to get out of the process (e.g.: print-only publication, CD-Rom based collection, etc.)?

• In what **file formats** should you store your documents (e.g.: Word, PDF, XML, etc.)?

• What kind of **infrastructure** do you have in place to store your documents (file system or database)?

• How do you provide your audience with **access to documents** (e.g.: Library, Website, etc.)?

• Do you plan to **reuse your documents** in future publications or on different media?

Structuring the workflow

Having identified your workflow objectives, you have to define:

DOCUMENT STANDARDS

If you want to automate part of your workflow, you have to make sure that standards (**e.g. templates, metadata, formats for texts and images**) are consistently applied. Otherwise, a lot of manual work has to be done in order to make a document compliant to your standards!

TOOLS

You need to identify the tools that best help you **to apply the standards**. Some standard tools can be used for the job (e.g., authors may use Microsoft Word just because it is widely used). Other tools have to be customised or built to fit your requirements.

KEY ROLES

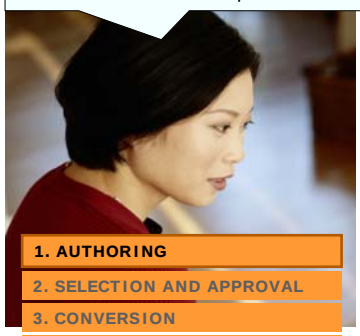
Authors, publications officers, information systems officers, librarians and Webmasters are among the key roles your staff will play in the workflow. Note that **roles do not necessarily correspond to the same number of staff members**: if you have simple needs, one person could play all roles.

Once standards, tools and goals have been established, tasks and procedures can be identified and assigned to the roles needed to implement the workflow.

Checklist for structuring a workflow

Using templates

I want to make sure that the researchers are going to provide useful information and that the conversion will be as fast and smooth as possible!



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

For the authoring stage, Ms. Lee needs a set of **document standards** for her organization that can be reused over time to create the same type of documents.

Here is how she can define her set of standards:

a) **Structure the document**
b) **Assign styles**
c) **Create a template**

Let's follow these steps using, for example, Microsoft Word.

Using templates

a) Structure the document

Name of the organization
Title of the document
Date of the meeting
Participants
Account of each discussed topic

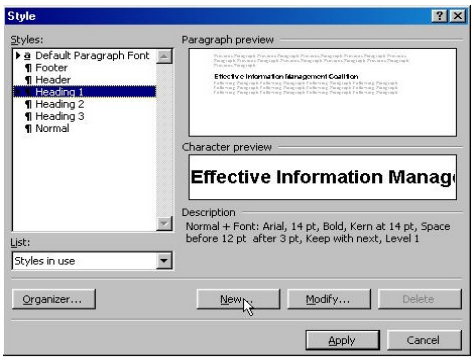
These are the contents needed for a meeting report.

Structuring a document means **identifying each part of the text (a block) as part of a structure** where each block is supposed to hold information that is related in a logical, hierarchical way to other blocks in the document.

For example, a book can have chapters which contain paragraphs, which in turn contain tables and captions for figures.

Using templates

b) Assign styles to each block



A formatting style should be assigned to mark the different blocks in order to facilitate the next stages in the process.

Look at this example: every time you assign the "Heading 1" style to the chapter-level headings, these will mark the chapter blocks of your document.

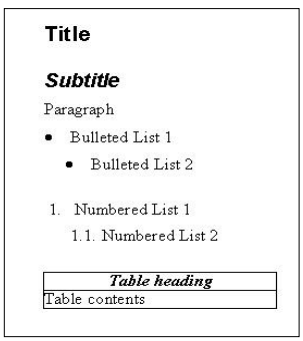
Choose styles carefully and assign them to your document blocks consistently: when you convert your document to HTML or XML, styles tell the conversion tool which HTML or XML elements should be used to correctly convert your document and preserve its structure.

This is a good investment at the authoring phase!

Consistent application of styles can be good for authoring as well. In Microsoft Word, you can build tables of contents quickly based on heading styles, or browse your document with the Document Map. If you need to create a PDF, bookmarks to the main sections marked with heading styles can be built automatically so readers can quickly browse your document.

Using templates

c) Create the template



How to create a style-based Word template

You can easily embed structure and format requirements in a **document template** for distribution to authors to create documents.

A workable document template can be created in Word with the minimum level of structure shown here.

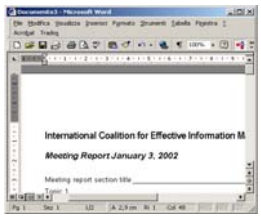
In adopting a style-based template, keep in mind that:

• Word uses a proprietary format: check for backward compatibility of new versions with older files;

• for complex templates, you need to programme macros to include in your document template;

• Word is useful for creating or editing, while XML is more advisable for structuring information for advanced processing (e.g. storage in a database, transformation, reusing components).

Conversion



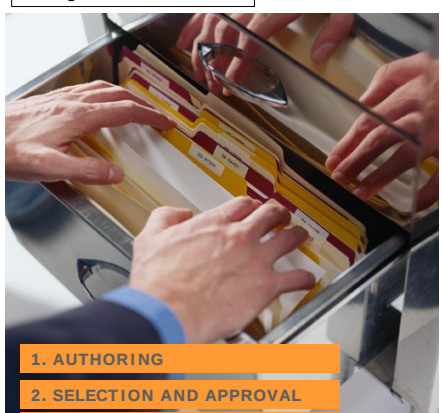
1. AUTHORING
2. SELECTION AND APPROVAL
3. CONVERSION
4. STORAGE
5. PROVIDING ACCESS

A text processing format like a Microsoft Word document is usually preferred for editing the content and the formatting **before the document is finally approved and selected** for conversion and publication.

The **conversion stage** can include different procedures depending on the file formats needed to visualize the final layout. Here are some **conversion standards**:

- **For PDF**: the compression options suitable to the intended use of the final output, e.g. to be read on screen or used for high-quality printing.
- **For HTML**: the HTML or XHTML definition for code validation, cascading style sheets for formatting and visual layout;
- **For XML**: a set of rules for mapping the template styles to the elements of the Document Type Definition; a Document Type Definition or a schema for validation; stylesheets for transformation into HTML, PDF or other formats.

Storage: file formats



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

Storage means keeping your documents in order, properly named, in a secure environment, in the most appropriate format for publication, reuse or conservation.

The most widely available **file formats** for electronic documents have a varying relevance to storage priorities.

The tables below summarise how suitable textual and image file formats are for the goals of preservation, reuse, access.

Table of storage formats for documents and images

Storage: file formats

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.

If your goal is...	Your decision should be:
Preserve content and look and feel of documents	To select a **software-independent format** for your documents whenever possible this will ensure that the content will be rendered in its integrity over time and regardless of the software utilized for its creation.
Reuse the documents and/or their components	Based on the **size and nature of the blocks** and on the format that allows you more **flexibility in transformation**.
Providing access to documents	Based on how your end users prefer to access your content. Relying on available software like **web browsers and free plugins** is likely to be more important than any consideration about proprietary formats. Because document addresses can change, providing access should take into account the issue of **persistence**. You might want to name your documents according to a scheme whereby they will remain available and accessible over time regardless of their location on the network.

Storage: file formats

For example, imagine that a book produced for print is to be reproduced on a CD-ROM and its components included in an online training course, slideshows and articles.

What is your main goal in identifying the most appropriate file formats?

- ○ Preservation
- ○ Reuse
- ○ Providing access

Click on your answer

Storage: file naming conventions

How will you **keep track** of versions and translations during the creation and conversion stages?

In a document workflow, storage also requires **keeping your documents in order and properly named**.

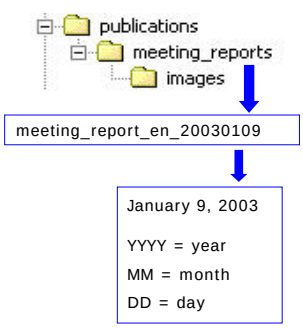
Even in a simple workflow, naming your files in a consistent way is a wise decision and will help you to:

- **prevent the loss** of documents and their components;
- **avoid renaming** for the sake of name compatibility with human comprehension, local drives and Internet servers, search, display, planning for database import of documents.

Storage: file naming conventions

It is helpful to define a set of **file naming conventions** and stick to them.

File naming conventions usually cover both the **directory structure** and the actual names **files** will be given. Here are some recommendations:



- Give **folders names** that help identify the files they contain.
- Give files **meaningful, memorable names**.
- **Dates** included in filenames should be written in reverse order and justified with a 0.
- Use **hyphens or underscore** to separate words.
- Do **not use spaces**: although supported by current Windows OSs, spaces are not tolerated in URLs.
- Indicate the **language of the document** content by using the 2-letter language code (e.g. en for English, fr for French, es for Spanish, ar for Arabic, zh for Chinese).

Storage: file naming conventions

More information about filenaming conventions

Moreover:

• Use letters or numbers as suffixes to mark **successive versions**: e.g. meeting_report_20030109**a** for the first version of the meeting report produced on January 9.

• File name **length** should be kept short as long as it allows for meaningful names. Eight characters is a limitation only if you are running on DOS or Windows 3.1

• For UNIX/Microsoft compatibility: write filenames in **lower case**.

• Do **not use** punctuation signs, such as: ,.;:#§*+!"|£$%&/()=?'^

Storage: file naming conventions

For example, 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

Click on your answer

Using a content management system

For large and complex workflow requirements, authoring, conversion and storage can also be approached with the adoption of a **content management system** where a core **database** and its related applications can help you to:

- apply the rules about formats, naming, versioning;
- provide access to different types of users based on their roles in the workflow;
- manage reuse of content;
- backup, archive and restore content.

Locators and identifiers

When you publish your documents on the Web, you are basically referencing them with a **URL** (**Uniform Resource Locator**), e.g.: **http//thelibrary.org/book.htm**.

The URL indicates where the document is located. However, what will happen if the documents are moved from one server to another?

The solution is to give your document a stable or **persistent identifier**, that identifies it as unique, regardless of how many copies are present on the Web or of the location where it is hosted.

Click on each button to find information on using identifier

Identifiers for internal publishing

An identifier is useful to track a document along the processing stage. For example, in FAO each publication is given a code called Job Number that uniquely identifies a document within FAO.

A publication can be identified as follows:

T1234A00.htm, where: **T1234** is a sequence that identifies that publication; **ar** is the language code (Arabic), **00** is the progressive numbering that identifies the first file of the publication (01, 02, etc).

Locators and identifiers

Identifiers for Internet publishing

If documents are made accessible online, it is important that:

- links to the documents are consistent and reliable;
- names are permanent;
- documents can be archived, e.g. their location changed or be preserved, while remaining available and accessible;
- multiple identical copies are identified as the same document.

Locators and identifiers

In practice, adopting an identifier system implies three factors:

1) An **identifier system**, e.g. choose what to call the documents;

2) A **system of resolution** to map the identifier to the document identified: when the identifier is used as a link, the resolution system will get users to the document.

3) Maintenance of access through **continued association** of the location with the identifier to make sure that the links continue to work over time.

Locators and identifiers

How the identifier system works

Once adopted, the identifier system works like this:

An article has been assigned a Digital Object Identifier (DOI)

Identifier: doi:10.1045/july95-arms

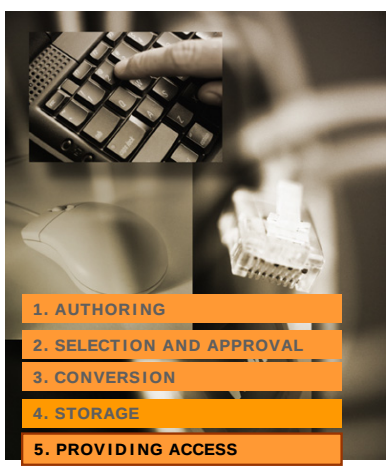
The access maintenance body provides the resolution service, in practice the URL that should be used to cite the article.

Resolver: http://dx.doi.org/10.1045/july95-arms

Clicking on the above URL takes you to the location where the article is published.

Locator: http://www.dlib.org/dlib/July95/07arms.html

Providing access to documents



The decision about how to give your **audiences access** to information is one of the key drivers in selecting and adopting standards along the workflow.

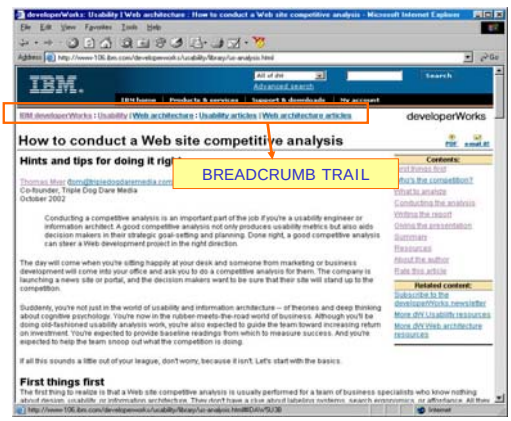
In an electronic document workflow the most natural option is often providing access to documents via the World Wide Web.

The simplest way to do it is to build a **static website**.

If the number of documents is high and search needs get complex, you can consider building a **dynamic website based on a database**. If your users have low bandwidth or no access to the Internet, you can consider releasing a **CD-ROM** version of your system.

Anyway, the goal is to support users in **finding and accessing documents** in easily understood ways.

Providing access to documents on the Web



To provide your users with **effective access to a website's content**, help them find their way easily:

- Provide a **site index**, e.g. an alphabetical list of pages and documents; alternatively, if you have a small collection, provide a **site map** with a visual overview of the main content areas of your systems.
- Provide a link to the home page and adopt the **breadcrumb trail**, e.g. the path of pages users have visited.
- Provide the document's **table of contents** with links to its sections.
- Label **links** in a descriptive way.

Providing access to documents on the Web

To design accessible sites, consider the following:

Write for the Web

- Use **headings**, **lists**, and **consistent** structure.
- Place **descriptive information** at the beginning of paragraphs, lists, etc.
- Write in a **simple, direct style** and block the text in **paragraphs** to help the eye scan it.

Use meaningful graphics

- Use **icons or graphics** (with a text equivalent) where they facilitate comprehension of the page.
- Provide a link to Acrobat reader **download** page for PDF document display.

Establish a dialogue with users

- Provide information about **your institution**.
- Make yourself **easy to contact**.

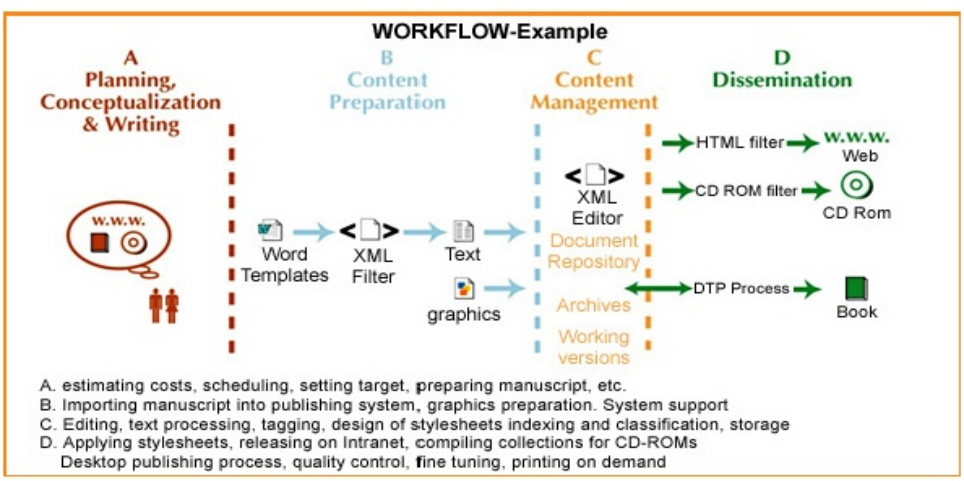
Maintaining the workflow

Once you have the standards, tools and roles in place, how do you keep it all together?

Level of complexity	How to mantain 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 **document management system** with workflow management capabilities.

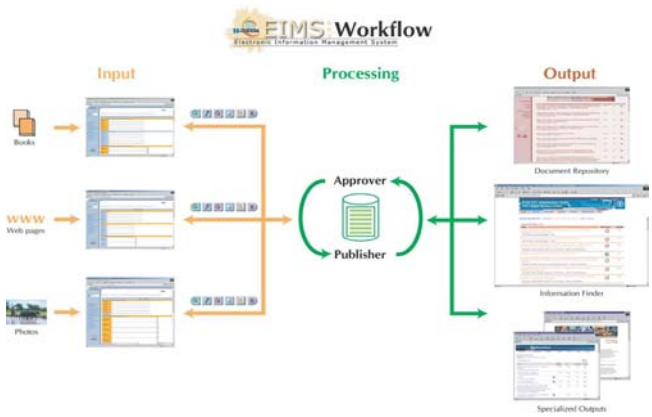
Document and workflow management systems can be used for providing access to documents to end users and thus cover the whole document workflow. How much of it is supported by the systems depends on the specific requirements and needs of your group and organization.

Example of a complex workflow



Approaches to workflow maintenance

A sample approach to complex workflows is given by the **FAO Document Repository**, a system for storage and dissemination of FAO documents and publications in electronic formats.



For the documents to get into the Repository, an electronic publishing workflow has been devised that is managed through the **EIMS**, the Electronic Information Management System.

EIMS tracks the cycle of a publication throughout the stages of creation, translation, conversion and publishing.

Approaches to workflow maintenance

Alternative approaches: Dspace and PKP

Dspace

DSpace is an open source digital asset management software platform that enables institutions to capture and describe digital works using a submission workflow module; distribute an institution's digital works over the Web through a search and retrieval system; and store and preserve digital works over the long term. DSpace is currently being implemented at the **Massachusetts Institute of Technology**.

http://dspace.org

PKP

The **Public Knowledge Project** is a research project at the **University of British Columbia** focused on improving access to scholarly publications and integrating it with complementary information. The PKP is currently developing and testing a number of online research management systems to improve the scholarly and public quality of academic research. These systems are designed not only to assist in the management and publishing of scholarly work, but to improve the indexing of research in online environments and create more connections with related online information. To ensure open, integrated, and well-indexed access to research the systems adhere to the Open Archives Initiative which provides a mechanism for linking research databases around the world.

http://www.pkp.ubc.ca/

Guidelines and procedures

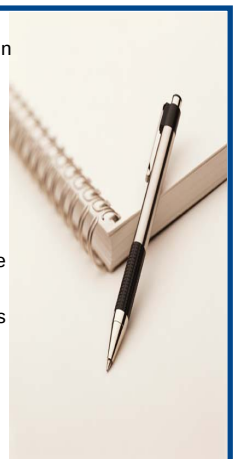
Here you can download and print the documents provided in this lesson.

You may use them as tools for your job.

- **Checklist for structuring workflow**
- **How to create a style-based Word template**
- **Template sample**
- **Table of storage formats for documents and images**

Summary

- Structuring an electronic document processing workflow relies on a **decision-making process** about final goals, document and conversion standards, tools and organizational roles.
- **Structured templates** facilitate conversion, storage and access to documents.
- **Preservation, reuse and access** set the priorities for deciding on which formats should be stored.
- **File naming conventions** should be used which facilitate human comprehension, assist in their location on local drives and Internet servers in terms of searching and display, as well as facilitate database import.
- **Identifiers** are names that identify documents as unique, regardless of how many copies are present on the Web or of the location where they are hosted.
- **Document management systems** with workflow management capabilities can track the status of documents along the workflow, ensure storage and provide access and search functionalities to end users.



Exercises

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

Good luck!

Exercise 1

These are the five main stages of creation and processing of documents to be disseminated through electronic and print media.

What is the correct sequence?

1	a
STORAGE	
AUTHORING	
CONVERSION	
PROVIDING ACCESS	
SELECTION AND APPROVAL	

Standard message

Exercise 2

When do you have to decide the 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.

Click on your answer

Exercise 3

You run a library or a documentation centre. You want to make sure that content, look and feel of documents will be displayed as originally intended over time and regardless of the software utilized for their creation.

What is your main goal?

- ○ Reuse
- ○ Preservation
- ○ Providing access

Click on your answer

Exercise 4

What are the benefits of using a document template based on Word styles?

- ☐ A document template provides a ready-made outline that helps authors write their content
- ☐ Styles simplify conversion to HTML and XML
- ☐ Using a template makes documents compatible with older versions of MS Word
- ☐ Table of contents can be built and updated quickly
- ☐ A template makes documents look nicer

Click on your answers

Exercise 5

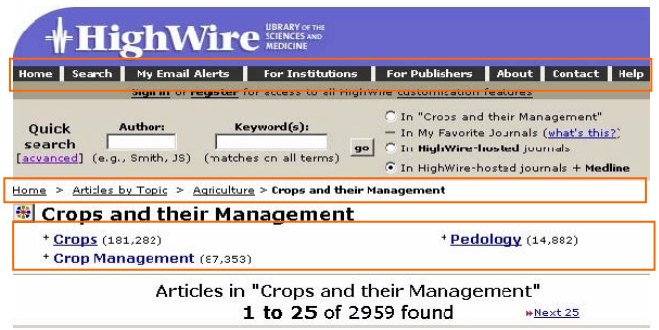
A URL is a name that uniquely identifies a document (or any other type of information resource) and will be forever associated with that document. It will ensure that when a document is moved, or its ownership changes, the links to it will continue to work.

- ○ True
- ○ False

Click on your answer

Exercise 6

Navigation bars help people moving around a website and access documents. They can come in many forms: look at this page and click on the breadcrumb trail.



1

2

3

Click on your answer

If you want to know more...

Structured Writing for a Single-Source Environment
(http://www.arbortext.com/html/webinars/webinar20.pdf)
Use styles to format text (http://www.shaunakelly.com/word/concepts/styles/index.html)
Best Practices for Digital Archiving - An Information Life Cycle Approach
http://www.dlib.org/dlib/january00/01hodge.html
Information Identifiers (http://www.elsevier.nl/inca/homepage/about/infoident/)
Key Concepts in the Architecture of the Digital Library
http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/July95/07arms.html#junewya4
http://www.nla.gov.au/padi/topics/36.html
http://www.nla.gov.au/nla/staffpaper/2001/dack.html
Searchtools.com – Search User Experience
http://www.searchtools.com/slides/baychi/index.html
http://www.searchtools.com/guide/index.html#form
Usable Web - Design Tips (http://usableweb.com/topics/000445-0-0.html)
Yale Web Style Guide (http://www.webstyleguide.com/index.html)
The Microsoft Valuable Professional Program website MS Word FAQ (http://www.mvps.org/word/)
Dspace, an open source digital library system developed jointly by MIT Libraries and Hewlett-Packard (http://dspace.org/)
Public Knowledge Project (PKP), a research project at the University of British Columbia (http://www.pkp.ubc.ca/)