Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 2. ELECTRONIC DOCUMENTS AND FORMAT

LESSON 8. CONVERSION BETWEEN FORMATS

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 :

- understand the **process involved in document conversion** from one format to another; and
- know more about the different ways of **converting documents**: from Word (doc) to HTML/PDF, from Word (doc) to XML, and XML to HTML/PDF.

What is document conversion?

I would like to display my Word document as a web page in HTML format, and also to print it from a paginated PDF file.

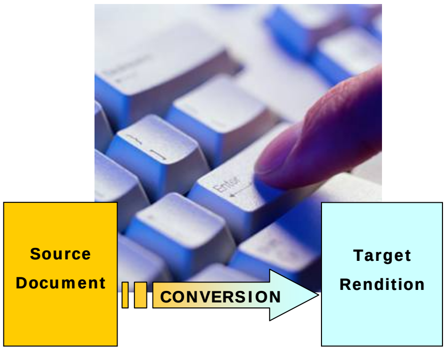
Sara, a member of her organization's Food Security department, has created a research document on desertification in Word format.

Now, she needs to publish it on the Web and allow users to read it in a browser as well as download and print it.

Therefore, she has to convert the documents from Word to HTML and PDF formats.

What is needed to do this? And how is it done?
Before proceeding, it's useful to know what **document conversion** is.

What is document conversion?



Document **conversion** is the transformation process applied to a source document in order to have different renditions (**target** renditions).

Conversion can be carried out:

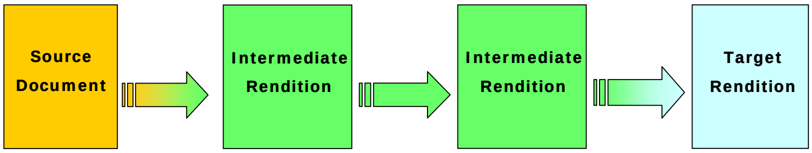
- **manually**, when a person creates the rendition by re-keying the document content, and inserting the mark-up necessary.

- using one or more **computer programs** that automatically convert the document from one format to another.

Often, conversion consists of one or more automated programs, together with manual intervention by users (**semi-automated transformation**).

What is document conversion?

The **semi-automated** transformation often takes two or more **separate transformation stages** (e.g. one manual and one automated) and connects them together to achieve the full transformation from source to target rendition.



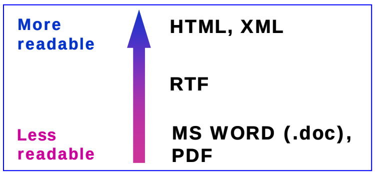
The output of each stage is called **intermediate rendition**. The intermediate rendition becomes the source format for the next rendition.

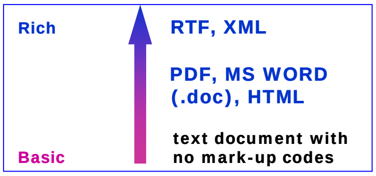
So conversion can also be used to have multiple renditions from a single source; it can also be used to move **from one source format to the next**.

What is document conversion?

Not all conversions have the same **level of complexity**.

It depends on the following factors:




Rich
Basic
RTF, XML
PDF, MS WORD (.doc), HTML
text document with no mark-up codes

READABILITY OF THE FORMAT

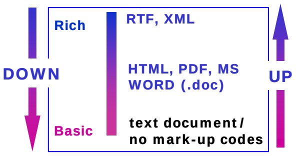
Plain text formats such as RTF, HTML or XML are **easy to read**: files in these formats can be opened and read in any plain text editing package. Binary formats such as Microsoft Word format are harder to read.

RICHNESS OF THE FORMAT

“Richness” refers to the amount of information that the mark-up is able to convey.

HTML conveys some information about the formatting but not as much as RTF or XML formats. In particular, XML also conveys information about the semantic structure of the document.

What is document conversion?



An up-transformation refers to going from a simple format to a richer one. The inverse is called a **down-transformation**.

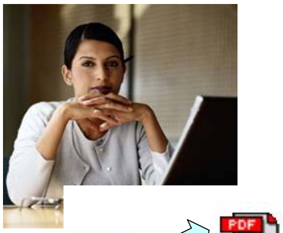
Which of the following transformations do you think is easier to carry out?

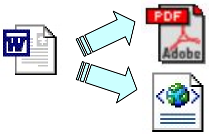
○ From XML to HTML (down).

○ From HTML to XML (up).

Please click on your answer

Conversion from a Word document to PDF / HTML





Let’s go back to Sara’s task.

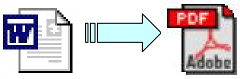
She has to convert:

- a Word document into PDF format; and
- a Word document into HTML format.

Let’s look at how she can do these conversions and what tools she needs.

We will start with the conversion from Word to PDF.

Conversion from a Word document to PDF / HTML



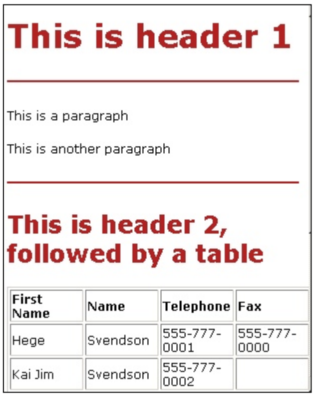
The Adobe Acrobat suite of tools (e.g. PDF Maker) can be used to:

- open the Word document, and
- save it as PDF file.

Moreover, any application that can print documents (like Microsoft Word) can also create a PDF by installing a **PDF print driver**.

Adobe’s own PDF print driver is called PDF Writer, but there are print drivers available from many other commercial and open sources available on the Web.

Conversion from a Word document to PDF/HTML



This is header 1

This is a paragraph

This is another paragraph

This is header 2, followed by a table

First Name	Name	Telephone	Fax
Hege	Svendson	555-777-0001	555-777-0000
Kai Jim	Svendson	555-777-0002	

A good tutorial on how to do CSS can be found on the W3 Schools website (www.w3schools.com).

Conversion **from Word to HTML** can be made in different ways, that can involve more or less manual work. However, some manual work is always required, so it’s a prerequisite to have a basic knowledge of HTML.

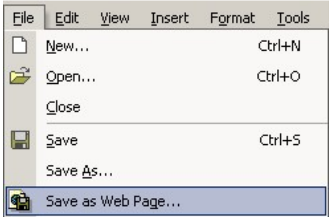
Before starting, you should analyze the document structure and create a **Cascading Style Sheet (CSS)**, a text file which define **how to display** HTML elements (e.g. titles, tables, lists, etc.).

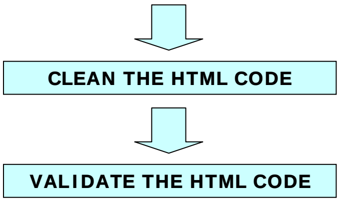
CSS can save you a lot of work, as it allows you control over the format of a group of Web pages all at once: for example, whenever you want to change the font in all the Web pages, you just have to change the CSS file.

CSS can be created by hand, or using tools like TopStyle which has a freely available version named TopStyle Lite:

www.bradsoft.com/topstyle/tslite/index.asp

Conversion from a Word document to PDF/HTML





CLEAN THE HTML CODE

VALIDATE THE HTML CODE

You can convert your Word document directly from Microsoft Word, by selecting the ‘**Save as HTML**’ (or ‘**Save as Web Page**’) option, available under the “File” menu.

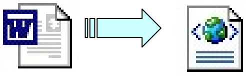
In this case, you have to **clean the resultant format**, as the program automatically adds a lot of useless information. If you don’t do this, the final file will be heavy and users could encounter some problems in displaying it on their browser.

You can clean your file using, for example, HTML Tidy, which is part of a free toolkit named **HTML Kit**.

HTML Tidy make your file cleaner, but you also should complete the process by deleting all the information that is not part of the document’s content.

Finally, it’s recommended that you **validate the HTML code**, to check it follows HTML standards. HTML Kit also provides a code validator.

Conversion from a Word document to PDF/HTML



Examples of converters:
Avanstar Transit
www.avantstar.com/solutions/transit/default.aspx
Logictran RTF Converter
www.logictran.net/

An optimized way to convert a Word document to HTML is by using dedicated tools, which can convert styled, template-based Word documents into clean and correctly formatted HTML, sometimes through an intermediate conversion to RTF.

These tools let you establish the conversion rules, e.g.:

- mapping Word styles to HTML elements,
- splitting the document into multiple pages,
- converting images to Web-compatible formats,
- preserving notes and cross-references in a document.

The disadvantage of using these tools is that they are **not free**, so you have to evaluate if it's worth buying one of them.

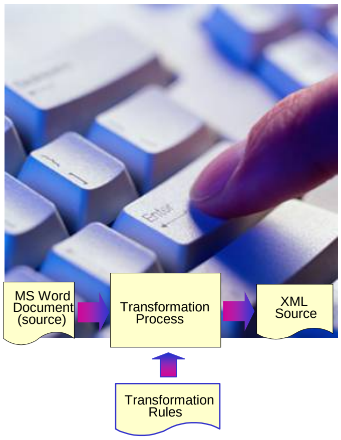
Using XML as a source format

A colleague told me about the usefulness of XML in the conversion processes. I would like to learn more about it...

Microsoft Word is often chosen as the original document creation application, and it can be used as source document to obtain other renditions.

However, many organizations are beginning to use XML to hold the source documents because it is easy to transform to other renditions; moreover, its mark-up captures the logical meaning of the content, it is open source and well defined with public specifications.

Conversion from Word to XML



MS Word Document (source)

Transformation Process

XML Source

Transformation Rules

There are a number of tools available on the market which can **plug in to Word** to help make the transformation to XML.
They generally use Word styles to make the transformation and rely on users of the word processor to apply word styles in a consistent manner.

In this case it is necessary that users have created Word documents **using styles and templates correctly**. If not, it is quite difficult to make a fully automated transformation from Word to XML.

Some organizations solve this problem by having a small team of people (the production or technical editorial team) who make manual corrections to the source Word documents before transformation and/or to the target XML documents after transformation.

Conversion from Word to XML

OpenOffice.org - Microsoft Internet Explor...
File Edit View Favorites Tools Help
Address http://www.openoffice.org
Go
OpenOffice.org Source Project
Home
Site Map
About Us
News
Downloads
Application
Source
Join/Login
OpenOff
OpenOffice
Download Oper
Windows | Linux |
Internet

Conversion to XML can also be made through an intermediate RTF or XHTML conversion.

Some organizations have developed their own application to do the conversions (**filters**), but for this the availability of one or more developers is necessary.

Also, an open source application like the Open Office suite (www.openoffice.org) can be used.
The Open Office suite can read Microsoft Word, Excel and Power Point files and can save to XML conforming to the Open Office DTD. Then, another transformation must be done to produce the target XML, conforming to the preferred DTD or schema.

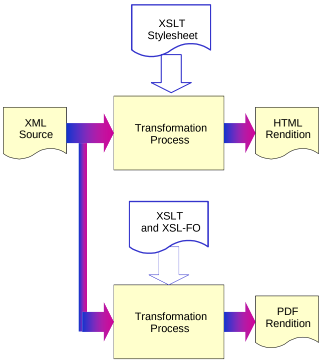
Converting from Word to XML

The transformation is 90% correct. Not bad... but how much time will we need to make it 100% correct?

One important point worth considering when choosing how to up-transform to XML is the **amount of time** you will spend fixing problems that result from an imperfect automated transformation from Microsoft Word or other less rich format.

Often it is actually easier and quicker to start from the beginning and **re-key the document as XML**.
There are many commercial re-keying agencies which guarantee a minimum error rate (e.g. one error in 20,000 characters) and a turnaround time from receipt of the source to return of the target XML documents. It's well worth considering such an approach, especially if your original source documents are only available in hardcopy.

Conversion from XML to HTML/PDF



One of the great advantages of XML is that it is very easy to transform XML mark-up to another format. The Extensible Stylesheet Language for Transformations (**XSLT**) offers a standard way to transform XML and there are many XSLT transformation processors available, both as open source and as commercial products.

There is also a standard way to transform XML into page-formatted renditions such as PDF, Postscript or RTF, the **XSL-FO**.
XSL-FO (XSL Formatting Objects) is a set of XML elements that represent objects such as pages, text blocks, tables, lists, footnotes, etc.

XSL-FO was published as the XSL standard by the W3C :

🌐 http://www.w3.org/TR/xsl/

Summary

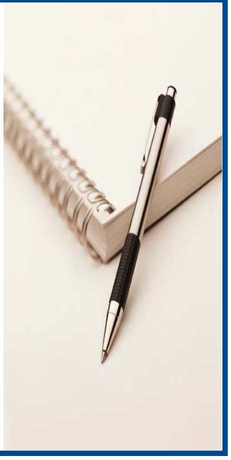
- Document conversion is the **transformation process** applied to a source document in order to create different target renditions.

- The transformation process may be **manual, automated** or **semi-automated**.

- The two factors in the mark-up of the source document that most affect the conversion process are the **readability** and the **richness** of the format.

- An **up-transformation** refers to going from a simple format to a richer one **(e.g. from Word .doc to XML)**. The inverse is called a **down-transformation (e.g. from XML to HTML)**.

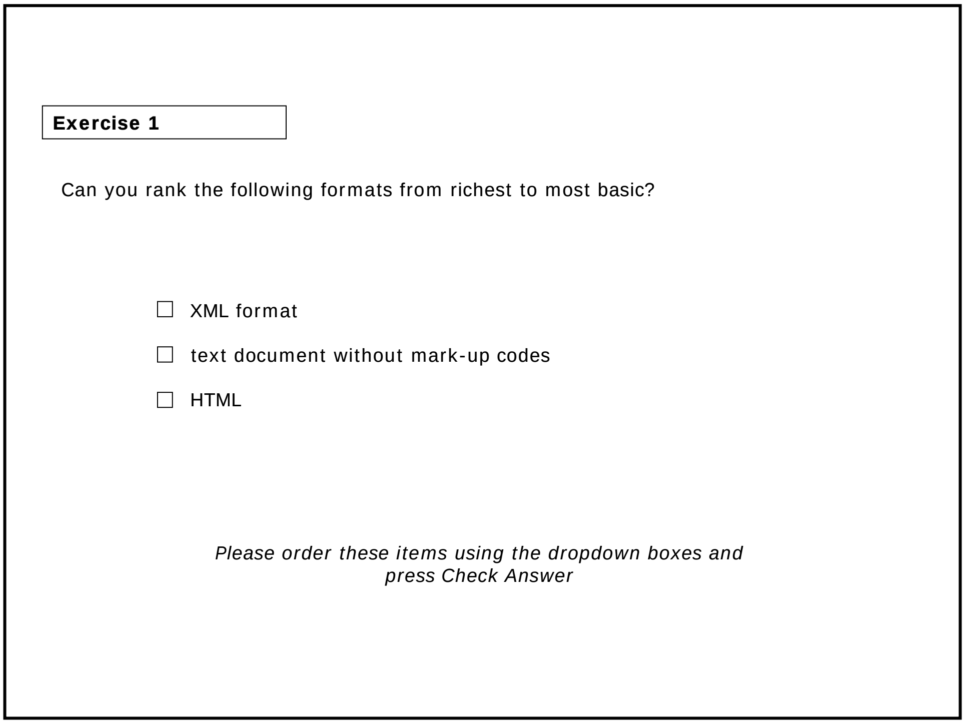
- **XML** is often used as **the primary source format** because it is an open, vendor neutral format, its mark-up captures the logical meaning of the content, it is well defined with public specifications, and it’s easy to transform into other renditions.



Exercises

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

Good luck!



Exercise 1

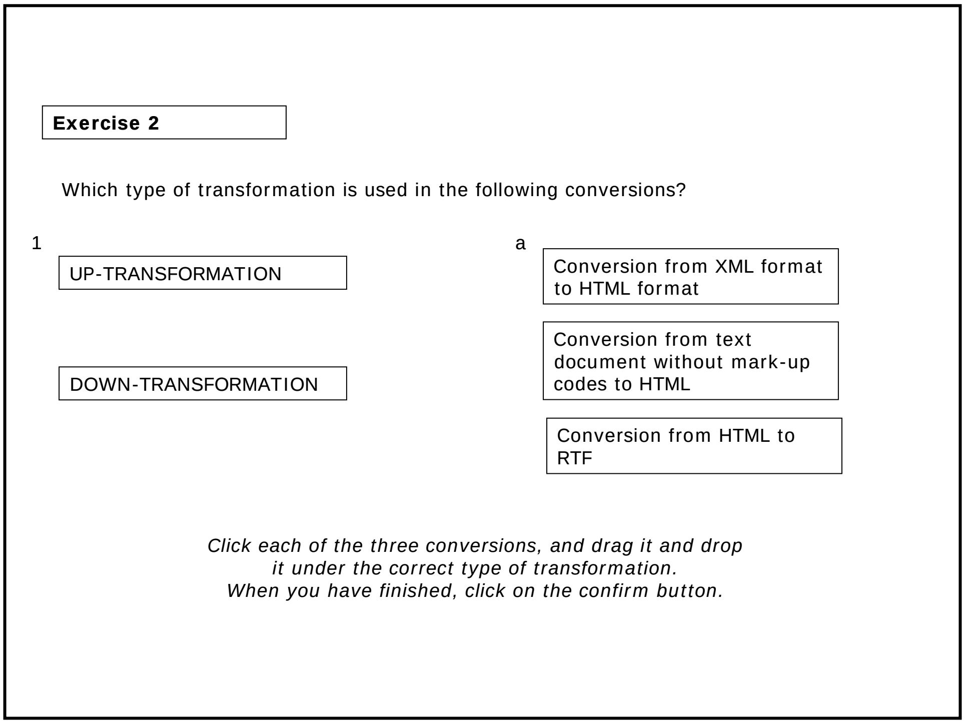
Can you rank the following formats from richest to most basic?

□ XML format

□ text document without mark-up codes

□ HTML

Please order these items using the dropdown boxes and press Check Answer



Exercise 2

Which type of transformation is used in the following conversions?

1

UP-TRANSFORMATION

DOWN-TRANSFORMATION

a

Conversion from XML format to HTML format

Conversion from text document without mark-up codes to HTML

Conversion from HTML to RTF

Click each of the three conversions, and drag it and drop it under the correct type of transformation.
When you have finished, click on the confirm button.

Exercise 3

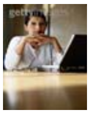
Which of the following procedures, used to convert a Word document to an HTML document, is cheaper?

- ○ Convert the file doc into an HTML format using the “Save as Web Page” option of Microsoft Word.
- ○ Convert the file doc into an XML format using dedicated tools like Avanstar Transit or Logictran RTF Converter.

.

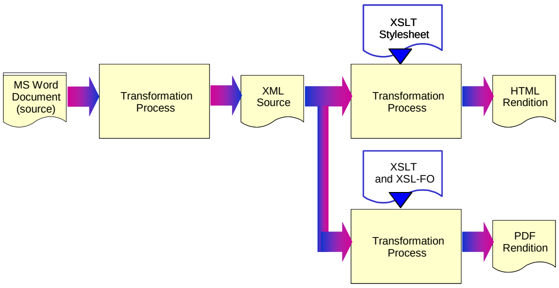
Please click on your answer

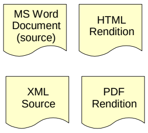
Exercise 4



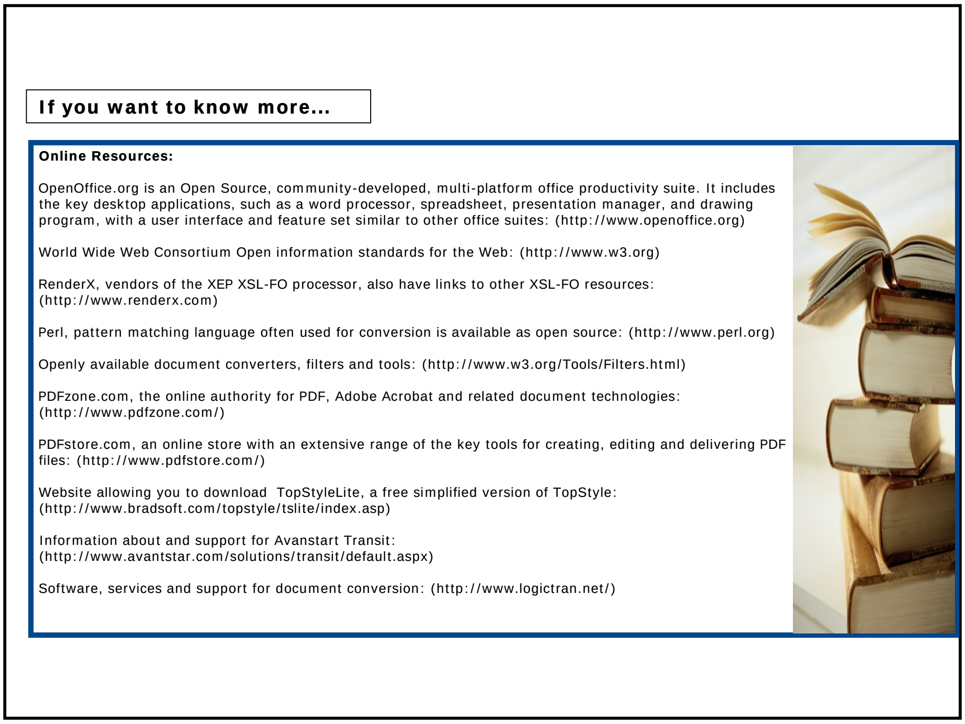
“My Word document must be published on our Web site. I also need to create a rendition for printing. To obtain this result I will use an intermediate rendition...”.

Which process is involved in this case?





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



If you want to know more…

Online Resources:

OpenOffice.org is an Open Source, community-developed, multi-platform office productivity suite. It includes the key desktop applications, such as a word processor, spreadsheet, presentation manager, and drawing program, with a user interface and feature set similar to other office suites: (http://www.openoffice.org)

World Wide Web Consortium Open information standards for the Web: (http://www.w3.org)

RenderX, vendors of the XEP XSL-FO processor, also have links to other XSL-FO resources: (http://www.renderx.com)

Perl, pattern matching language often used for conversion is available as open source: (http://www.perl.org)

Openly available document converters, filters and tools: (http://www.w3.org/Tools/Filters.html)

PDFzone.com, the online authority for PDF, Adobe Acrobat and related document technologies: (http://www.pdfzone.com/)

PDFstore.com, an online store with an extensive range of the key tools for creating, editing and delivering PDF files: (http://www.pdfstore.com/)

Website allowing you to download TopStyleLite, a free simplified version of TopStyle: (http://www.bradsoft.com/topstyle/tslite/index.asp)

Information about and support for Avanstart Transit: (http://www.avantstar.com/solutions/transit/default.aspx)

Software, services and support for document conversion: (http://www.logictran.net/)