
HTMLtoXML Converter (HTML2XML)

User's guide

Version e1.10

Preface:

Each patent office in the world, including the World Intellectual Property Organization (WIPO) and the triple pole (EPO, USPTO and JPO), is promoting the computerization of patent documents (in XML format) for the following advantages:

- Efficient paperwork (efficient formality examination and efficient use of search)
- Providing better service to applicants (using globalized patent documents)

For these reasons, the patent offices centering on the WIPO are developing tools (PCT-SAFE Editor by WIPO and PatXML by EPO). However, xml filing has not become sufficiently widespread among applicants worldwide.

To solve these problems, the Japanese Patent Office will release the HTMLtoXML Converter (commonly called HTML2XML) to the world. At the JPO, xml e-filing of domestic patent applications has surpassed a rate of more than 97%, while HTML2XML is based on the technology of personal computer application software (JPO-PAS) provided to Japanese patent applicants. Please use HTML2XML for the XML computerization of patent documents to be applied in English in each country concerned.

<u>Chapter 1 Overview</u>	<u>4</u>
1.1 What is HTMLtoXML Converter (commonly called HTML2XML)?	4
1.2 Features of HTML2XML	4
1.3 Work Flow	5
1.4 Operation Environment	5
<u>Chapter 2 Operating Procedure</u>	<u>6</u>
2.1 Starting Word	6
2.2 Entering Data	6
2.3 Saving a Document in HTML Format	7
2.4 Converting to XML with HTML2XML	8
2.5 Importing to PCT-SAFE	9
<u>Chapter 3 Document Samples and Notes</u>	<u>11</u>
3.1 Important Notes	11
3.2 Linking Images	13
3.3 Linking Array Tables	14
<u>Chapter 4 Typical Error Messages</u>	<u>15</u>
4.1 Typical Error Dialog Boxes	15
4.2 Typical Errors Displayed by [Display Check result]	16
<u>Appendix</u>	<u>17</u>
Appendix 1 Installation and Uninstallation	17
Appendix 2 Using Word Macros	18
Appendix 3 Headings Writable to Documents	19
Appendix 4 Usable Images	21
Appendix 5 Usable File Paths	21
Appendix 6 HTML Tags Supported by HTML2XML22	

Notice:

- **This software may be revised without prior notice.**
- **The contents of this manual shall not be disclosed in any way or reproduced in any media without the express written permission of FUJITSU LIMITED.**
- **The contents of this manual may be revised without prior notice.**

Adobe, Acrobat Reader, and Acrobat are trademarks or registered trademarks of Adobe Systems Incorporated in the U.S. and other countries. Microsoft, Windows, and Internet Explorer are registered trademarks of Microsoft Corporation in the U.S. and other countries. Other product names are product names, trademarks or registered trademarks of the respective companies.

Microsoft Windows Vista Home Basic, Home Premium, Business, and Ultimate are collectively referred to as Windows Vista. Microsoft(R) Windows(R) XP Professional operating system and Microsoft(R) Windows(R) XP Home Edition operating system are collectively referred to as Windows XP.

This software is based in part on the work of the Independent JPEG Group.

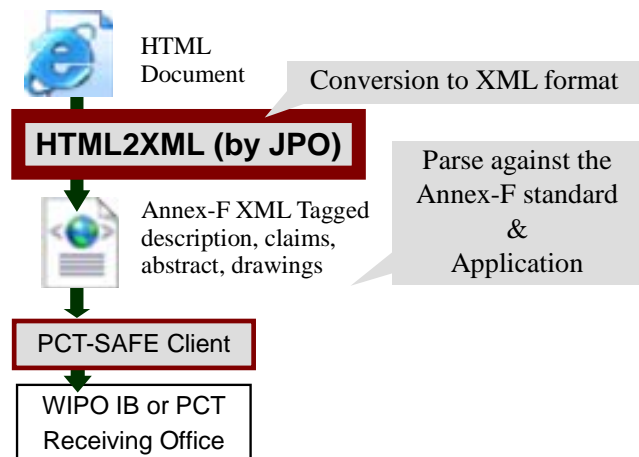
Chapter 1 Overview

1.1 What is HTMLtoXML Converter (commonly called HTML2XML)?

HTML2XML converts HTML documents prepared in Word into XML format. Converting HTML documents into XML format can make XML documents that comply with the PCT e-filing standard (Annex F of the PCT Administrative Instructions).

Note:

The documents here refer to description, claims, abstract and drawings.



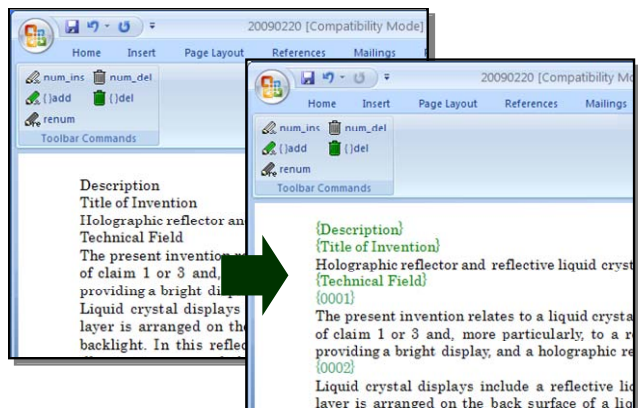
1.2 Features of HTML2XML

HTML documents can be prepared and converted into XML format through typical Word operations.

HTML2XML is not an XML authoring tool. In other words, no special knowledge and operations are necessary to write in XML. An HTML document can be prepared for input by simply creating a document with headings enclosed in { } or [] with Word and saving it in HTML format.

Word macros to help assign paragraph numbers and enter { } and [] are provided as samples.

HTML2XML comes with Word macros to support the preparation of documents. To convert a document with HTML2XML, paragraph numbers must be inserted and headings enclosed in { } or []. Word macros support these tasks. A document as shown in the figure can be prepared by simply clicking buttons.



Others

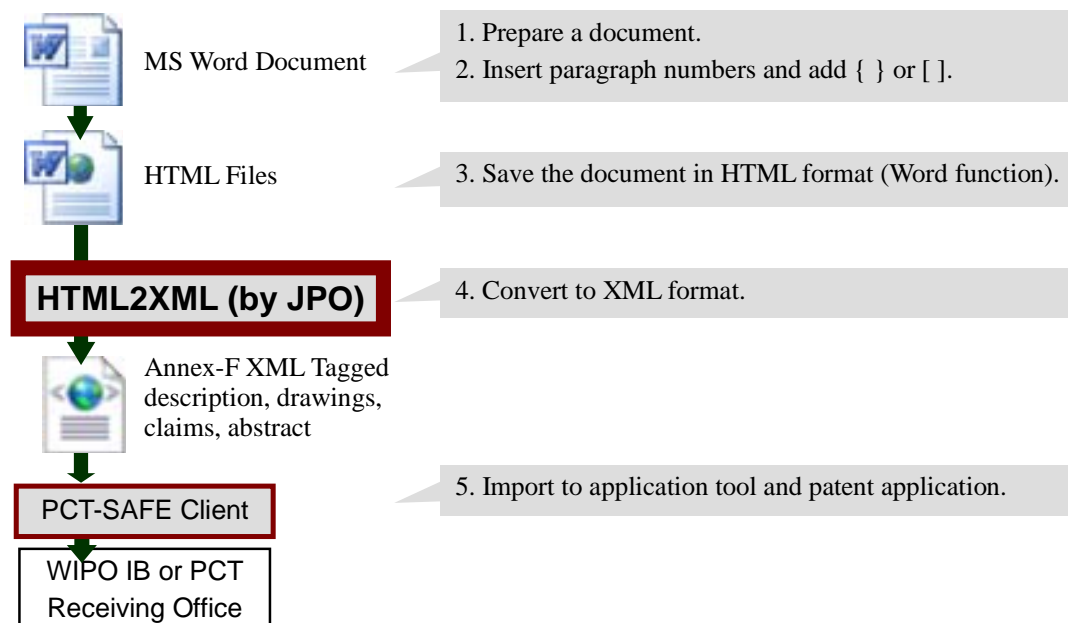
- XML documents compliant with the PCT e-filing standard (Annex-F) are created.
- Converted XML documents can be checked in PDF format before sending a patent application.
- CAF format that is applied from January 2009 is supported.
- PCT-SAFE can be started from HTML2XML. This function is useful for applying the documents converted with HTML2XML by using PCT-SAFE.

1.3 Work Flow

This section introduces the general work flow.

Hint:

- For details on the operation of each procedure, see Chapter 2, "Operating Procedure."
- For details on Word documents to be prepared in Procedures 1 and 2, see Chapter 3, "Document Samples and Notes."
- For details on installing HTML2XML, see Appendix 1, "Installation and Uninstallation."



1.4 Operation Environment

OS	Windows XP(SP3), Windows Vista(SP1)
Editor	Microsoft Word 2003, 2007

Chapter 2 Operating Procedure

This chapter explains the procedure from preparing a document to converting and importing it to PCT-SAFE.

2.1 Starting Word

1) Start Word.

2) When using Word 2007, click the Microsoft Office Button () and select [Open].

When using Word 2003, select [Open] from [File] on the menu bar.

3) Select a template.

Hint:

- This folder provides the templates.
- Open the template from Word. Double-clicking the icon of a template displays the template in a browser because the provided templates are in HTML format, and you cannot edit it in the browser.

2.2 Entering Data

Hint:

For details on Word operation, refer to the manual supplied with the product.

1) Enter the data in a common format.

Notes on entry are summarized separately (see Chapter 3). Because likely misunderstood headings are provided, be sure to check such heading before entering data.

2) Insert paragraph numbers and enclose headings in { } or [].

Hint:

- For details on where paragraph numbers are required, see Appendix 3, "Headings Writable to Documents."
- For details on Word macros, see Appendix 2, "Using Word Macros."

2.3 Saving a Document in HTML Format

Word 2007

- 1 Select [Save As] from the Microsoft Office Button ().
- 2 Select [Web Options] from [Tools].
- 3 Set the following on the [Encoding] tab:
 - Code: Unicode (UTF-8)
 - Always save Web pages in the default encoding: Uncheck the check mark.
- 4 Specify the file storage location and file name.

Note:

Only ASCII characters can be used for the file path and file name.

- 5 Save the file with [Web Page, Filtered].

Note:

Do not select [Single File Web Page].



Word 2003

- 1 Select [Options] from [Tools].
- 2 Select [Web Options] from [General].
- 3 Select the following on the [Encoding] tab:
 - Code: Unicode (UTF-8)
 - Always save Web pages in the default encoding: Uncheck the check mark.
- 4 Select [Save As] from [File].
- 5 Specify a file storage location and file name.

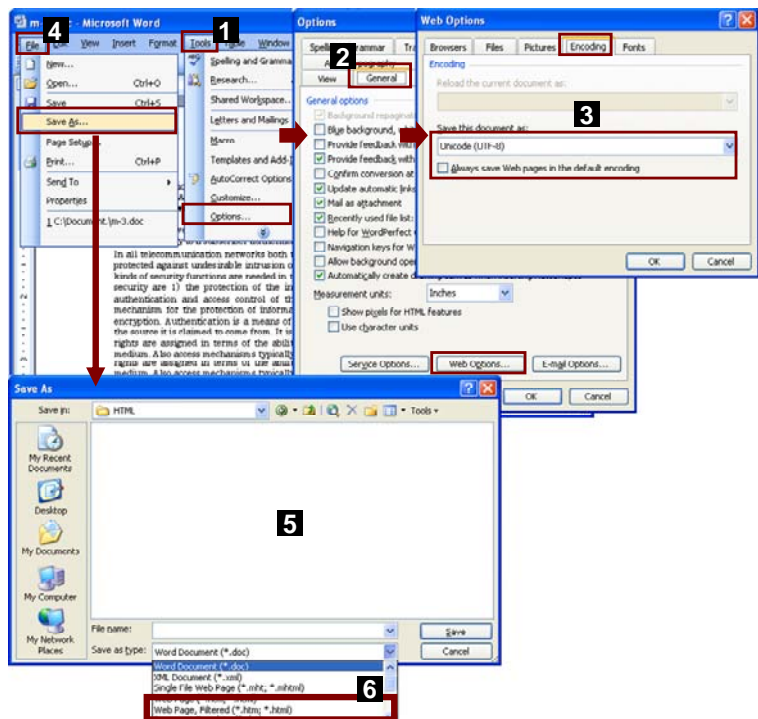
Note:

Only ASCII characters can be used for the file path and file name.

- 6 Save the file with [Web Page, Filtered].

Note:

Do not select [Single File Web Page].



2.4 Converting to XML with HTML2XML

Note:

- After conversion with HTML2XML, a folder with the name "HTML-file-name.out" is created in the same folder as the original HTML file, where a set of files is stored. Do not delete the files in that folder or change their names. Changing the folder configuration may result in abnormal system operation.
- For information about operations likely to be misunderstood, see Chapter 4, "Typical Error Messages."

- 1 Start HTML2XML ([All Programs] - [JPO] - [HTMLtoXML Converter] from [Start] ([start])).
- 2 Specify an HTML file to be converted (enter the path or select the file from [Browse]).
- 3 Specify [Country/Receiving Office] (enter the abbreviation of the country or select the country from the pull-down list).
- 4 Select a language for patent application.
- 5 Click [Convert to XML] to convert to XML.
- 6 Check the conversion result.

Hint:

If WARNING or ERROR is displayed, check the details by clicking [Display Check result]. Later, correct the original file, save it again in HTML format, and convert it to XML.

- 7 Check the path of the conversion result XML file.

Hint:

- Use PCT-SAFE to specify the XML document with the path that is displayed here.
- Select the path, right-click it, and select [Copy]. The path can be copied. This is useful for specifying the file path in PCT-SAFE.
- To check the conversion result, click [Display application-body] and [Display Seq. list] (reference information).

- 8 Click [Run PCT-SAFE] to start PCT-SAFE.

Note:

If PCT-SAFE is not yet installed, clicking [Display application-body] and [Run PCT-SAFE] displays an error message and PCT-SAFE cannot be started. Install PCT-SAFE.

2.5 Importing to PCT-SAFE


Hint:

For details on using PCT-SAFE, refer to the PCT-SAFE manual.
This section explains how to import the XML file created with HTML2XML.

- 1 Select [production mode] in the dialog box that appears at PCT-SAFE startup.
- 2 Select a template.
- 3 Select [Contents].
- 4 Select [Fully electronic].
- 5 Select [Application body].
- 6 Click [Yes] and specify the file.

Hint:

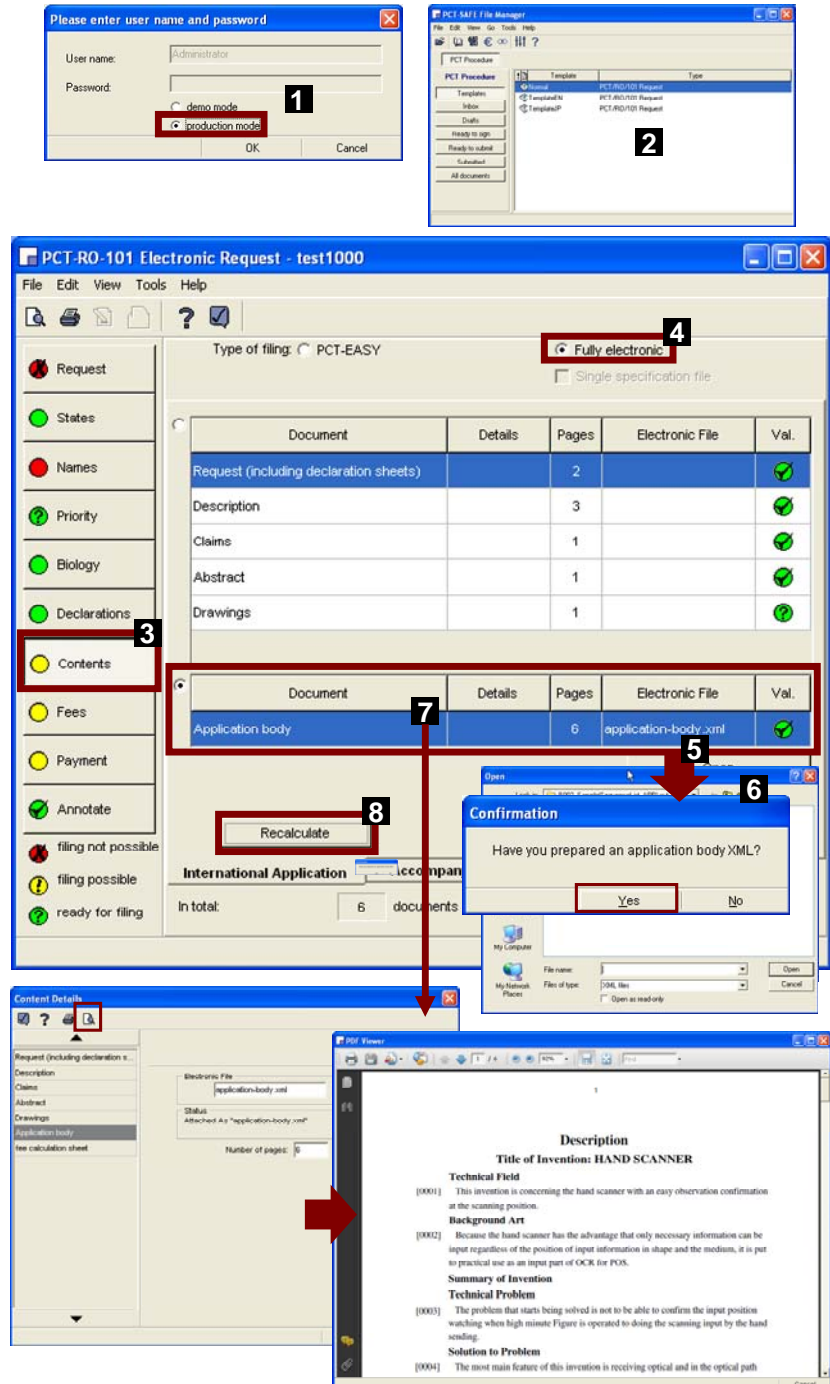
Here, specify the file with the path that is displayed in the HTML2XML window. If the file path was copied in HTML2XML, the file can be specified by pasting the path to [File name].

- 7 Check whether the attached XML file has an error (by double-clicking [Application body] and then clicking ).

Note:

Be sure to check before sending the patent application.

- 8 Click [Recalculate] to calculate the total number of pages and recalculate the fee.



The screenshots show the following steps:

1. Login dialog box with 'production mode' selected.
2. File Manager showing template selection.
3. Main application window with 'Fully electronic' selected and 'Contents' highlighted in the left sidebar.
4. Document list table with 'Application body' selected.
5. Confirmation dialog box with 'Yes' selected.
6. Open dialog box for file selection.
7. 'Recalculate' button highlighted.
8. 'Content Details' window showing 'Application body' with 'Number of pages: 6'.
9. PDF Viewer showing the patent document content.

Document	Details	Pages	Electronic File	Val.
Request (including declaration sheets)		2		✓
Description		3		✓
Claims		1		✓
Abstract		1		✓
Drawings		1		✓

Document	Details	Pages	Electronic File	Val.
Application body		6	application-body.xml	✓

Hint:

The procedure to include an array table in PCT-SAFE is shown below.

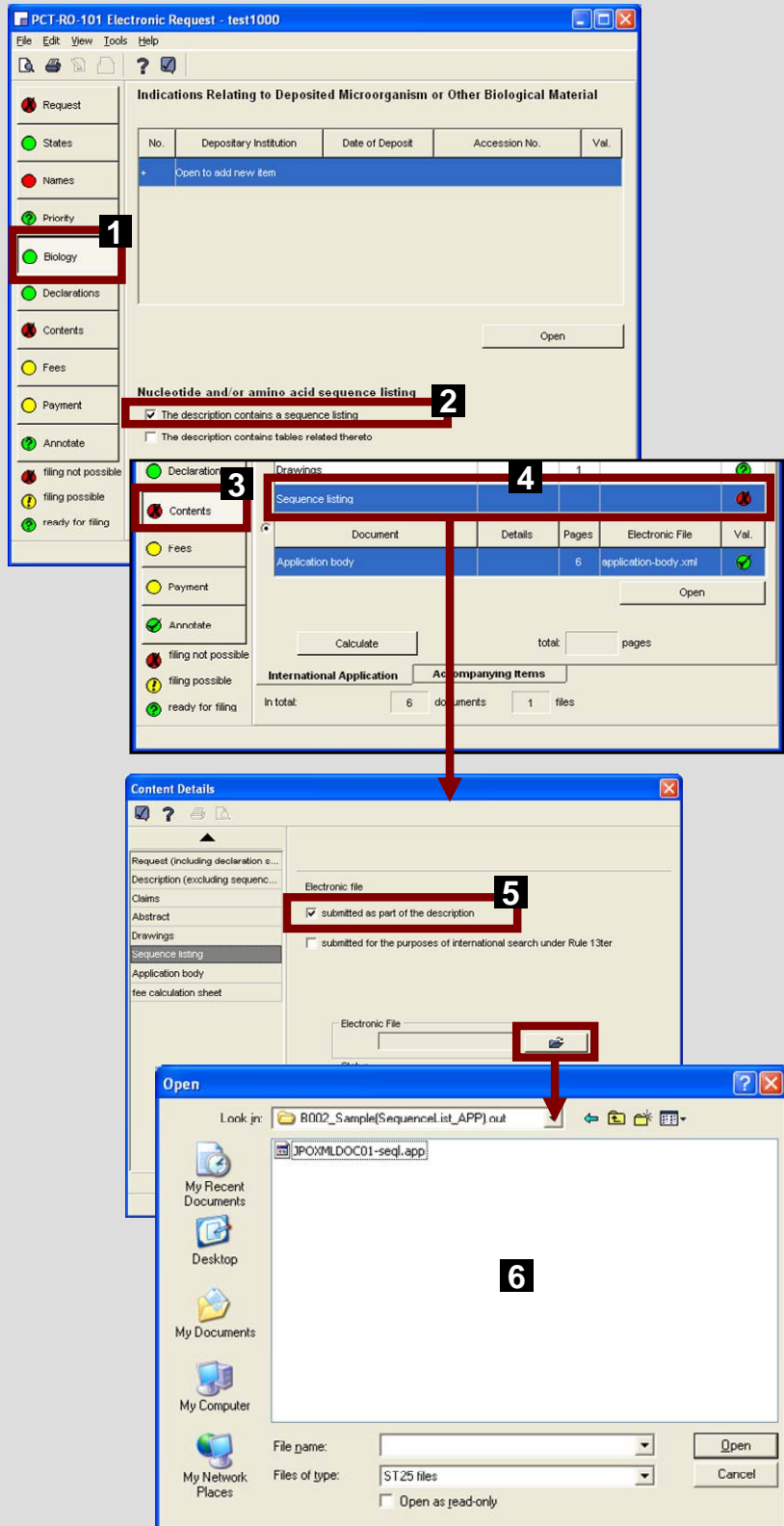
The array table is displayed only when [Array table in description] is selected in the biology edit window.

- 1 Select [Biology].
- 2 Select [The description contains a sequence listing].
- 3 Select [Contents].
- 4 Double-click [Sequence listing].
- 5 Select [submitted as part of the description].
- 6 Specify a file.

Hint:

Here, specify the array table file for the path displayed in the HTML2XML window (ST.25). If the file path was copied in HTML2XML, the file can be specified by pasting the path to [File name].

- 7 Click [Recalculate] to calculate the total number of pages and recalculate the fee.



Chapter 3 Document Samples and Notes

3.1 Important Notes

A sample document to be created in Word is given below. Points likely to be misunderstood are included. Be sure to check these points.

Hint:

Other contents likely to be misunderstood are also listed in Chapter 4, "Typical Error Messages." Also check these contents.

Note:

- Write description, claims, abstract and drawings to one Word file.
- HTML tags need not be written (do not write them).
- Do not use Word functions (headers, footers, indents and styles). They may be converted incorrectly.
- For details on headings that can be included in a common format, see Appendix 3, "Headings Writable to Documents."

Note:

Add paragraph numbers for conversion with HTML2XML.

To automate this work, use the Word macros of HTML2XML. For details on the procedure, see Appendix 2, "Using Word Macros."

Note:

For conversion with HTML2XML, enclose an heading in { }.

To automate this work, use the Word macros of HTML2XML. For details on the procedure, see Appendix 2, "Using Word Macros."

Note:

Do not enter a character to the right of an heading.

Note:

To enter a serial number, insert a space between an heading and number.

:
 :
 {Description of Embodiments}
 {0007}
 The purpose of inputting the image
 {Example 1}
 {0008}
 Fig. 1 is one cross section of the ex
 {0009}
 {Math. 1}
 Figure of mathematical expression
 {0010}
 {Table 1}
 Figure of table
 {Industrial Applicability}
 {0011}
 This invention can be applied to the usage to which...
 {Reference Signs List}
 {0012}
 1 Hand scanner housing
 2 Input document
 3 One dimension image sensor
 4 Wide lens
 5 Lighting lamp
 {Citation List}
 {Patent Literature}
 {0013}
 {PTL 1}
 Patent 2003-00
 {Non Patent Literature}
 {0014}
 {NPL 1}
 Patent Company Publication "Variations of Hand ...
 :
 {Sequence Listing}
[Sequence List.TXT](#)

Note:
 Paste an image file with a link. Do not paste it by pressing Ctrl + v.
 For details on the procedure, see Section 3.2, "Linking Images."

Note:
 The usable image files are restricted.
 For details, see Appendix 4, "Usable Images."

Note:
 To attach an array table, create the array table file separately and link to it.
 For details on the procedure, see Section 3.3, "Linking Array Tables."

3.2 Linking Images

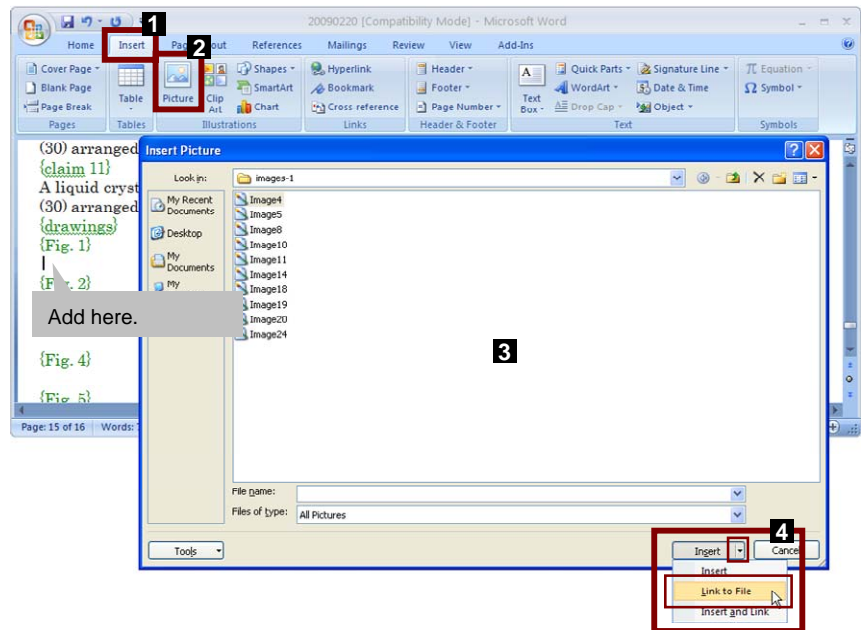
Paste an image file with a link. Do not paste it by pressing Ctrl + v.

Note:

After an image is linked, do not move or delete the image file and HTML document. Otherwise, error "The image file is missing" is displayed.

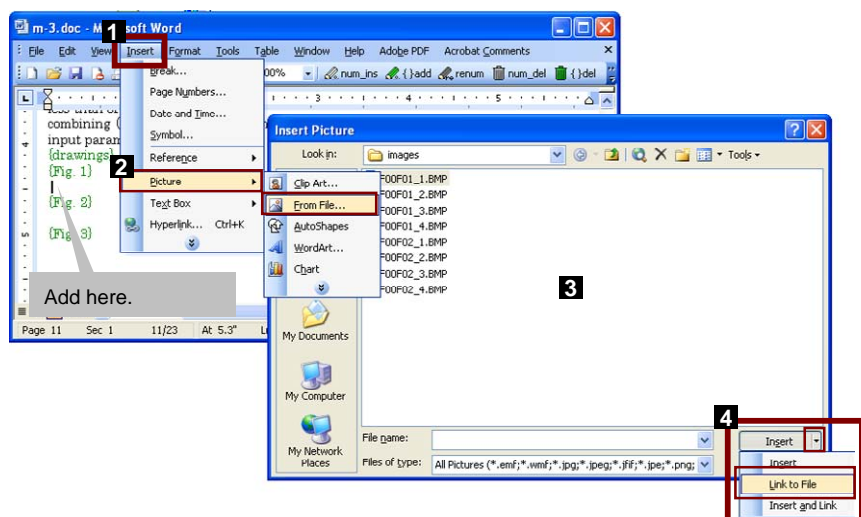
Word 2007

- 1 Select [Insert].
- 2 Select [Picture].
- 3 Select a file.
- 4 Select [Link To File].



Word 2003

- 1 Select [Insert].
- 2 Select [Picture] - [From File].
- 3 Select a file.
- 4 Select [Link To File].

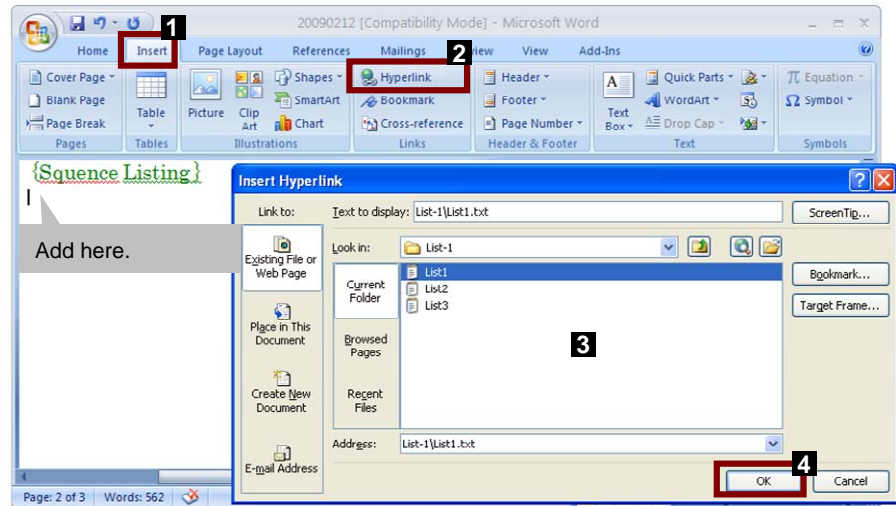


3.3 Linking Array Tables

To attach an array table, create the array table file separately and link to it.

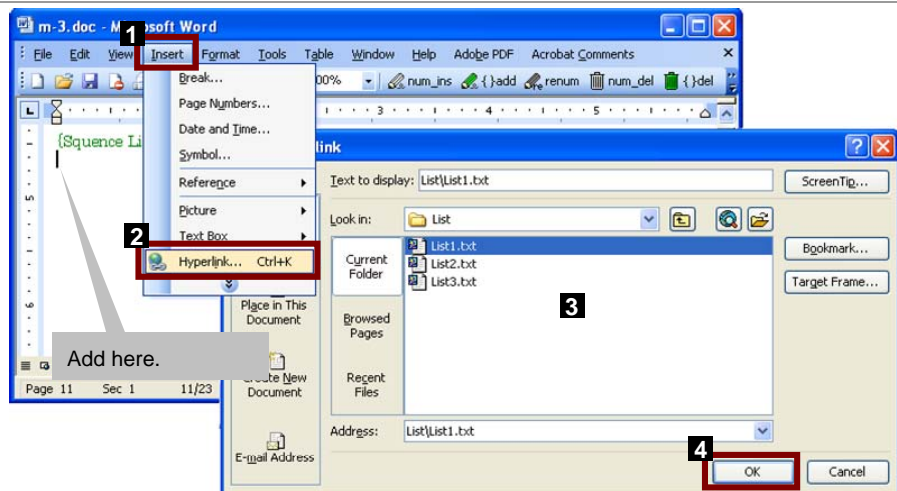
Word 2007

- 1 Select [Insert].
- 2 Select [Hyperlink].
- 3 Select a file.
- 4 Click [OK].



Word 2003

- 1 Select [Insert].
- 2 Select [Hyperlink].
- 3 Select a file.
- 4 Click [OK].



Chapter 4 Typical Error Messages

4.1 Typical Error Dialog Boxes

Clicking [Convert to XML] displays an error (1).	
Message	V6XCF03007 The path including the em-size character is invalid.
Cause	An HTML file specified with HTML2XML includes non-ASCII character code.
Response	Change the file path and name so as not include a non-ASCII character. Then convert the file again.

Clicking [Convert to XML] displays an error (2).	
Message	<ul style="list-style-type: none"> • V6XCF04005 Please Set Country. • V6XCF04006 Please Set Language.
Cause	Information is insufficient in specifying HTML2XML.
Response	Specify the missing information according to the error message. Then convert the file again.

Clicking [Display application-body] displays an error (1).	
Message	V6XCF04204 The PDF display software is not installed, the detailed statement XML cannot be displayed.
Cause	The PDF display tool is not installed.
Response	Install the PDF display tool such as Adobe Reader.

Clicking [Display application-body] displays an error (2).	
Message	V6XCF04101 Because PCT-SAFE Information cannot be acquired, the detailed statement cannot be displayed.
Cause	PCT-SAFE is not installed.
Response	Install PCT-SAFE.

Clicking [Run PCT-SAFE] does not start PCT-SAFE.	
Message	V6XCF04102 Because PCT-SAFE Information cannot be acquired, it cannot be started.
Cause	PCT-SAFE is not installed.
Response	Install PCT-SAFE.

4.2 Typical Errors Displayed by [Display Check result]

Message sample	V6XCB04019 The character code of the input file is invalid. Only UTF-8 format.
Cause	The HTML file contains a non-UTF-8 code.
Response	To save a document in HTML with Word, use Unicode(UTF-8). Because the default value is Unicode in Word 2007, be sure to change it. For details on the procedure, see Section 2.3, "Saving a document in HTML Format."
Message sample	Error: V6XCBR6004 The Item is not described in the document. The beginning of documents must be the [item name].
Cause	The heading is not enclosed in { }.
Response	Enclose the heading in { }. For details on the procedure using Word macros, see Appendix 2, "Using Word Macros."
Message sample	Error: V6XCBS3001 A certain problem exists in generated XML. Necessary item doesn't exist. Please confirm the part before <claims>.(line=67)[(doc-page (description,claims,abstract),drawings))]
Cause	The statement order is incorrect ("claims" is stated first in the case of this message sample.)
Response	Follow the sequence, description, claims, abstract, and drawings. For details on common formats, see Appendix 3, "Headings Writable to Documents."
Message sample	<ul style="list-style-type: none"> • Error: V6XCBI4510 The Image file is missing. image file name [F00F02_1.BMP] (line=65) • Error: V6XCBI2100 The link file of the sequence listing is not found. You must check the file. file name [Sequence_List.APP] information[The HTML file and The sequence listing files are put in the same folder.] (line=63)
Cause	The file, linked from HTML, is not found. The image and array table files may have been deleted or moved.
Response	If a link destination file is deleted or moved, a document saved in HTML is disconnected from it. If the file is found deleted or moved, replace it. If it cannot be replaced, re-create the correct set of HTML files.

Appendix

Appendix 1 Installation and Uninstallation

Installation


Note:

Before installing HTML2XML, install PCT-SAFE.

- 1) Download the HTML2XML installer.

Hint:

The installer can be downloaded from the following URL:
[http:// www.inpit.go.jp/english/](http://www.inpit.go.jp/english/)

- 2) To start installation, double-click the installer icon ().


Hint:

- Do not install the converter in a common folder.
- In Windows Vista, the functions (installer and environment setting) needed to use a registry are enabled, the entire screen becomes dark, and the user account control window is displayed. In the user account control window, be sure to check that the issuer is National Center for Industrial Property Information and Training, and then click the [Continue] button. To interrupt installation, click the [Cancel] button.
- If you overwrite the installed HTML2XML, the installation destination cannot be changed. To change the installation destination, uninstall the converter once, and then install it again.

Uninstallation

Note:

- **Before starting uninstallation, disable all HTML2XML functions.**
- **After quitting HTML2XML, you should immediately restart Windows and then start uninstallation.**
- **HTML and XML files are not deleted. If these files are not required, delete them manually.**

- 1) From [Start] () , open Control Panel.
- 2) Click [Add or Remove Programs].
- 3) From the list, select [HTMLtoXML Converter] and click [Remove].

Appendix 2 Using Word Macros

To convert a document with HTML2XML, paragraph numbers must be inserted and headings enclosed in { } in the HTML document for input. To automate this operation, use the Word macros attached to HTML2XML as samples.

This section explains how to use these macros.

1) Double-click the Word macro icon.






Hint:

This folder provides the Word macros together with the templates.

2) Enter document contents.

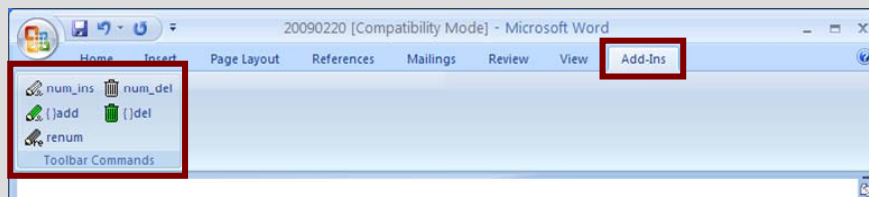
3) Click the Word macro buttons.

The function of each button is as follows:

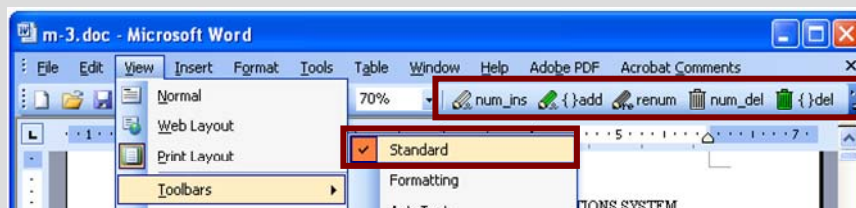
 num_ins	Inserts paragraph numbers.	 num_del	Deletes paragraph numbers.
 { }add	Adds { }.	 { }del	Deletes { }.
 renum	Reassigns paragraph numbers. Use this function when changing the statement sequence or deleting/adding paragraph numbers.		

Hint:

- Word 2007 includes the buttons in [Add-Ins].



- Word 2003 includes the buttons in [Standard] in [Toolbars].



Note:

Depending on the user environment, Word security settings restrict macro execution. If the following error occurs during macro execution, review security risks, and change the setting by following Help message instructions.



Appendix 3 Headings Writable to Documents

This section explains common formats for the following:

- Document names
- Headings that can be entered next to document names (and whether paragraph numbers are included)
- Item contents that can be entered next to headings

Headings that can be entered

Document name	Heading	Paragraph number	Heading or item content within Paragraph number			
{Description}	{Title of Invention} or {Title}		Character string (lowercase characters not allowed) *4			
	{Technical Field} or {Field}	{9999} *2	Character string *4			
			Image			
			{Chem. n} *3	Image		
			{Math. n} *3	Image		
			{Table n} *3	Image		
	{Background Art} or {Background}	{9999} *2	Same as above *1			
	{Summary of Invention} or {Summary}	{Technical Problem}	{9999} *2	Same as above *1		
				{Solution to Problem}	{9999} *2	Same as above *1
						{Advantageous Effects of Invention}
		{Brief Description of Drawings}	{9999} *2	{Fig. n} *3	Character string *4	
	{Description of Embodiments}	{Examples} or {Example 9} *2	{9999} *2	Same as above *1		
				{Industrial Applicability}	{9999} *2	Same as above *1
	{Reference Signs List}		{9999} *2	Same as above *1		
	{Reference to Deposited Biological Material}		{9999} *2	Same as above *1		

{Sequence Listing Free Text}		{9999} ^{*2}	Same as above ^{*1}
{Citation List}			
		{Patent Literature}	
		{9999} ^{*2}	{PTL n} ^{*3}
			Character string
		{Non Patent Literature}	
		{9999} ^{*2}	{NPL n} ^{*3}
			Character string
{Sequence Listing}			Link
{Claims}			
		{Claim 9} ^{*3}	Same as above ^{*1}
{Abstract}			
			Character string ^{*4}
			Image
{Drawings}			
		{Fig. n} ^{*3}	Image

*1 Refer to {Technical Field}

*2 Enter 9 or 9999 in serial numbers. Left-justify 9999 with zeros.

*3 Enter "n" using alphanumeric characters, "-", ".", and "(", ")".

*4 As character modifiers, an underline (<U>-</U>), superscript (⁻) and subscript (₋) can be used.

Appendix 4 Usable Images

Enter drawings, tables, mathematical expressions, chemical formulas, and external characters as image data (except where expressed by modifiers such as superscripts and subscripts). Note that the usable images are restricted. The restrictions are as follows:

Usable image format	Color	Pixel density and maximum image size (W x H: X x Y)
BMP	Monochrome binary If the color is not monochrome binary, an error occurs.	300 dpi and 400 dpi (400 dpi when the number of dots exceeds 300 dpi) mm conversion : 170 x 255 mm 300 dpi : 2007 x 3011 dots 400 dpi : 2677 x 4015 dots
GIF	Monochrome binary A non-monochrome binary color is converted by HTML2XML. Check the result by printing from PCT-SAFE.	
JPEG (JFIF only) If the format is not JFIF, an error occurs.	Gray scale	

Hint:

- Attached images are converted respectively as shown below by HTML2XML. Print the images from PCT-SAFE and check whether the display is correct.
<BMP and GIF>TIFF binary and <JPEG>
- HTML2XML judges an image size from the number of dots (ignoring the image resolution (dpi information)). If the number of image dots is known, HTML2XML can determine size (cm) using the following calculation expressions (1 inch = 2.54 cm).
Image of 300 dpi: number of dots/300 x 2.54 ≈ length (cm)
Image of 400 dpi: number of dots/400 x 2.54 ≈ length (cm)
- If an original image is created at 300 dpi, but the number of dots in the width exceeds 2,007 and the number of dots in the height exceeds 3,011, 400 dpi is assumed.
- The size of a gray scale JPEG becomes 8 to 10 times greater than that of monochrome binary GIF or BMP format. To create graphic charts, diagrams, and chemical formulas other than substitute photographs for drawings, use GIF and BMP.

Appendix 5 Usable File Paths

Only ASCII characters can be used in Word, HTML and XML file names. All characters included in a file path must also be ASCII characters.

Appendix 6 HTML Tags Supported by HTML2XML

This section explains the HTML tags supported by HTML2XML.

Chapter 2 explained how to create an HTML file with Word. To create an HTML file using other methods, only use the tags supported by HTML2XML.

Hint:

Uppercase and lowercase tags may be mixed.

Basic structure tags

One set of these tags is required for each HTML document.

Beginning and end of HTML document	<HTML>entire document</HTML>	<ul style="list-style-type: none"> • Enter the tags in this sequence. • Do not enter characters before the <HTML> tags.
Beginning and end of header part	<HEAD>header part of document</HEAD>	
Beginning and end of title	<TITLE>document title</TITLE>	
Beginning and end of document entity	<BODY>entire document content</BODY>	

Character modifier tags

These tags modify characters.

Beginning and end of underline	<U>target character</U>	A superscript and subscript cannot be used together with one character.
Beginning and end of superscript	^{target character}	
Beginning and end of subscript	_{target character}	
Beginning and end of pre-format	<PRE>target character</PRE>	

Line feed tags

These tags are used to specify line feed position in documents.

Line feed tag	 	If there is no but a line feed control code after an effective display character, the line feed control code is handled similar to (except for English description when "Suppress unintended forcible line feed in English description" is specified in the environment setting).
Line feed control code	0x0d0x0a	

Image tags

These tags are used to include image data such as figures and tables in documents.

File name (path may be included) to included image		<ul style="list-style-type: none"> • Leave a single-byte space between IMG and SRC. • Do not place another image tag or character (including space) on the same line with an image tag.
--	------------------------------------	---

Link tags

These tags are used to link an array table file into a document.

File name (path may be included) to link destination file		-
---	--	---