DOCX2XML Converter Functional Specification Document

#### DOCUMENT VERSION 1.1

#### Working draft for consultation

Table of Contents

[1. Introduction 2](#_Toc211870294)

[1.1. Purpose 2](#_Toc211870295)

[1.2. Project Scope Details 2](#_Toc211870296)

[1.3. Assumptions, Recommendations and Constraints 3](#_Toc211870297)

[2. Requirements Overview 4](#_Toc211870298)

[2.1. Use case 4](#_Toc211870299)

[3. Software detail design 5](#_Toc211870300)

[3.1. XML Resources 5](#_Toc211870301)

[3.2. Filing System and Converter Detailed Design Example 6](#_Toc211870302)

[3.3. Exception Handling and Error Reporting 6](#_Toc211870303)

[4. References 7](#_Toc211870304)

[ANNEX I: IP Office DOCX2XML Conversion Details 8](#_Toc211870305)

[ANNEX II: Section Headings for Patent Applications in DOCX 9](#_Toc211870306)

# Introduction

At its tenth session held in November 2022, the Committee on WIPO Standards (CWS) approved that the Digital Transformation Task Force (DTTF) explore DOCX to XML conversion as an option for Offices (see paragraphs 105 and 106 of document CWS/10/22). Based on the DTTF survey results discussed in document CWS/10/15, almost half the respondents indicated that they accept DOCX files. At its eleventh session held in December 2023, the CWS noted that the DTTF had conducted a review and performed an analysis of the functionality of DOCX to XML converters in use at the United States Patent and Trademark Office and at the International Bureau for ePCT. At the same session, the CWS approved the DTTF to prepare a proposal for the recommendation on a common requirements specification for a DOCX to XML converter (see paragraphs 56 to 58 of document CWS/11/28). It would be very beneficial for the intellectual property (IP) offices to use a conversion tool with commonly agreed-upon functionality to support the efforts of accepting and processing patent application information in structured text format, thereby advancing towards the common goal of using WIPO Standard ST.36 or ST.96 as the standard XML format. This document and its Annex I consolidate and summarize the feedback provided by IP offices about the functionality of the DOCX converter used at their Office. It is relevant to note that more than half of the survey respondents indicated that they provide initial content-based validation information to the applicants, at the time of filing, prior to submission. Annex I includes a list of existing DOCX converters, outlining consistencies and differences for reference.

## Purpose

The purpose of this document is to provide a description of the functional requirements for a DOCX2XML converter, which will convert patent applications submitted in DOCX format to XML. The converter should generate output in compliance with WIPO Standard ST.36 or ST.96. The DOCX is a word processing file format based on the Office Open XML (OOXML) and is governed by international standards ECMA-376 and ISO/IEC 29500.

## Project Scope Details

A user should be able to create a new submission with one or more related DOCX documents via a filing system. If a submitted document results in DOCX validation errors or warnings in the filing system, it should return feedback to the submitter for resubmission. The filing system may generate and send a corrected DOCX file to the DOCX2XML converter for processing and conversion. After the conversion process, the system should return a success or failure feedback message to the reviewer at the IP office. Failure message covers non-compliance to business rules which happen before actual conversion process. Additionally, this document also explains how content such as tables and math equations should be converted into XML as referred to in section 3.1 below.

Note: Corrected DOCX refers to producing a DOCX copy (to be used for transformation to ST.36 or ST.96 XML) from the original copy by applying all the business rules and removing any metadata like author, comments, bookmarks, and making list numbering in the original document explicit.

### Supported input document types

Documents produced by editors that can save or export the content in the DOCX format are the following:

* Microsoft Word 2007 or higher (for the PC or Mac);
* Google docs;
* Office online;
* LibreOffice; and
* Pages for Mac.

## Assumptions, Recommendations and Constraints

* The conversion process may not be able to convert all DOCX features to the WIPO Standard ST.36 or ST.96 format, and practically support only a very limited range of basic formatting. Notably, this does not include fonts, font color, font size or any form of specific positioning of text elements on the page such as tabs, other than some limited support through tables, using images instead of text or the <pre> tag.
* This document does not specify any specific content management system solutions.
* DOCX file should contain section headers as specified in Annex II.

# Requirements Overview .96

A screenshot of a computer

AI-generated content may be incorrect.

The **above diagram** is an example of a DOCX application conversion process system workflow, and is described in detail below.

Upon the user uploading the DOCX document using a filing system, the DOCX2XML converter will process the conversion. The following steps illustrate the associated flow:

1. Filing System validates the DOCX document against IP office business rules and returns any Error(s) or Warning(s) if they are present;
2. Corrected DOCX generated by filing system will be provided to DOCX2XML converter; and
3. If no Error(s) or Warning(s) are identified during the validation process, the converter service then converts the corrected DOCX document into XML, and the XML document should be validated against WIPO Standard ST.36 or ST.96 schemas.

## Use case

The use case below describes how the filing system processes the DOCX document. The filing system may split the DOCX document into multi section documents based on the section header (Claims, Description, Abstract and Drawings). Please see Annex II in this document for section headings.

### DOCX Document Processing Flow

1. Submitter uploads DOCX file as application content to a filing system;
2. Filing system performs validations (file name, extension, etc.);
3. Filing system may convert original DOCX to corrected DOCX (after compliance with business rules);
4. Corrected DOCX Success or Failure feedback is sent to the submitter;
5. Filing system stores original DOCX (prior to correction);
6. Corrected DOCX is converted to an XML compliant with WIPO Standard ST.36 or ST.96 via DOCX2XML converter; and
7. WIPO ST.36 or ST.96 XML conversion Success or Failure feedback is sent to the IPO reviewer.

# Software detail design

## XML Resources

### Mapping of Major DOCX Constructs to WIPO ST.36 or ST.96

| **Major DOCX Construct** | **XML equivalent Construct** |
| --- | --- |
| Tables | (OASIS CALS) tables |
| Math equations | Embedded math equation object will be converted into MathML format, others will be converted to image |
| ChemDraw objects (some IP offices may not support ChemDraw) | Converted to image |
| Drawings, SmartArt, charts, shapes (some IP offices may not support) | Converted to image |
| JPEG, TIFF, PNG, GIF, EMF, WMF and SVG images (some IP offices may not support SVG format) | Converted to image |

## Filing System and Converter Detailed Design Example

DOCX2XML filing system and converter process flow chart

A screenshot of a computer

AI-generated content may be incorrect.

## Exception Handling and Error Reporting

Below is a table of errors and warnings which may occur during DOCX processing.

**DOCX filing system and converter service errors and warnings:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Message Category** | **Examples** | **Corrected DOCX** | **Generated XML** |
| Error: system level | Virus or password protected | No | No |
| Error: content level | No parts, invalid font | No | No |
| Warning: uncorrected | Oversized abstract | Yes | Yes |
| Warning: corrected | Remove comments, track changes | Yes | Yes |
| Error: no WIPO ST.36 or ST.96 XML conversion | Unsupported image format, validation errors | Yes | No |
| Happy Path | Most cases | Yes | Yes |

# References

ISO/IEC 29500 <https://www.iso.org/standard/71691.html>

ECMA-376 <https://ecma-international.org/publications-and-standards/standards/ecma-376/>

WIPO Standard ST.36 <https://www.wipo.int/documents/d/standards/docs-en-03-36-01.pdf>

WIPO Standard ST.96 <https://www.wipo.int/documents/d/standards/docs-en-03-96-01.pdf>

# ANNEX I: IP Office DOCX2XML Conversion Details

The IP offices DOCX2XML conversion details (XSLX), is available at: <https://www.wipo.int/edocs/mdocs/cws/en/cws_13/cws_13_12-annexii.xlsx>

# ANNEX II: Section Headings for Patent Applications in DOCX

Annex II provides exemplary section headings as a guide for filing submissions based on WIPO ST.36 or ST.96. These headings will assist the converter in auto-detection of the various sections of the application such as the Description, Claims, Abstract and Drawings. A generic heading element can be used for headings not in the list below.

|  |  |
| --- | --- |
| **Item Number** | **Section Headings** |
| 1 | Technical Field |
| 2 | Background Art |
| 3 | Disclosure |
| 4 | Invention Summary |
| 5 | Drawing Description |
| 6 | Embodiment Description |
| 7 | Best Mode |
| 8 | Invention Mode |
| 9 | Industrial Applicability |
| 10 | Reference Sign |
| 11 | Deposited Biological Material |
| 12 | Sequence Listing |
| 13 | Citation |
| 14 | Claims |
| 15 | Abstract |
| 16 | Drawings |

[End of Annex and of document]