Spécification fonctionnelle du convertisseur DOCX2XML

#### Version 1.1

#### Projet de document de travail pour consultation

Table des matières

[1. Introduction 2](#_Toc212717308)

[1.1. Objet 2](#_Toc212717309)

[1.2. Précisions relatives à la portée du projet 2](#_Toc212717310)

[1.3. Hypothèses, recommandations et contraintes 3](#_Toc212717311)

[2. Aperçu des exigences .96 4](#_Toc212717312)

[2.1. Cas d’utilisation 4](#_Toc212717313)

[3. Conception détaillée du logiciel 5](#_Toc212717314)

[3.1. Ressources XML 5](#_Toc212717315)

[3.2. Exemple de conception détaillée d’un système de dépôt et d’un convertisseur 6](#_Toc212717316)

[3.3. Gestion des exceptions et signalement des erreurs 7](#_Toc212717317)

[4. Références 7](#_Toc212717318)

[ANNEXE I: Détails de la conversion DOCX2XML par office de propriété intellectuelle 8](#_Toc212717319)

[ANNEXE II: Titres des sections pour les demandes de brevet dans DOCX 9](#_Toc212717320)

# Introduction

À sa dixième session, tenue en novembre 2022, le Comité des normes de l’OMPI (CWS) s’est déclaré favorable à ce que l’Équipe d’experts chargée de la transformation numérique de se penche sur les possibilités de conversion du format DOCX en XML pour les offices (voir les paragraphes 105 et 106 du document CWS/10/22). D’après les résultats de l’enquête réalisée par l’équipe d’experts et présentés dans le document CWS/10/15, près de la moitié des offices ayant répondu ont indiqué accepter les fichiers DOCX. À sa onzième session, tenue en décembre 2023, le CWS a noté que l’équipe d’experts avait procédé à un examen et réalisé une analyse de la fonctionnalité des convertisseurs DOCX vers XML utilisé par l’Office des brevets et des marques des États‑Unis d’Amérique et le Bureau international pour le système ePCT. À cette même session, le CWS s’est déclaré favorable à ce que l’équipe d’experts établisse une proposition de recommandation relative à une spécification d’exigences communes pour un convertisseur DOCX vers XML (voir les paragraphes 56 à 58 du document CWS/11/28). Il serait très avantageux pour les offices de propriété intellectuelle d’utiliser un outil de conversion doté de fonctionnalités communément admises afin de faciliter l’acceptation et le traitement des informations relatives aux demandes de brevet sous forme de texte structuré, ce qui permettrait de progresser vers l’objectif commun consistant à utiliser les normes ST.36 ou ST.96 de l’OMPI comme format XML standard. Le présent document et son annexe I regroupent et résument les commentaires fournis par les offices de propriété intellectuelle concernant les fonctionnalités du convertisseur DOCX utilisé dans leur office. Il convient de noter que plus de la moitié des offices ayant répondu à l’enquête ont indiqué qu’ils fournissaient aux déposants des informations initiales de validation basées sur le contenu, au moment du dépôt, avant la soumission. L’annexe I contient une liste des convertisseurs DOCX existants, indiquant les similitudes et les différences à titre de référence.

## Objet

Le présent document a pour objet de décrire les exigences fonctionnelles d’un convertisseur DOCX2XML, qui convertira les demandes de brevet soumises au format DOCX en XML. Le convertisseur doit générer un résultat conforme à la norme ST.36 ou ST.96 de l’OMPI. Le DOCX est un format de fichier de traitement de texte fondé sur Office Open XML (OOXML) et régi par les normes internationales ECMA‑376 et ISO/IEC 29500.

## Précisions relatives à la portée du projet

L’utilisateur doit être en mesure de créer une nouvelle soumission avec un ou plusieurs documents DOCX connexes via un système de dépôt. Si le document soumis donne lieu à des erreurs ou des avertissements de validation DOCX dans le système de dépôt, un message doit être renvoyé à l’auteur afin qu’il le soumette à nouveau. Le système de dépôt peut générer un fichier DOCX corrigé et l’envoyer au convertisseur DOCX2XML pour traitement et conversion. Après la conversion, le système doit renvoyer un message indiquant la réussite ou l’échec de l’opération au réviseur de l’office de propriété intellectuelle. Le message d’échec concerne la non‑conformité aux règles métier qui s’appliquent avant le processus de conversion proprement dit. De plus, le présent document explique également comment les contenus tels que les tableaux et les équations mathématiques doivent être convertis en XML, comme indiqué à la section 3.1 ci‑dessous.

Note : Le fichier DOCX corrigé désigne la production d’une copie DOCX (qui sera utilisée pour la conversion au format XML selon la norme ST.36 ou ST.96) à partir de la copie originale, en appliquant toutes les règles métier et en supprimant toutes les métadonnées concernant l’auteur, les commentaires, les signets, etc., et en rendant explicite la numérotation des listes dans le document original.

### Types de documents pris en charge

Les documents créés par des éditeurs pouvant enregistrer ou exporter le contenu au format DOCX sont les suivants :

* Microsoft Word 2007 ou version ultérieure (pour PC ou Mac);
* Google docs;
* Office en ligne;
* LibreOffice; et
* Pages pour Mac.

## Hypothèses, recommandations et contraintes

* Il se peut que le processus de conversion ne permette pas de convertir toutes les fonctionnalités DOCX au format de la norme ST.36 ou ST.96 de l’OMPI et ne puisse prendre en charge, dans la pratique, qu’un nombre très limité de formats de base. Il convient de noter que cela n’inclut pas les polices, la couleur ou la taille des polices, ni aucune forme de positionnement spécifique des éléments de texte sur la page, telle que les onglets, à l’exception d’une prise en charge limitée via des tableaux, l’utilisation d’images à la place du texte ou la balise <pre>.
* Ce document ne définit aucune solution spécifique en matière de système de gestion de contenu.
* Le fichier DOCX doit contenir les titres des sections tels que spécifiés à l’annexe II.

# Aperçu des exigences .96



Le **diagramme ci‑dessus** est un exemple de flux de travail d’un système de conversion d’applications DOCX, décrit en détail ci‑dessous.

Une fois que l’utilisateur a téléchargé le document DOCX à l’aide d’un système de dépôt, le convertisseur DOCX2XML effectue la conversion. Les étapes suivantes illustrent le processus correspondant :

1. le système de dépôt valide le document DOCX par rapport aux règles métier de l’office de propriété intellectuelle et renvoie les éventuelles erreurs ou avertissements s’il y en a;
2. le fichier DOCX corrigé généré par le système de dépôt sera fourni au convertisseur DOCX2XML; et
3. si aucune erreur ou aucun avertissement n’est identifié pendant le processus de validation, le service de conversion convertit alors le document DOCX corrigé en XML, et le document XML doit être validé par rapport aux schémas relevant des normes ST.36 ou ST.96 de l’OMPI.

## Cas d’utilisation

Le cas d’utilisation décrit ci‑dessous explique comment le système de dépôt traite le document DOCX. Le système de dépôt peut diviser le document DOCX en plusieurs sections en fonction du titre de la section (revendications, description, abrégé et dessins). Veuillez consulter l’annexe II du présent document pour connaître les titres des sections.

### Flux de traitement des documents DOCX

1. L’auteur télécharge le fichier DOCX en tant que contenu de la demande dans un système de dépôt.
2. Le système de dépôt effectue des validations (nom de fichier, extension, etc.).
3. Le système de dépôt peut convertir le fichier DOCX original en fichier DOCX corrigé (après conformité avec les règles métier).
4. Un message indiquant la réussite ou l’échec de l’opération de conversion du fichier DOCX est envoyé à l’auteur.
5. Le système de dépôt conserve le fichier DOCX original (avant correction).
6. Le fichier DOCX corrigé est converti en un fichier XML conforme à la norme ST.36 ou ST.96 de l’OMPI à l’aide du convertisseur DOCX2XML.
7. Un message indiquant la réussite ou l’échec de la conversion XML conforme à la norme ST.36 ou ST.96 de l’OMPI est envoyé au réviseur.

# Conception détaillée du logiciel

## Ressources XML

### Correspondance entre les principales structures DOCX et la norme ST.36 ou ST.96 de l’OMPI

| **Principale structure DOCX** | **Structure XML équivalente** |
| --- | --- |
| Tableaux | Tableaux (OASIS, CALS) |
| Équations mathématiques | Les objets sous forme d’équations mathématiques intégrées seront convertis au format MathML, les autres seront converties en images. |
| Les objets ChemDraw (certains offices de propriété intellectuelle peuvent ne pas prendre en charge ChemDraw) | Conversion en image |
| Les objets ChemDraw (certains offices de propriété intellectuelle peuvent ne pas prendre en charge ChemDraw) | Conversion en image |
| Dessins, SmartArt, graphiques, formes (certains offices de propriété intellectuelle peuvent ne pas les prendre en charge) | Conversion en image |

## Exemple de conception détaillée d’un système de dépôt et d’un convertisseur

Schéma fonctionnel du système de dépôt DOCX2XML et du convertisseur



## Gestion des exceptions et signalement des erreurs

Vous trouverez ci‑dessous un tableau répertoriant les erreurs et avertissements pouvant survenir lors du traitement des fichiers DOCX.

**Erreurs et avertissements liés au système de dépôt DOCX et au service de conversion :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Catégorie de message** | **Exemples** | **DOCX corrigé** | **XML généré** |
| Erreur : niveau système | protection antivirus ou mot de passe | Non | Non |
| Erreur : niveau contenu | aucune partie, police non valide | Non | Non |
| Avertissement : non corrigé | abrégé trop long | Oui | Oui |
| Avertissement : corrigé | supprimer les commentaires, changements apparents | Oui | Oui |
| Erreur : pas de conversion XML conforme à la norme ST.36 ou ST.96 de l’OMPI | format d’image non pris en charge, erreurs de validation | Oui | Non |
| Chemin heureux | La plupart des cas | Oui | Oui |

# Références

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

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

Norme ST.36 de l’OMPI [https://www.wipo.int/documents/d/standards/docs‑fr‑03‑36‑01.pdf](https://www.wipo.int/documents/d/standards/docs-fr-03-36-01.pdf)

Norme ST.96 de l’OMPI [https://www.wipo.int/documents/d/standards/docs‑fr‑03‑96‑01.pdf](https://www.wipo.int/documents/d/standards/docs-fr-03-96-01.pdf)

# ANNEXE I : Détails de la conversion DOCX2XML par office de propriété intellectuelle

Les détails de conversion DOCX2XML par office de propriété intellectuelle peuvent être consultés à l’adresse suivante : [https://www.wipo.int/edocs/mdocs/cws/fr/cws\_13/cws\_13\_12‑annexii.xlsx](https://www.wipo.int/edocs/mdocs/cws/fr/cws_13/cws_13_12annexii.xlsx)

# ANNEXE II : Titres des sections pour les demandes de brevet dans DOCX

L’annexe II fournit des exemples de titres de sections à titre indicatif pour le dépôt de demandes sur la base des normes ST.36 ou ST.96 de l’OMPI. Ces titres aideront le convertisseur à détecter automatiquement les différentes sections de la demande, telles que la description, les revendications, l’abrégé ou les dessins. Un titre générique peut être utilisé pour les titres qui ne figurent pas dans la liste ci‑dessous.

|  |  |
| --- | --- |
| **Numéro d’élément** | **Titres des sections** |
| 1 | Domaine technique |
| 2 | Technique antérieure |
| 3 | Divulgation |
| 4 | Résumé de l’invention |
| 5 | Description des dessins |
| 6 | Description d’un mode de réalisation |
| 7 | Meilleur mode d’exécution |
| 8 | Mode de réalisation de l’invention |
| 9 | Possibilité d’application industrielle |
| 10 | Signe de référence |
| 11 | Matériel biologique déposé |
| 12 | Listage des séquences |
| 13 | Citation |
| 14 | Revendications |
| 15 | Abrégé |
| 16 | Dessins |

[Fin de l’annexe et du document]