Brevets français contemporains

Texte intégral océrisé

Format et structure des informations
Description des fichiers

Date de création : 05/03/2013 – Version 1.2

Auteur : Fenny Versloot-Spoelstra

Europäisches Patentamt
European Patent Office
Office européen des brevets

INPI

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

SOMMAIRE

1. PERIMETRE

La licence « **Brevets français contemporains – Texte intégral océrisé** » contient le texte complet océrisé de la description et des revendications des demandes de brevets français publiées par l'INPI **à partir de 1981**.

Ces données sont au format normalisé **XML ST36**.

Evolution des formats

Le format d'échange décrit dans ce document **remplace le format d'origine** qui reposait sur des extractions des bases de données « Texte intégral EPOQUE » structurées selon le format propriétaire EBCDIC de l'OEB.

Le nouveau format d'échange repose sur **des extractions issues de la base de données de référence de l'OEB pour les documents en texte intégral** (FTM : Full Text Master database). Les données sont au format normalisé **XML OMPI/ST36 (WIPO/ST36)**.

Fichiers de l'arriéré / fichiers courants

La base de données de référence de l'OEB pour les documents en texte intégral (FTM) est une base de données relativement récente, totalement opérationnelle depuis 2008/2009. Elle a remplacé les bases de données « Texte intégral EPOQUE » (format EBCDIC).

Lors de la migration des nombreuses bases de données d'origine vers la nouvelle base de données de référence FTM, deux stratégies ont été appliquées. Lorsque cela a été possible, la base de données de référence pour le texte intégral (FTM) a été reconstruite en partant de la source d'origine et en rechargeant les fichiers d'arriéré avec le nouveau format. Dans certains cas, le contenu d'origine des bases EPOQUE a dû être directement reconverti au nouveau format.

Par conséquent, la richesse du XML peut varier selon qu'il s'agit de données issues de fichiers de l'arriéré ou de fichiers courants. Les données des fichiers courants seront systématiquement produites avec la richesse offerte par la définition du nouveau schéma XML. Les documents plus anciens des fichiers de l'arriéré peuvent dans certains cas, être structurés dans un format XML limité du fait des contraintes du format propriétaire d'origine EBCDIC.

2. FORMAT D'ECHANGE

A. CARACTERISTIQUES

Le format d'échange du texte intégral contient :

1. **Les données bibliographiques suivantes :**
   - 1.1. Identifiant de publication
   - 1.2. Date de publication
   - 1.3. Identifiant de dépôt
   - 1.4. Date de dépôt

2. **Les différentes parties du texte intégral :**
   - 2.1. Description
   - 2.2. Revendications
   - 2.3. Abrégé

3. **Le traitement de l'attribut "status"**

Contenu

L'extraction est limitée à une extraction en texte intégral par publication.

Si la même publication a été fournie plusieurs fois à l'OEB, le mécanisme d'export consiste à sélectionner les données fournies par la source qui est de meilleure qualité.

L'extraction est limitée aux fournisseurs et aux sources avec lesquels l'OEB a un accord pour diffuser l'information à des tiers.

"Status"

L'attribut **"status" ou " field level change indicator "** sert à indiquer quelles informations ont été modifiées depuis le dernier échange.

B. DONNEES BIBLIOGRAPHIQUES

L'attribut **"status"** sur les données bibliographiques est renseigné par le status=D - <bibliographic-data status="D"> - lorsque :

- une publication a été supprimée de la base de données DOCDB
- une publication est devenue "void" dans DOCDB

Si, de manière exceptionnelle, tout le texte intégral d'une publication donnée a été retiré, l'attribut "status" au niveau de <bibliographic-data> sera aussi renseigné avec la valeur "D".

Un nouveau codage de l'identifiant de publication dans la base DOCDB est renseigné par une combinaison des status=D et =C :

- <bibliographic-data status="D"> pour l'ancien identifiant
- <bibliographic-data status="C"> pour le nouvel identifiant

Toute modification des autres données bibliographiques – modification de la date de publication, de l'identifiant de dépôt ou de la date de dépôt - sera renseignée par le status=A.

REMARQUE : *les modifications apportées aux données bibliographiques seront signalées uniquement pour les publications dont le texte intégral est disponible dans la base FTM.*

C. PARTIES DU TEXTE INTEGRAL

L'attribut "status" au niveau des différentes parties du texte intégral (description, revendications, abrégé) sera renseigné par :

- status="C" lorsqu'une partie a été ajouté
- status="A" lorsqu'une partie a été remplacé

REMARQUE : *le status="D" n'est pas prévu. La suppression d'une partie du texte intégral entraine de façon implicite qu'elle ne sera plus incluse dans l'extraction.*

Composant Texte intégral en plusieurs langues

Quand une partie de texte intégral est ajoutée ou remplacée pour une langue donnée, **l'attribut «status» est renseigné dans cette langue**, au niveau de la partie correspondante. Par exemple :

- pour la partie description en langue EN d'un document qui est créé :
  <description lang="de"> ... </description>
  <description lang="en" status='C'> ... </description>
  <description lang="fr"> ... </description>
- pour la partie description en langue DE d'un document qui a été remplacé :
  <description lang="de" status='A'> ... </description>
  <description lang="en"> ... </description>

<description lang="fr"> ... </description>

D. FORMATAGE ET VALIDATION

L'échange reposera toujours sur un **document complet** comprenant la série complète des images – si disponibles – et les différentes parties composant le texte intégral du document

Images

Les images sont au **format TIFF**.
Chaque image associée au document en texte intégral est unique, les doublons sont éliminés. Seules les images qui sont pleinement référencées dans le texte sont prises en compte.

Texte

Le texte est au format UTF8.
Les documents sont soumis à une validation XML exhaustive et tout caractère qui ne serait pas au format UTF8 est détecté lors de cette étape.

Cette validation permet de vérifier :
- que le texte est bien formé
- l'encodage UTF-8
- la conformité au schéma fulltext-documents.xsd

E. TRAITEMENT DES ERREURS

Les documents présentant des irrégularités sont inclus. La décision d'éliminer le document ou de le traiter tel quel est laissée à l'utilisateur. Ces documents comportant des erreurs sont fournis aux utilisateurs de manière séparée. Ils sont en outre stockés par l'OEB qui les analyse et décide de la suite à y donner.

Images

Certaines images peuvent être correctement référencées dans un document XML mais ne pas être physiquement présentes dans la base de données FTM.

Texte

SI le texte de certains documents n'est pas validé pour des questions de forme ou de conversion de caractères, la partie du texte intégral concernée est encapsulée dans CDATA.

Eléments du texte intégral non inclus

Les éléments du texte intégral suivants ne sont pas traités :
- <drawing> : dessin
- <sequence-listing> : liste de séquences
- <tables-external-doc> : tableaux

Abrégé <abstract>

L'abrégé (<abstract>) est inclus uniquement s'il est contenu dans les données en texte intégral fournies à l'OEB. Pour certains pays, le texte intégral fourni ne comporte pas d'abrégé alors qu'on s'attend à en trouver un. Il s'agit d'une question de conception et non d'une erreur.

Élément <application-reference> - Attribut "doc-id"

L'attribut "doc-id" a été introduit en vue d'une prochaine utilisation. Il comportera un identifiant unique et stable qui permettra, dans le futur, de faire le lien de manière fiable entre diverses bases de données de l'OEB.

3. ORGANISATION DES FICHIERS

La structure d'organisation des fichiers livrés est conforme à la norme de l'OEB en la matière.

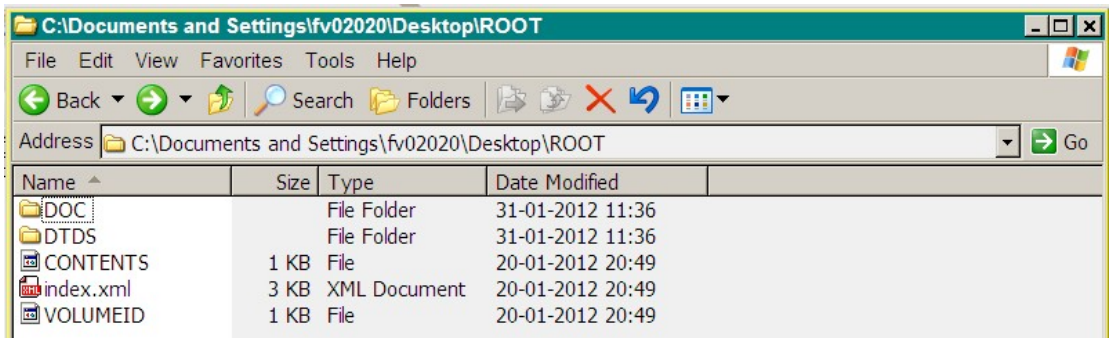
Archives

Il existe un fonds d'archives par pays pour les données valides.
Éventuellement, il existe un fonds d'archives supplémentaire par pays, contenant les données comportant des erreurs. Exemple :

- Ftm_fulltext_CCYYww_CC_nnnn.zip
- Ftm_fulltext_CCYYww_CC_nnnn_errors.zip

Organisation des répertoires

La structure du répertoire est similaire à celle de la base de données DOCDB/XML.

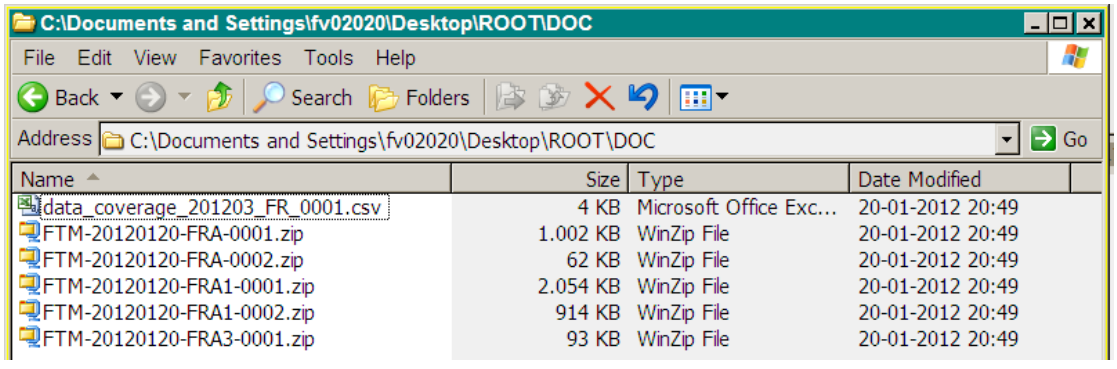


Répertoire DOC

On y trouve un ou plusieurs fichiers d'une taille maximale donnée, zippés.

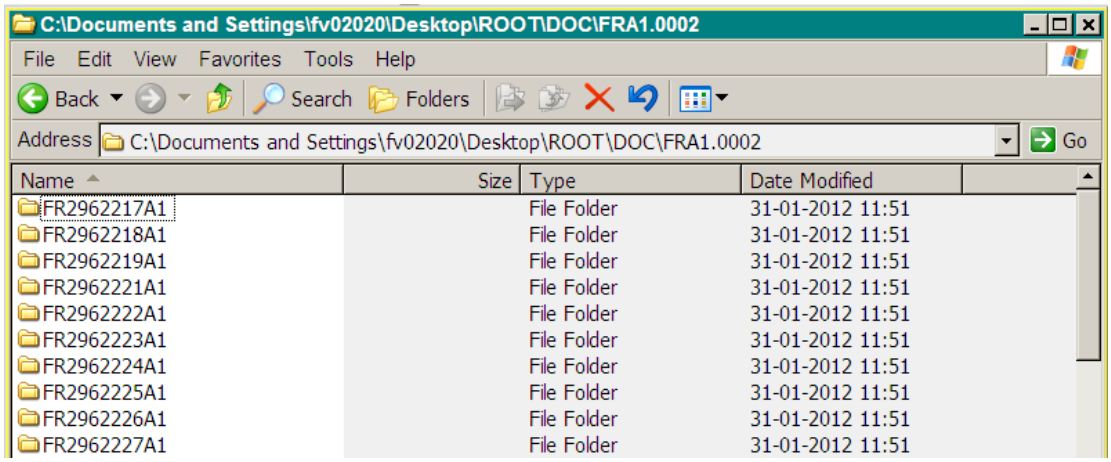
Chaque fichier est identifié par un code pays, un code type (kind code) et un numéro de séquence.

Outre les fichiers, le répertoire DOC comporte également un rapport sur les statistiques et le périmètre couvert.



Dossiers

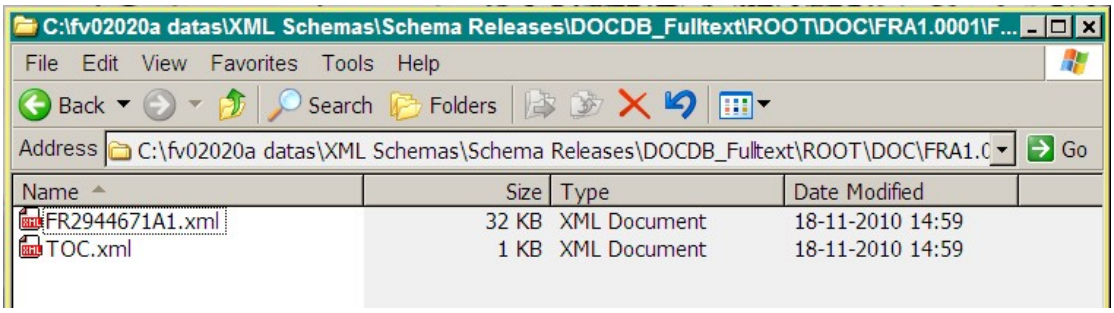
Il existe un dossier par document en texte intégral.

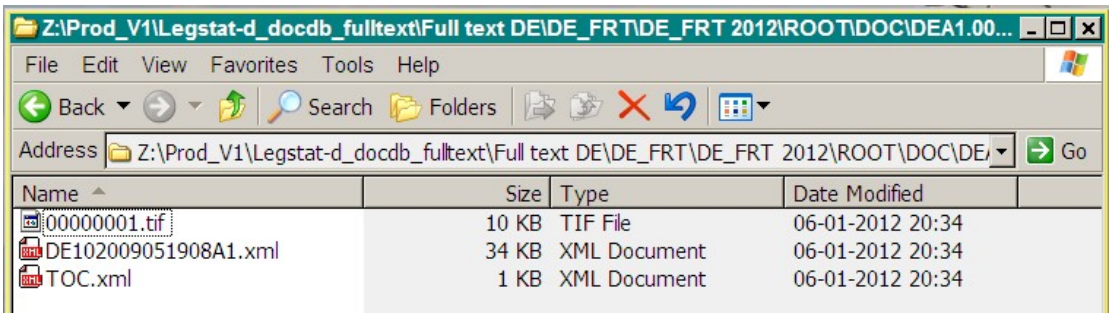


Contenu d'un dossier

Chaque fichier comprend :

- le document en XML
- le cas échéant, les images associées, référencées dans le document XML
- une table des matières





Le document XML

Un document XML peut contenir plusieurs occurrences d'une même partie du texte intégral, chacune étant identifiée par l'indication de langue :

<publication-reference>
<document-id>
<country>EP</country>
<number>2000000</number>
<kind>A1<kind>
<date> … </date>
<document-id>
</publication-reference>
<description lang="de"> ... </description>
<description lang="en"> ... </description>
<description lang="fr"> ... </description>

4. ANNEXE : HISTORIQUE DES VERSIONS

Version 1.2 datée du 5 mars 2013.

www.inpi.fr

INPI

INPI Direct
0820 210 211
(0,09 € TTC/min)

00 33 171 087 163
(depuis l'étranger)

www.inpi.fr

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE