Contemporary Fre n patents

Ocerized full text

Information format and structure File description

Creation date: 05/03/2013 - Version 1.2 Author:

Fenny Versloot-Spoelstra

Europäisches Patentamt
European Patent Office
Office européen des brevets

INPI

CONTENTS

1. PERIMETER

The "**Brevets français contemporains - Texte intégral océrisé**" license contains the full text of the description and claims of French patent applications published by INPI **from 1981 onwards**.

This data is in the standardized **XML ST36** format.

▶ Format evolution

The exchange format described in this document **replaces the original format**, which was based on extractions from the EPOQUE full-text databases, structured according to the EPO's proprietary EBCDIC format.

The new exchange format is based on **extractions from the EPO's Full Text Master database** (FTM). The data is in the standardized **XML OMPI/ST36 (WIPO/ST36)** format.

▶ Backlog files / current files

The EPO reference database for full-text documents (FTM) is a relatively new database, fully operational since 2008/2009. It has replaced the full-text EPOQUE databases (EBCDIC format).

When migrating the numerous original databases to the new FTM reference database, two strategies were applied. Where possible, the full-text reference database (FTM) was rebuilt from the original source, reloading the backlog files with the new format. In some cases, the original content of EPOQUE databases had to be directly reconverted to the new format.

As a result, the richness of the XML may vary depending on whether the data comes from backlog or current files. Data from current files will systematically be produced with the richness offered by the new XML schema definition. Older documents from backlog files may, in some cases, be structured in a limited XML format, due to the constraints of the original proprietary EBCDIC format.

2. EXCHANGE FORMAT

A. FEATURES

The full-text exchange format contains :

1. **The following bibliographical data:**
   - 1.1. Publication identifier
   - 1.2. Publication date
   - 1.3. Deposit ID
   - 1.4. Filing date

2. **The different parts of the full text :**
   - 2.1. Description
   - 2.2. Claims
   - 2.3. Abstract

3. **Processing the "status" attribute**

▶ Contents

Extraction is limited to one full-text extraction per publication.

If the same publication has been supplied to the EPO more than once, the export mechanism consists of selecting the data supplied by the source of higher quality.

Extraction is limited to suppliers and sources with whom the EPO has an agreement to distribute information to third parties.

▶ "Status"

The **"status" or "field level change indicator"** attribute is used to indicate which information has changed since the last exchange.

B. BIBLIOGRAPHICAL DATA

The **"status"** attribute on bibliographic data is filled with status=D - <bibliographic-data status="D"> - when :

- a publication has been removed from the DOCDB database
- a publication has become "void" in DOCDB

If, exceptionally, the full text of a given publication has been removed, the "status" attribute in <bibliographic-data> will also be set to "D".

A new encoding of the publication identifier in the DOCDB database is provided by a combination of status=D and =C :

- <bibliographic-data status="D"> for the old identifier
- <bibliographic-data status="C"> for the new identifier

Any modification of other bibliographic data - modification of publication date, deposit identifier or deposit date - will be indicated by status=A.

***NOTE:** changes to bibliographic data will be reported only for publications whose full text is available in the FTM database.*

C. PARTS OF THE FULL TEXT

The "status" attribute in the various parts of the full text (description, claims, abstract) will be filled with :

- status="C" when a part has been added
- status="A" when a part has been replaced

***NOTE:** status="D" is not provided. Deleting a part of the full text implies that it will no longer be included in the extraction.*

▶ Multi-language full text component

When a part of the full text is added or replaced for a given language, **the attribute "status" is entered in this language**, in the corresponding section. For example :

- for the EN language description part of a document that is created :
  <description lang="de"> ... </description>
  <description lang="en" status='C'> ... </description>
  <description lang="en"> ... </description>
- for the DE language description part of a document that has been replaced :
  <description lang="de" status='A'> ... </description>
  <description lang="en"> ... </description>

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

D. FORMATTING AND VALIDATION

The exchange will always be based on a **complete document** including the full set of images - if available - and the various parts making up the full text of the document.

▶ Images

Images are in **TIFF format**.
Each image associated with a full-text document is unique, and duplicates are eliminated. Only images that are fully referenced in the text are taken into account.

▶ Text

The text is in UTF8 format.
Documents undergo an exhaustive XML validation, and any characters not in UTF8 format are detected during this stage.

This validation verifies :

- that the text is well formed
- UTF-8 encoding
- compliance with the fulltext-documents.xsd schema

E. ERROR HANDLING

Documents with irregularities are included. The decision to eliminate the document or to process it as is is left to the user. Documents containing errors are provided to users separately. They are also stored by the EPO, which analyzes them and decides what action to take.

▶ Images

Some images may be correctly referenced in an XML document but not physically present in the FTM database.

▶ Text

IF the text of certain documents is not validated for reasons of form or character conversion, the relevant part of the full text is encapsulated in CDATA.

▶ Full-text elements not included

The following full-text elements are not processed:

- <drawing> : drawing
- <sequence-listing>: list of sequences
- <tables-external-doc>: tables

▶ Abstract <abstract>

The abstract (<abstract>) is included only if it is contained in the full-text data supplied to the EPO. For some countries, the full text supplied does not include an abstract when one is expected. This is a matter of design and not an error.

▶ <application-reference> element - "doc-id" attribute

The "doc-id" attribute has been introduced for future use. It will contain a unique and stable identifier which will enable, in the future, a reliable link to be made between various EPO databases.

3. ORGANIZATION OF FILES

The organizational structure of the delivered files complies with the EPO standard.

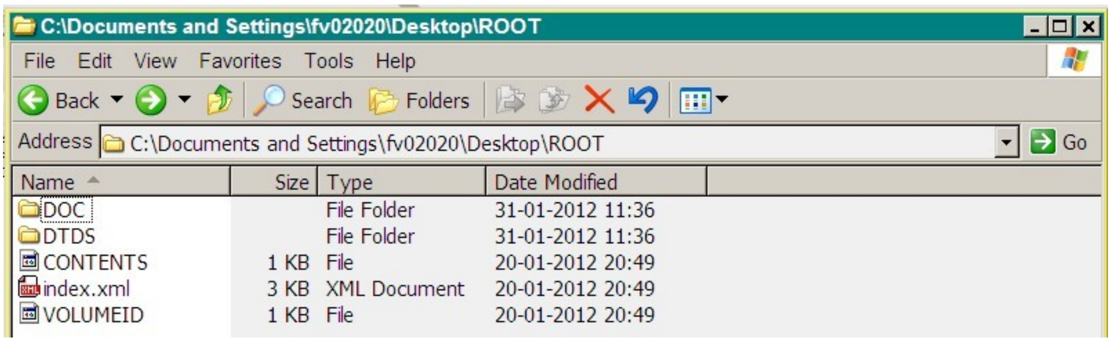
Archives

Each country has its own archive for valid data.
There may also be an additional archive for each country, containing data containing errors. Example:

- Ftm_fulltext_CCYYww_CC_nnnn.zip
- Ftm_fulltext_CCYYww_CC_nnnn_errors.zip

Directory organization

The directory structure is similar to that of the DOCDB/XML database.



C:\Documents and Settings\fv02020\Desktop\ROOT

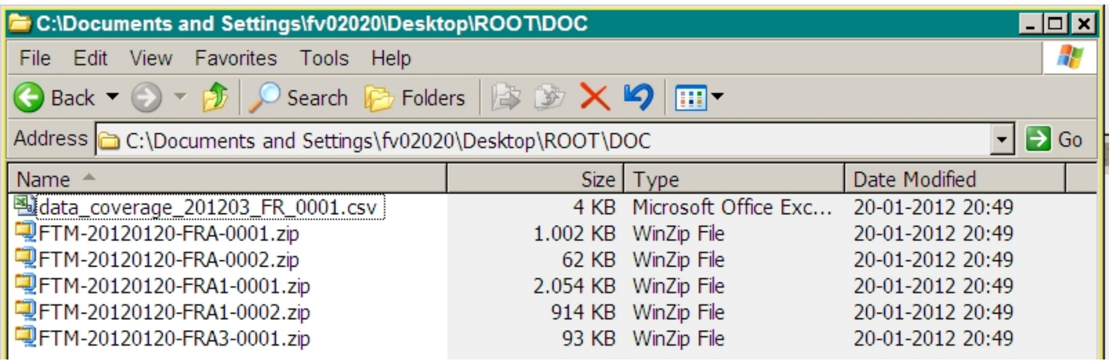
Name	Size	Type	Date Modified
DOC		File Folder	31-01-2012 11:36
DTDS		File Folder	31-01-2012 11:36
CONTENTS	1 KB	File	20-01-2012 20:49
index.xml	3 KB	XML Document	20-01-2012 20:49
VOLUMEID	1 KB	File	20-01-2012 20:49

DOC directory

It contains one or more zipped files of a given maximum size.

Each file is identified by a country code, a kind code and a sequence number.

In addition to the files, the DOC directory also includes a report on the statistics and scope covered.

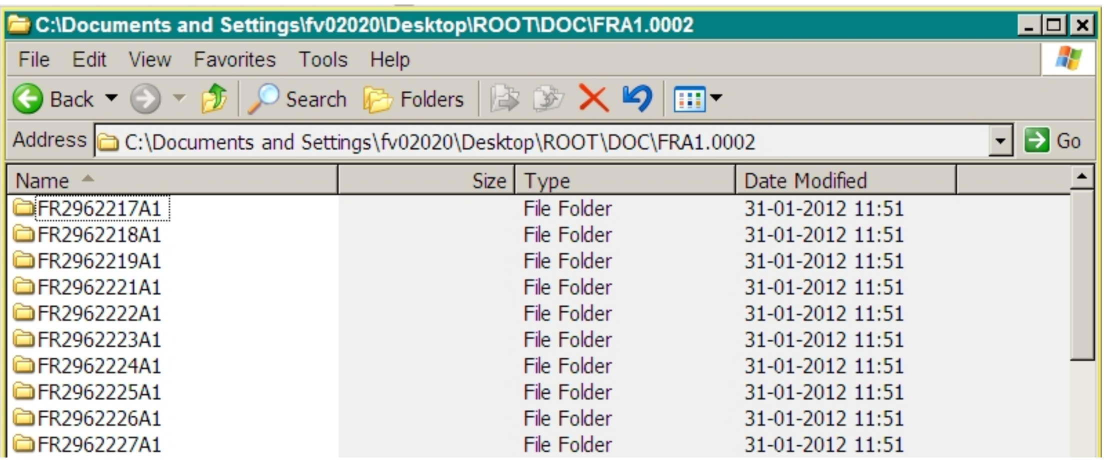


C:\Documents and Settings\fv02020\Desktop\ROOT\DOC

Name	Size	Type	Date Modified
data_coverage_201203_FR_0001.csv	4 KB	Microsoft Office Exc...	20-01-2012 20:49
FTM-20120120-FRA-0001.zip	1.002 KB	WinZip File	20-01-2012 20:49
FTM-20120120-FRA-0002.zip	62 KB	WinZip File	20-01-2012 20:49
FTM-20120120-FRA1-0001.zip	2.054 KB	WinZip File	20-01-2012 20:49
FTM-20120120-FRA1-0002.zip	914 KB	WinZip File	20-01-2012 20:49
FTM-20120120-FRA3-0001.zip	93 KB	WinZip File	20-01-2012 20:49

Files

There is one folder for each full-text document.



C:\Documents and Settings\fv02020\Desktop\ROOT\DOC\FRA1.0002

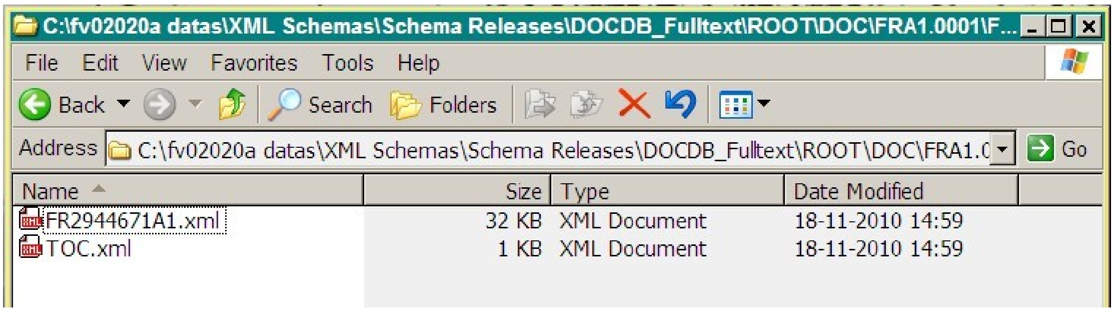
Address C:\Documents and Settings\fv02020\Desktop\ROOT\DOC\FRA1.0002

Name	Size	Type	Date Modified
FR2962217A1		File Folder	31-01-2012 11:51
FR2962218A1		File Folder	31-01-2012 11:51
FR2962219A1		File Folder	31-01-2012 11:51
FR2962221A1		File Folder	31-01-2012 11:51
FR2962222A1		File Folder	31-01-2012 11:51
FR2962223A1		File Folder	31-01-2012 11:51
FR2962224A1		File Folder	31-01-2012 11:51
FR2962225A1		File Folder	31-01-2012 11:51
FR2962226A1		File Folder	31-01-2012 11:51
FR2962227A1		File Folder	31-01-2012 11:51

▶ Contents of a folder

Each file includes :

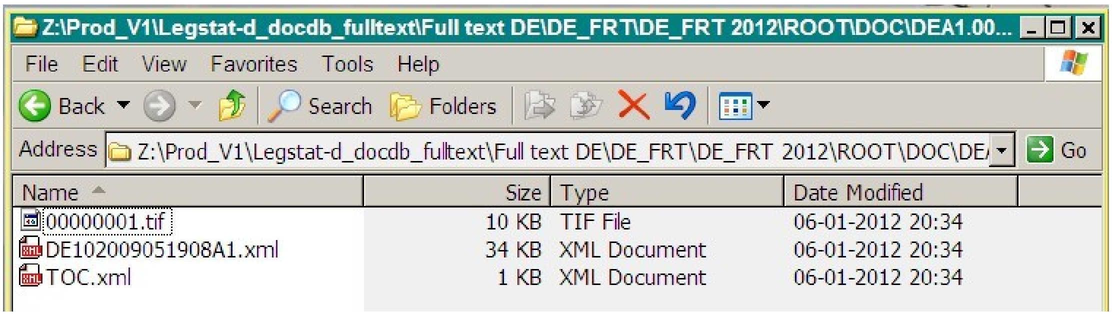
- the document in XML
- any associated images referenced in the XML document
- a table of contents



C:\fv02020a datas\XML Schemas\Schema Releases\DOCDB_Fulltext\ROOT\DOC\FRA1.0001\F...

Address C:\fv02020a datas\XML Schemas\Schema Releases\DOCDB_Fulltext\ROOT\DOC\FRA1.(

Name	Size	Type	Date Modified
FR2944671A1.xml	32 KB	XML Document	18-11-2010 14:59
TOC.xml	1 KB	XML Document	18-11-2010 14:59



Z:\Prod_V1\Legstat-d_docdb_fulltext\Full text DE\DE_FRT\DE_FRT 2012\ROOT\DOC\DEA1.00...

Address Z:\Prod_V1\Legstat-d_docdb_fulltext\Full text DE\DE_FRT\DE_FRT 2012\ROOT\DOC\DE/

Name	Size	Type	Date Modified
00000001.tif	10 KB	TIF File	06-01-2012 20:34
DE102009051908A1.xml	34 KB	XML Document	06-01-2012 20:34
TOC.xml	1 KB	XML Document	06-01-2012 20:34

▶ The XML document

An XML document can contain several occurrences of the same part of the full text, each identified by the language indication :

```
<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="en"> ... </description>
```

4. APPENDIX: HISTORY OF VERSIONS

Version 1.2 dated March 5, 2013.

www.inpi.fr

INPI

INPI Direct
0820 210 211
(€0.09 incl. tax/min.)

00 33 171 087 163
(from abroad)

www.inpi.fr