DOCUMENT: Full Text in ST.36/XML

AUTHOR: Fenny Versloot-Spoelstra

PURPOSE: Description of functional components

DISTRIBUTION: Patent Information/Vienna
External User community

VERSION: 1.2.1

PRODUCT-ID: T03.02

PROJECT: FTM

Document Control

Amendment History

Version	Date	Revisor	Description
1.2	05-03-2013	F. Versloot	Added product description
1.2.1	01-08-2016	F. Versloot	Minor changes, clarification on images

References

No.	
[1]	T04 FTM ST36 Exchange - Technical Design
[2]	T03 ST.36 Package Standard 1.0
[3]	Schema : fulltext-documents-v1.2.2.xsd -

Table of Contents

1 INTRODUCTION

This documents describes the process of generating the Full Text Raw Data Product "Fulltext in ST.36/XML" in functional components :

2. Back-ground
3. Extraction
4. Formatting and Validation
5. Error Handling
6. Packaging

2 BACK-GROUND

2.1 Legacy Product versus XML

"Full-text in ST36/XML" is replacing the legacy product that was based on extractions from the EPOQUE full-text databases. In this legacy product the data exchanged was represented in EPO proprietary EBCDIC.
The new product is based on extractions from the FTM - the EPO's master database for full-text. In this new product the data is represented in XML that is fully compliant with WIPO/ST36.

2.2 Legacy Database versus FTM

The EPOQUE full-text databases have been specifically designed to support the EPO examiners in their search and examination efforts. These efforts do not necessarily require rich tagging inside the full-text. The EPO proprietary format is rich in content but relatively flat in format.
The Full Text Master database has been designed to serve all equally well, both users inside the EPO and users outside. The Full Text Master will maintain the richness of the source by retaining the embedded XML of the original.

2.3 Back-file versus Front-file

The Full Text Master database is a relatively young database. It did not come into full swing until 2008/2009. Until that time the EPOQUE full-text databases were the main repositories for full-text at the EPO:
In the migration from many legacy databases to one master database two strategies have been applied. Where possible the Full Text Master has been populated by going back to the original source and re-loading the back-files in the new format. Only when there was no other option but to start from the legacy databases have the contents of the EPOQUE databases been converted into the new format.
As a consequence the richness of the XML might vary between back-file and front-file data. The front-file data will invariably be represented in the richness that the XML schema definition is offering. The older documents in the back-file may be represented in an XML format limited by the constraints of the proprietary EBCDIC format.

3 EXTRACTION

3.1 Features

Extraction of fulltext will accommodate

1. exchange of a minimum of bibliographic data
    1.1. publication-identification
    1.2. publication-date
    1.3. application-identification
    1.4. application-date
2. exchange of full set of fulltext components
    2.1. description
    2.2. claims
    2.3. abstract
3. population of attribute "status" on component level

3.2 Content

Extraction will be limited to one set of full text data per publication.
In the event where more than one set of full text has been supplied within one exchange period for a given publication, the export mechanism is to select data provided by the best quality source.

Extraction will be limited to those providers and sources for which the EPO have an agreement to disperse the information to third parties.

3.3 "Status"

Attribute "status" or "field level change indicator" serves to indicate which data-components have been modified since the last exchange.

3.3.1. Bibliographic data

Attribute "status" on the bibliographic data component will be populated with status=D - `<bibliographic-data status="D">` - when:

- a publication has been deleted from DOCDB
- a publication has been made "void" in DOCDB

In he rare event that all the fulltext for a given publication will have been removed, attribute "status" on <bibliographic-data> will also be populated with value "D".

A re-key of the publication identifier in DOCDB will be reflected by a combination of status=D and status=C :

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

Any changes to other elements in the bibliographic data - changes to publication-date, application-identifier or application-date - will be reflected by status=A.

NOTE THAT modifications to the bibliographic data will only be signalled for publications that have fulltext available in the Full Text Master

3.3.2. Full Text Component

Attribute "status" on full text component level will be populated with

- status="C" when a full text component has been added
- status="A" when a full text component has been replaced

NOTE THAT status="D" is not being catered for. The delete of a fulltext component is indicated by implication. The deleted component will be missing from the set of fulltext components exchanged

3.3.3. Full Text Component in Multiple Languages

When a full text component in one given language is added or replaced, attribute "status" will be populated on the fulltext component in that given language, e.g:

- EN full text description for a given document has arrived

```
<description lang="de"> ... </description>
<description lang="en" status='C'> ... </description>
<description lang="fr"> ... </description>
```

- DE full text description for a given document has been replaced

```
<description lang="de" status='A'> ... </description>
<description lang="en"> ... </description>
<description lang="fr"> ... </description>
```

4 FORMATTING AND VALIDATION

The exchange will always be based on the full image, ie. a fulltext document included in the exchange will always come with a full set of images - when available - and a full set of components.

4.1 Images

Images – when present – will be in TIFF format.
Each image will be unique to the fulltext document, duplicates will be dropped.
Only images that are fully referenced in the text are taken into consideration.

4.2 Text

Text will be in UTF8.
Any instances of characters not in UTF8 will be detected at XML validation time.

Every text document is systematically run against a full-blown XML validation.
This validation will verify :

- well-formedness
- UTF-8 encoding
- compliance with schema fulltext-documents:xsd

5 ERROR HANDLING

In case of deficiencies the deficient document will be included. The choice either to discard the document or to process it with its deficiences is left to the user.
Documents in error will be provided to the users in a dedicated package.
Documents in error will also be logged in an EPO repository for further analysis and action by the EPO.

5.1 Images

Images may be fully referenced in an XML document, but not physically present in the Full Text Master and therefore not included in the package.

5.2 Text

Text may not have passed validation on well-formedness, character conversion for given entities may not have been successful. The component at fault will be encapsulated in CDATA.

6 DESIGN FEATURES

6.1 Fulltext Components Not Supported

The following fulltext components are not supported :

- <drawing>
- <sequence-listing>
- <tables-external-doc>

6.2 Fulltext Component <abstract>

The fulltext component <abstract> will only be included when the data supplier has provided the EPO with an abstract in its fulltext data feed to the EPO. It may occur for given countries that the fulltext document does not contain an abstract where an abstract would have been expected. This is a design feature rather than an error.

6.3 Element <application-reference> - Attribute "doc-id"

Attribute "doc-id" has been introduced for use in the very near future. It will contain a stable and unique identifier that - in the future - will allow for linking up a number of EPO raw data products through the application in a reliable way.

7 PACKAGING

The packaging strategy opted for is compliant with the EPO Packaging Standard.

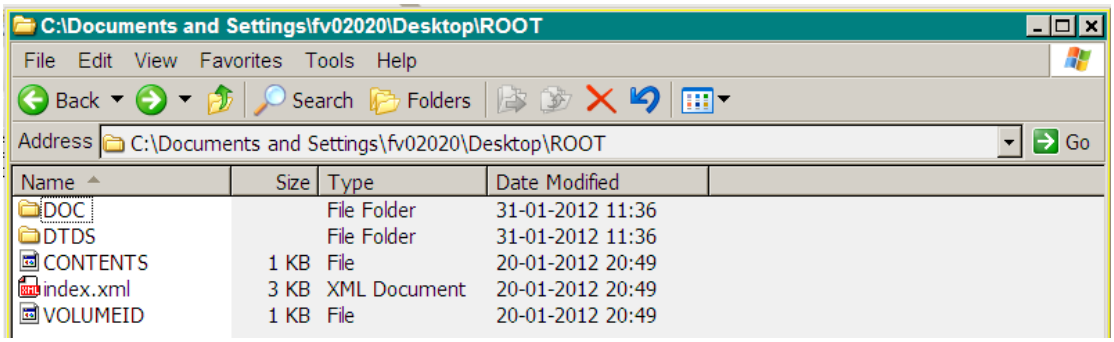
7.1 Archive

One archive per country for valid data.

Potentially an additional archive containing data in error per country for valid data.

Eg. `Ftm_fulltext_CCYYww_CC_nnnn.zip`
`Ftm_fulltext_CCYYww_CC_nnnn_errors.zip`

7.2 Root node

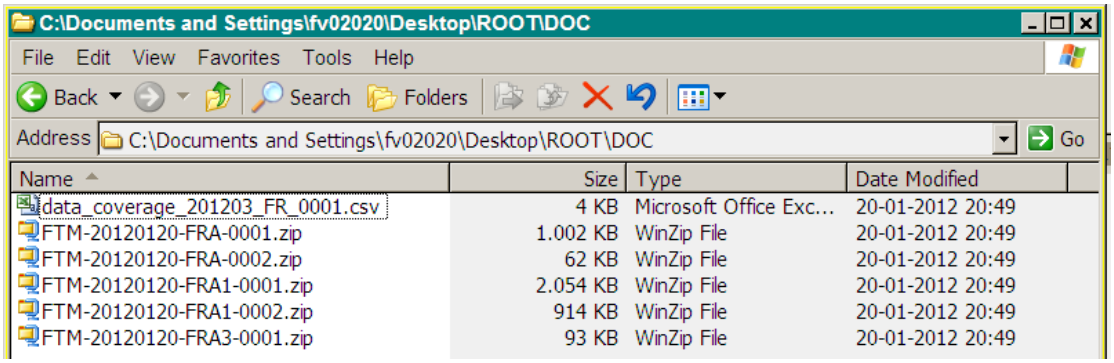


7.3 Directory DOC

One or more packages of a given maximum size, zipped.

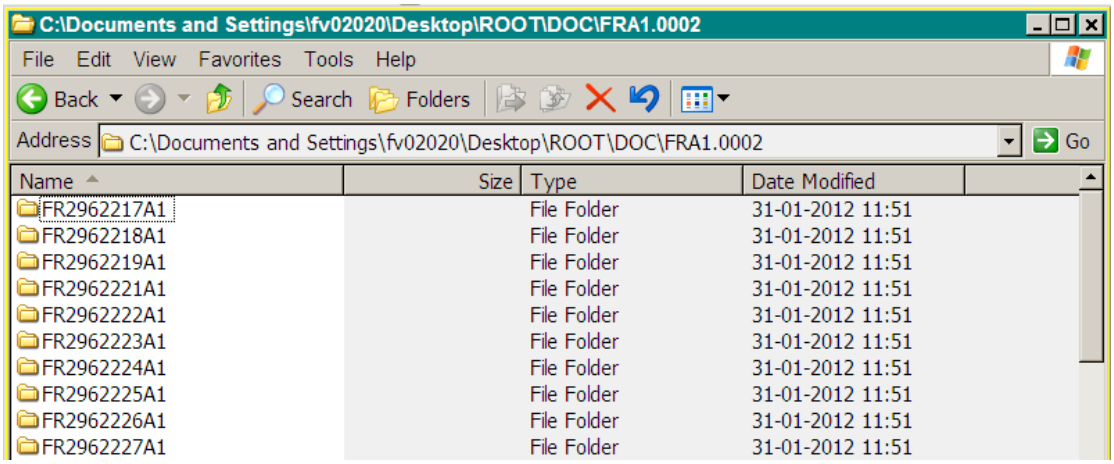
Each package identified by country-code, kind-code and sequence-number.

In addition to the packages, directory DOC will also contain a report on statistics and data coverage.



7.4 Package

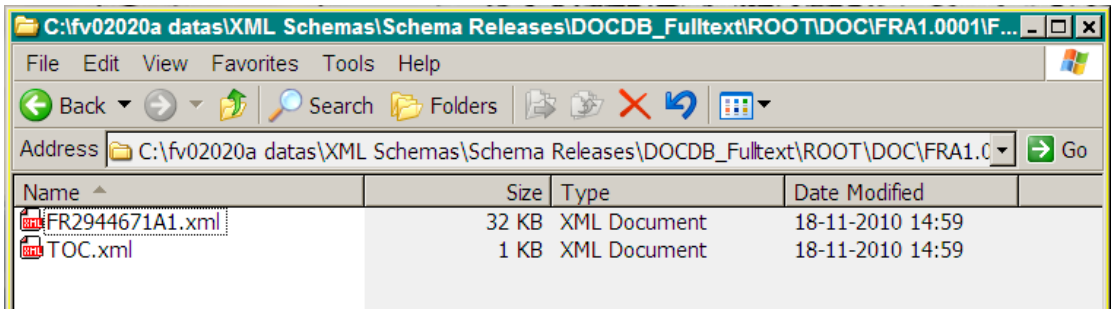
One or more package files, each package-file relating to one fulltext document.

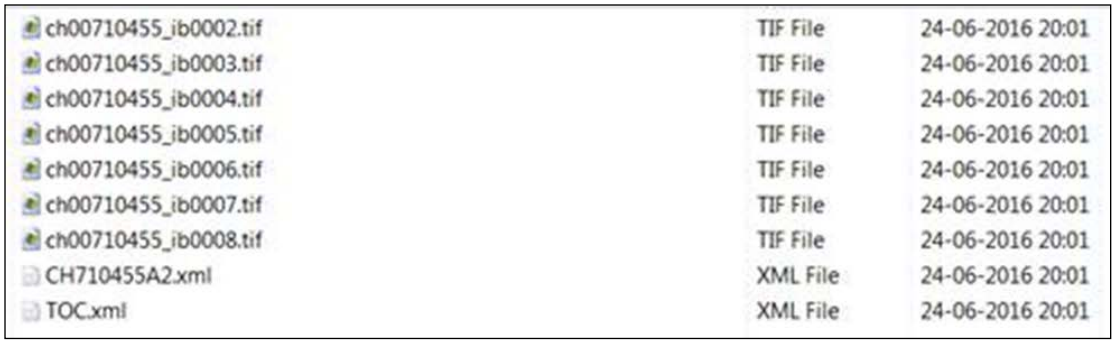


7.5 Package-file

Each package-file containing :

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





7.6 The XML Document

The XML document may contain multiple occurrences of a given component, further identified by language :

```
<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>
```