EP full-text data for text analytics

User guide

Author: EPO – Electronic Publication and Dissemination

Document version 1.0

Save date: 14 March 2019

Revision sheet

Doc. version	Date	Revision description	Scope
1.0	13.03.2019	First version	

Table of contents

1. About this document

This document describes the product EP full-text data for text analytics.

Chapter 2 provides general information about the product, the licence conditions and how to download the data.

Chapters 3 describes the packaging, chapter 4 the content.

Chapter 5 outlines ways of making best use of the data.

Chapter 6 contains information about learning resources relating to patent information and about the support provided by the EPO.

Chapter 7 explains how the data for the product has been created.

2. About EP full-text data for text analytics

2.1. Introduction

The EPO provides users with EP publication data as character-coded full-text in the product EP full-text data (https://www.epo.org/searching-for-patents/data/bulk-data-sets/data.html#tab-1). This product consists of all EP publications in XML, PDF and TIFF format (where available), making it a very comprehensive data set. The only difficulty is that extracting only the publication texts in bulk format is extremely cumbersome.

This is what motivated the EPO to create the derivative product EP full-text data for text analytics. It is tailored to the needs of users who work with text for analysis, machine learning, artificial intelligence, translations and similar purposes.

The design goals have been:
- simple data structure and packaging
- usability for multiple purposes
- complete text data (but without bibliographic data or TIFF and PDF images)
- full XML structure preserved
- easy access
- free
- open licence

EP full-text data for text analytics contains all EP publications, from 1978 until the end of January 2019[1]. The data set, which will be updated annually, comprises approximately 5.8 million EP publications and is about 210 GB in size.

For more product information and user documentation, go to https://www.epo.org/searching-for-patents/data/bulk-data-sets/text-analytics.html .

2.2. Cost and licence

EP full-text data for text analytics is free and can be accessed without prior registration at the EPO.

The EPO grants permission to use EP full-text data for text analytics data under the "Creative Commons Attribution 4.0 International Public license" (further information), allowing you to freely use and share the data.

You must give due credit to the European Patent Office, provide a link to the licence, and indicate whether any changes to the data have been made.

---

[1] DOCDB, INPADOC, EP Register back files (PATSTAT's source data) and EP full-text data are produced at the end of January each year, so EP full-text data for text analytics will be in sync with the PATSTAT Spring Edition releases of the corresponding year.

2.3. Download

The data is stored on the Google Cloud Platform (GCP) in the bucket `epo-patentinformation`. To download it, you must be a registered Google user with a billing project. The data itself is free, but charges apply for downloading the data from GCP, depending on its volume[2]. These charges will be billed by Google to your billing project.

There are several ways to download a file or a complete directory: See https://cloud.google.com/storage/docs/access-public-data .

If you want to download all the files from the subdirectory, you can use Google’s `gsutil` tool as described below:

To list the content of the bucket `epo-patentinformation`, you could use a command like:

```
gsutil –u “your-billing-project-ID” ls –r gs://epo-
patentinformation
```

To download a directory from the bucket to your local storage:

```
gsutil –u “your-billing-project-ID” cp –r gs://epo-
patentinformation/dir-to-download c:\temp
```

To download a file from the bucket to your local storage:

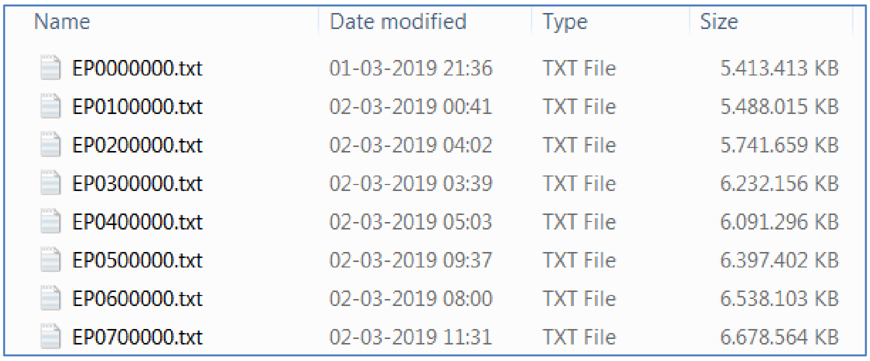
```
gsutil –u “your-billing-project-ID” cp gs://epo-
patentinformation/dir/file-to-download c:\temp
```

---

[2] See the pricing for Google Cloud Platform here: https://cloud.google.com/storage/pricing.
The egress charge for data to worldwide locations (except China and Australia) was USD 12 for 100 GB in March 2019.
Note that the EPO accepts no liability for the completeness and accuracy of this information.

3. Packaging

EP full-text data for text analytics comprises around 35 data files of about 5-8 GB each.

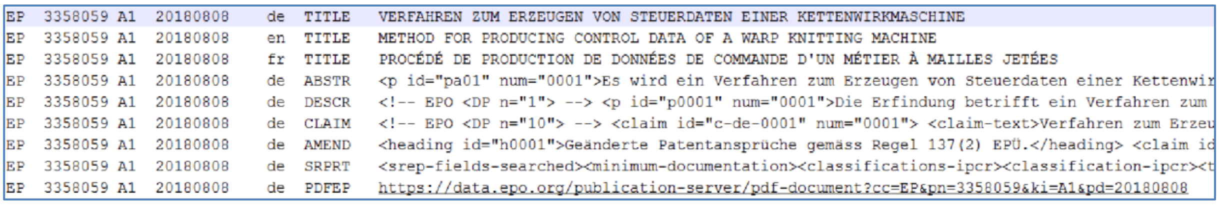


Name	Date modified	Type	Size
EP0000000.txt	01-03-2019 21:36	TXT File	5.413.413 KB
EP0100000.txt	02-03-2019 00:41	TXT File	5.488.015 KB
EP0200000.txt	02-03-2019 04:02	TXT File	5.741.659 KB
EP0300000.txt	02-03-2019 03:39	TXT File	6.232.156 KB
EP0400000.txt	02-03-2019 05:03	TXT File	6.091.296 KB
EP0500000.txt	02-03-2019 09:37	TXT File	6.397.402 KB
EP0600000.txt	02-03-2019 08:00	TXT File	6.538.103 KB
EP0700000.txt	02-03-2019 11:31	TXT File	6.678.564 KB

Figure 1: Data files of EP full-text data for text analytics

- the name of each file follows the format "EP`nn`00000.txt", for example "`EP2300000.txt`".
- each file contains the publications associated with 100 000 publication numbers. For example, file `EP2300000.txt` contains all publications in respect of publication numbers 2300000-2399999. There may be (and often are) more than one publication with a given publication number.

4. Content



```
EP  3358059 A1  20180808   de  TITLE   VERFAHREN ZUM ERZEUGEN VON STEUERDATEN EINER KETTENWIRKMASCHINE
EP  3358059 A1  20180808   en  TITLE   METHOD FOR PRODUCING CONTROL DATA OF A WARP KNITTING MACHINE
EP  3358059 A1  20180808   fr  TITLE   PROCÉDÉ DE PRODUCTION DE DONNÉES DE COMMANDE D'UN MÉTIER À MAILLES JETÉES
EP  3358059 A1  20180808   de  ABSTR   <p id="pa01" num="0001">Es wird ein Verfahren zum Erzeugen von Steuerdaten einer Kettenwir
EP  3358059 A1  20180808   de  DESCR   <!-- EPO <DP n="1"> --> <p id="p0001" num="0001">Die Erfindung betrifft ein Verfahren zum
EP  3358059 A1  20180808   de  CLAIM   <!-- EPO <DP n="10"> --> <claim id="c-de-0001" num="0001"> <claim-text>Verfahren zum Erzeu
EP  3358059 A1  20180808   de  AMEND   <heading id="h0001">Geänderte Patentansprüche gemäss Regel 137(2) EPÜ.</heading> <claim id
EP  3358059 A1  20180808   de  SRPRT   <srep-fields-searched><minimum-documentation><classifications-ipcr><classification-ipcr><t
EP  3358059 A1  20180808   de  PDFEP   https://data.epo.org/publication-server/pdf-document?cc=EP&pn=3358059&ki=A1&pd=20180808
```

Figure 2: Example of a file fragment

The data format is as follows:

- every line of the text file represents a text portion of the publication.
- all parts of a single publication are in consecutive lines. Each line ends with a CR/LF.
- the files are tab-separated value files containing key-value pairs.
    - the key consists of:
      - the publication authority (will always have the value "EP")
      - the publication number (a seven-digit number)
      - the publication kind[3]
      - the publication date
      - the language of the text component (de, en, fr; xx means unknown)
      - the text type
    - the text itself depends on the text type. It contains, where appropriate, XML tags for better structure. You will find the DTD applicable to all parts of the publication at: http://docs.epoline.org/ebd/doc/ep-patent-document-v1-5.dtd All tabs in the text have been replaced by a space character because the tab is used as a field delimiter.
- some or all of these text types occur:

TITLE	a title
ABSTR	the abstract
DESCR	the description of the invention
CLAIM	a set of claims
AMEND	a set of amended claims
ACSTM	a statement of the amended claims
SREPT	the search report
PDFEP	the URL of the publication as a PDF document in the European Publication Server

The table below shows you which text types occur in which kinds of publication:

---

[3] See https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/definitions.html for help.

	Text type							
Publn. kind	**TITLE**	**ABSTR** (abstract)	**DESCR** (description)	**CLAIM** (set of claims)	**AMEND** (set of amended claims)	**ACSTM** (amended claims statement**)**	**SRPRT**$ (search report)	**PDFEP** (URL of the PDF document of the EP)
A1	3	1*	1*	1*	✓*	✓*	1*$	1*
A2	3	1*	1*	1*	-	-	-	1*
A3	3	1	-	-	-	-	1$	1
A4	not included							
A8	3	1	-	-	-	-	-	1
A9	3	1	1	✓Δ	✓Δ	✓Δ	✓Δ	1
B1	3	-	1	3+	-	-	-	1
B2	3	-	1	3+	-	-	-	1
B3	3	-	1	3+	-	-	-	1
B8	3	-	-	-	-	-	-	1
B9	3	-	1	3+	-	-	-	1

Explanation (numbers and symbols):

1	occurs exactly once
3	occurs exactly three times (i.e. in all three official languages)
3+	occurs three times (or as multiples of three due to country-specific claims)
✓	may occur
-	does not occur
*	not for Euro-PCTs which are not re-published
$	search reports occur only for publications published in 2012 or later
Δ	only if already in the corrected publication

These general rules apply:

- lines belonging to the same publication are kept together.
- the publications are not sorted. For example, the lines of the B1 publication will not immediately follow the lines of the A1 application but they will be in the same file because they have the same publication number.
- A4 publications (supplementary search reports) are not included.
- no personal information (e.g. inventor names; the examiner name in a search report) is included.

- the URL in the PDFEP component refers to the European Publication Server. Its fair use policy applies.

  Note: In addition to PDF, you can retrieve the publication in XML and HTML format from the publication server. For guidance, see the "European Publication Server REST services reference guide" at https://www.epo.org/searching-for-patents/data/web-services/publication-server.html.

For details of how text has been extracted from XML publications, see section 7 Extraction from XML publications – what, where and how.

5. Use cases

EP full-text data for text analytics has been designed to make the text of EP publications easily accessible and suitable for multiple purposes. To filter the publications by patent office, time period, technical classification, applicant, etc., you first have to identify the relevant publication numbers. You can do so by using one of the numerous EPO data sets and tools (available for free or by subscription) or others available from commercial providers.

Some relevant EPO products and services are:

With European scope:
- European Publication Server and its REST interface
- Open Patent Services (OPS Register service; a web service and OPS Published services for EP full text)
- EP full-text search
- EP linked open data
- PATSTAT EP Register data

With worldwide scope:
- Espacenet
- Global Patent Index (GPI)
- Open Patent Services (OPS worldwide services; a web service)
- PATSTAT Global data

For more information on bulk data products, go to the table "Main features of EPO bulk data subscription products and services" accessible from near the bottom of page https://www.epo.org/searching-for-patents/data/bulk-data-sets.html.

5.1. Combining it with PATSTAT data

There are at least two ways of combining PATSTAT and EP full-text data for text analytics.

- **Using PATSTAT to identify relevant publications**
  PATSTAT can be used to create a list of publication numbers which fulfil some user-defined criteria by writing an appropriate SQL query statement and by downloading the list of publication numbers via the feature "Download result table" or "Download PATSTAT subset".
  As a second step, any appropriate tool can be used to extract the text of the identified publications from EP full-text data for text analytics and process it as required.

- **Loading text data into the PATSTAT database**
  EP full-text data for text analytics actually consists of a set of tab-separated value files, which can be loaded into a conventional database as a single table. This table will consist of some identifying fields (i.e. the *key* attributes) and a long field

holding the text. If all publications are loaded into the database, then this table will be huge. Depending on the way the text is then queried, it might be useful to apply a full-text index to the text field.

Once all the data is available in table format, processing it using the SQL query language becomes straightforward.

6. Introduction to patent information and EPO help desk

6.1. Introduction to patent information

This free e-learning tool helps you to understand patent information and its uses: https://e-courses.epo.org/pluginfile.php/32401/mod_resource/content/1/pi-tour/index.html

6.2. Glossary of patent-related terms

If you are not familiar with the terminology of the patent domain, you will find a glossary of EPO, patent and IP-related terms and abbreviations in www.epo.org/service-support/glossary.html.

6.3. Contact for help, feedback and reporting of errors

A help desk is provided at patentdata@epo.org

7. Extraction from XML publications – what, where and how

This section explains what content appears in each text type and shows you the XML element in the XML publication from which the data is extracted.

You will find the DTD of the publication XML at:
http://docs.epoline.org/ebd/doc/ep-patent-document-v1-5.dtd

7.1. Text types

Text type TITLE: title

- includes everything within the `<B542>` element
- the language code of the title is taken from the `<B541>` element

Text type ABSTR: abstract

- includes everything within the `<abstract>` element

Text type DESCR: description

- includes everything within the `<description>` element

Text type CLAIM: claim set

- includes everything within the `<claims>` element, which in turn contains multiple `<claim>` elements

Text type AMEND: set of amended claims

- includes everything within the `<amended-claims>` element, which in turn contains multiple `<claim>` elements

Text type ACSTM: amended claims statement

- includes everything within the `<amended-claims-statement>` element

Text type SREPT: search report

- includes everything within the `<srep-for-pub>` element
- the language code of the search report is taken from the `@lang` attribute of the `<search-report-data>` element

Text type PDFEP: link to the PDF publication

- the URL is constructed as follows: https://data.epo.org/publication-server/pdf-document?cc=EP&pn`=<publn-number>`&ki`=<publn-kind>`&pd=`<optional publn-date yyyymmdd>`
- "Non-republished Euro-PCT publications" do not have a PDF representation. They are A1 or A2 publications containing a single character which is an asterisk in the B003EP tag. So the tag looks exactly like this: `<B003EP>*</B003EP>.` Example: publication EP 3102070 A1.

7.2. Page breaks

Page breaks are included in the XML full text as XML comments such as `<!– EPO <DP n="9"> -->` , which indicates the start of document page 9. These page numbers refer to the original document (e.g. the application as filed) and not to the newly laid out PDF document, which is retrievable from the European Publication Server, e.g. via the link in the text type PDFEP.