

How to load PATSTAT data into your database

PATSTAT Team / EPO Vienna
Version 1.18

Contents

1.	About this document.....	1
2.	Understanding the PATSTAT delivery.....	2
2.1.	Data files	2
2.2.	Documentation and related material	2
3.	Creating the database structure	3
4.	Loading files into the database	3
5.	Testing the data import.....	4
6.	Version history	4

1. About this document

The latest version of this document can be found at <http://www.epo.org/patstat> .

This document explains how to load PATSTAT's data files into your database. You won't need this information if you subscribe to PATSTAT Online because it is a ready-to-use online application for patent analytics.

Data loading, which is an essential part of the process, can vary according to the individual DBMS¹ and is therefore not detailed in this document.

This document is applicable to all data products in the PATSTAT suite, these being

- PATSTAT Global and
- PATSTAT EP Register

The data loading process is split into several stages, and these are described in the following:

- Section 2: Understanding PATSTAT delivery
- Section 3: Creating the database structure
- Section 4: Loading files into the database
- Section 5: Testing the data import

¹ Database Management System

2. Understanding the PATSTAT delivery

This section describes the files and folders which you have downloaded

- manually with your browser from EPO's download platform (<https://publication-bdds.apps.epo.org/raw-data/products>), or
- programmatically via the REST API (instructions provided in the download platform product folder).

It is recommended that you make use of the fingerprint value associated with each file to check that the files have not been corrupted during download or copy.

You first need to unzip all downloaded files.

The delivery consists of two parts:

- data files
- documentation and related material

2.1. Data files

You will find the data files in the folder `Data`. For every table there is one or more zipped data file(s). To keep all files (if unzipped) smaller than about 7 GB, larger tables are split into multiple files. The unzipped data files have the file extension `.csv`.

Each `csv` file contains one header record containing the column names of the table, followed by multiple data records.

- All characters are in Unicode with UTF-8 encoding
- The files are in MS-DOS format (i.e., each line ends with CR/LF)
- Values are delimited by a comma ","
- Values of text attributes (which are of a string type) are enclosed in double quotes, like: "Smith, John". Within string values there never are double quotes.
- Values of non-text attributes (which are of a number or date type) are *not* enclosed in quotes, like: 123, 2019-12-31
- The decimal separator is the point ".", e.g., 0.125
- Line breaking characters (LF, CR) within a text attribute are replaced by " \n ". They occur most frequently in the abstract text and the NPL bibliographic text.
- Except "\n", no other characters are escaped.

The size of the data sets (from the 2024 Autumn Edition) is roughly as follows:

Data set	Size of zipped files	Size of unzipped files
PATSTAT Global	69,6 GB	410 GB
PATSTAT EP Register	5 GB	41 GB

2.2. Documentation and related material

In the folder `Documentation_Scripts` you will find various documents to get started and to assist you in using PATSTAT effectively. Please familiarise yourself with them.

3. Creating the database structure

The sample database structure provided by the EPO is documented in detail in

- the Data Catalog for PATSTAT Global and
- the Data Catalog for PATSTAT EP Register.

The Data Catalogs are part of the delivered documentation and are also available on the EPO website.

You can use the proposed database schema as it is or adapt it to your needs. The EPO also provides SQL scripts to generate the database, the tables, and some basic database indexes. You will find these SQL scripts in the subfolder `CreateScripts` in the documentation folder. Internally, the EPO uses the MS SQL Server 2018, which is why the scripts are in T-SQL (the Microsoft SQL dialect). If you work with another DBMS, you may need to adapt the syntax.

To create the database structure, you need to:

- 1) Run the database creation script
- 2) Run the table creation scripts

This creates empty tables which will be populated with data in the next step.

4. Loading files into the database

To store data, the database needs a certain amount of disk space (the figures are taken from an MS SQL Server installation for PATSTAT 2024 Autumn Edition and may differ to your DBMS):

Data set	Database size app.
PATSTAT Global	537 GB
PATSTAT EP Register	40 GB

Please keep in mind that you need additional disk space for log files (at least during the data loading phase as log files can become very large), extra database indexes, and any supplementary tables that you may want to create.

The method you use to import PATSTAT data files into your database tables depends on your DBMS. Each DBMS provides different tools or commands to perform this task, so we cannot provide any more details here.

When configuring the import tool or creating data loading scripts, you need to specify the format of the source and the target data. To do this, you need the information given in Section 2.1 “Data files” of this document, the Data Catalog, and the table creation scripts (see above).

Here are some tips for loading data:

- The abstracts in table `TLS203_APPLN_ABSTR` usually take a long time to load and require quite a lot of disk space so there is no need to load them if you do not intend to use them.

You can import them at a later stage, if needed.

- The data structures may change slightly from one PATSTAT edition to the next. Instead of creating your data loading scripts each time when there are changes, it is a good practice to adapt them as required. Database schema changes are documented in the Data Catalogs, in the section "History of major changes to tables and attributes".
- When using MySQL: You must select character set `utf8mb4` because MySQL's character set `utf8` is restricted to 3 bytes (cf. <https://dev.mysql.com/doc/refman/5.5/en/charset-unicode-utf8mb4.html>)

The next step is to:

3) Use an appropriate tool to load PATSTAT data files into the DBMS of your choice

Now you can create indexes on the attributes that are used most often for filtering or for joining tables. Primary key indexes have been created during table creation. Additional indexes are specified in the index creation scripts in the sub-folder `CreateScripts` in the documentation folder. You may use these scripts or change or extend them to suit your needs. To speed up data loading, the general recommendation is to create the indexes *after* the data has been loaded, so the next step is:

4) Create indexes

5. Testing the data import

For most errors during the data loading process, you will get a warning from your data loading tool. However, some errors, e.g., forgetting to load a specific data file, will not be identified during data loading. Therefore, as a final test, you should verify that all tables are complete.

To support this process, the PATSTAT documentation folder contains test scripts that count the number of rows for each table, and a list showing the expected outcome. You will find these scripts in the sub-folder `TestScripts` in the documentation folder.

Final step:

5) Run the script to count the rows of each table and compare with the specification

6. Version history

Version 1.00: 2015-10-06: First published version

Version 1.01: 2015-11-16: Note on UTF8 character set in MySQL (section 4)

Version 1.02: 2015-12-09: Note on format of decimal numbers (section 2.1)

Version 1.03: 2016-10-19: CR/LF are replaced by "`\n`"

Updated list of user-provided loading scripts

New product names used

Adaption to delivery via memory sticks and download via EPO's data platform

Version 1.04: 2016-10-28 Note added about WinZip
Version 1.05: 2017-04-13 Title changed
Version 1.06: 2017-05-05 Editorial changes
Version 1.07: 2017-10-25 Update of links to EPO home page
Version 1.08: 2018-04-01 Link to SQLite loading script added; Adapt product names
Version 1.09: 2018-09-28 Link to loading scripts updated
Version 1.10: 2019-01-30 Link to loading scripts updated
Version 1.11: 2019-10-18 Link to loading scripts updated
Version 1.11: 2019-11-06 Link to loading scripts updated
Version 1.12: 2020-03-16 Data files now have the extension `.csv`
Version 1.13: 2020-06-12 Link to loading script for MS SQL server has been added
Version 1.14: 2020-09-16 Minor corrections
Version 1.15: 2022-11-15 Data size figures updated, new download platform
Version 1.16: 2023-10-20 Data size figures updated, physical storage media removed
Version 1.17: 2024-05-13 Data size figures updated, forum links removed
Version 1.17: 2024-10-31 Data size figures updated, editorial changes