## Revision sheet

<table>
<thead>
<tr>
<th>Doc. version</th>
<th>Date</th>
<th>Revision description</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>18.12.2017</td>
<td>First version</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>16.05.2018</td>
<td>Editorial update and clarifications</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>28.06.2018</td>
<td>Updated example of section 2</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>20.03.2019</td>
<td>Link corrected in section 2.2</td>
<td></td>
</tr>
</tbody>
</table>
# Table of contents

1. **Introduction** ................................................................. 4  
   1.1. About this document 4  
   1.2. About linked data 4  
   1.3. About *Linked open EP data* 5  
   1.4. Fair use 5  
   1.5. Accessing *Linked open EP data* 6  
   1.6. Contact for help, feedback and reporting of errors 6  
   1.7. Glossary of patent-related terms 6  

2. **Getting started** ............................................................ 7  
   2.1. Retrieving information about an application 7  
   2.2. Retrieving information about multiple publications 9  
   2.3. Filtering and sorting; customised views and return formats 11  
   2.4. Querying with SPARQL 12  

3. **Linked open EP data web pages** .................................... 14  
   3.1. About 14  
   3.2. API (application programming interface) 14  
   3.3. SPARQL query 16  
   3.4. Documentation 17  
   3.5. Download 17
1. Introduction

1.1. About this document

This manual provides information on the use of the EPO's patent information product **Linked open EP data**.

Chapter 1 introduces the concept of linked data and provides general information about the product **Linked open EP data**.

Chapter 2 sets out easy-to-follow steps to get you started.

Chapter 3 covers the functionalities of the web pages associated with the product.

1.2. About linked data

Linked data is data which has a very simple structure and which is made available on the Internet. The data is structured in threes or triples, like short sentences containing a subject, a predicate and an object. These simply structured facts can express information like

- Patent 12345 has IPC code A61K7/00.
- Person abcd has name John Smith.

Individual data items can be related to each other, even if they have been created by a separate organisation or are located on different servers, e.g. a data item can combine two of the above facts with a third one:

- Patent 12345 has applicant Person abcd.

Linked data is based on the technologies of the World Wide Web. The web as we know it, however, is a web of documents. Each document is actually a web page containing text, images, videos, etc., which are intended to be read and understood by humans. By contrast, linked data creates a web of data in the form of the above-mentioned triples, and this data is intended to be read and processed by machines.

Linked data allows the web to be seen as a giant global database containing a vast number of individual facts that can be combined in any number of ways.

Linked data is based on three simple principles which Tim Berners-Lee, who coined the term "linked data", presented in 2009:

1. **All kinds of conceptual things, they have names now that start with "HTTP".**

   Note: In our data set, these conceptual things are applications, families, inventors, CPC symbols, etc.
2. If I take one of these HTTP names and I look it up ... I will get back some data in a standard format which is kind of useful data that somebody might like to know about that thing, about that event.

Note: All you need is a web browser: we will do this in sections 2.1-2.3.

3. When I get back that information, it's not just got somebody's height and weight and when they were born; it's got relationships. And when it has relationships, whenever it expresses a relationship, then the other thing that it's related to is given one of those names that starts with HTTP.

Note: All you need is a web browser: we will do this in chapter 2.

1.3. About Linked open EP data

Linked Open EP data is a data product provided by the EPO. It contains EP publications with their bibliographic and family information. It also contains some basic information on non-EP patents, which are related to EP patents, e. g. because they are a priority of an EP document or they are in the same family as an EP application.

The CPC hierarchy is also provided. All this data is interlinked. Links are also provided to other patent data sets, if they are published as linked open data by their respective authority. Currently this is only the case for South Korea.

The data is updated weekly. It is covered by the open standard licence CC BY 4.0.

The main purpose of Linked open EP data is to make this data available for bulk download to be used again in other products and applications.

The product also comes with a simple application programming interface (API), allowing you to consult reference data, explore that data and try out ideas on a small scale. A SPARQL interface enables you to analyse the data.

1.4. Fair use

The data browser, linked data API and the SPARQL endpoint are provided for occasional use. They allow you to consult reference data, explore that data and try out ideas on a small scale. You can use them to look up patent data with a known URI, to see what data is available or to try out a simple API call or SPARQL query. They are not designed to provide a backend database service for production applications.

Note: Please see http://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-2 for the conditions of use.
1.5. Accessing Linked open EP data

To access Linked open EP data go to

- the EPO home page www.epo.org > Searching for patents > Data > Linked open EP data. or
- directly to http://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1

and click Open. No login or registration is required.

![Figure 1: Product page](image)

1.6. Contact for help, feedback and reporting of errors

A dedicated discussion forum is available (www.epo.org/forums) where you can ask and respond to questions or share experiences. Product news will also be announced via this forum.

A help desk is provided at patentinformation@epo.org.

1.7. Glossary of patent-related terms

Users of Linked open EP data might not be 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.
2. Getting started

This section contains a brief introduction to what you can do with Linked open EP data.

2.1. Retrieving information about an application

Each application has a unique identifier which looks like a URL and has this structure:

https://data.epo.org/linked-data/doc/application/cc/nnnnnnnn

where

cc stands for the Office code.
To begin with, use "EP" because our data set contains all EP applications, but only some applications from other offices

nnnnnnnn stands for the application number. For EP, these are always eight digits, with leading zeros if required.

Example:


Note: Section 3.4 describes where you will find information on how to retrieve data about other business objects like publications, families, inventors, CPC classification, etc.

Unlike the web as you know it, this URL (which is actually called a URI) refers to data and not to a web page. So entering the URI given in our example above in the address line of your web browser returns data about the application in question. This data is structured so that machines can process it easily, but it may not necessarily be easy for humans to process it. So to make it easier to read, our data is nicely rendered in your browser.
You will see a table with two columns. The second column probably includes other nested two-column tables. The first column is a property of the item; the second column is the value of this property.

In our example, there is a property called "application number" and it has the value 00122581. Likewise, the property "filing date" has the value Oct 17, 2000.

All underlined text is a link and is therefore clickable:
- Clicking a property displays a short explanatory "comment" about this property.
- If the value in the right-hand column is underlined, it is another URI identifying another linked data resource. If you click the link, the browser will display the property / value pairs of this resource.
For example, you can click a patent publication:

![Figure 3: A publication of that application](image)

This will display information about this publication:

![Figure 4: Information about publication](image)

You can click other links or go back to the previous screen via the browser's back button.

### 2.2. Retrieving information about multiple publications

An example of an identifier for a publication is:


Entering this publication identifier in the address line of your Internet browser will retrieve data about the EPO's A1 publication number 1382399. The hyphen at the end of this identifier indicates that no publication date needs to be given, because there may only be a single A1 publication for any EP publication. In contrast, some kinds of application, like the corrections A8, A9, B8, and B9, may occur multiple times, so instead of a hyphen the publication date has to be added to the URI.

Removing a part of the tail of the URI indicated above will retrieve all EP publications that have the number 1382399:
Figure 5: All four EP publications that have the number 1382399: A1, B1, B8, B2 publications.

The B8 publication is shown with this identifier:


Because there may be multiple B8 publications (i.e. multiple corrections) with the same number, the publication date 2007-10-10 must be part of the identifier.
You could go one step further and again shorten the URI identifier by removing the number. You will get this identifier:


It will retrieve a list of all EP applications. Because there are several million EP publications, only the (random) first few will be displayed.

### 2.3. Filtering and sorting; customised views and return formats

The following example will give you an idea of what you can do by adding parameters to the URI. Cut and paste the following URI into the address line of your browser (it should be a single line; remove all line breaks if required) and wait a couple of seconds:

https://data.epo.org/linked-data/data/publication/EP.html?_pageSize=50&_page=0&_view=none&min-publicationDate=2017-07-01&_properties=titleOfInvention,publis

The browser will display the requested data. For better readability, each part of this string is explained separately.

<table>
<thead>
<tr>
<th>URI</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>.html</td>
<td>Return data in HTML format</td>
</tr>
<tr>
<td>pageSize=50</td>
<td>Retrieve 50 entries</td>
</tr>
<tr>
<td>page=0</td>
<td>Retrieve the first (50) entries</td>
</tr>
<tr>
<td>view=none</td>
<td>Apply no view; for views see also section 3.3</td>
</tr>
<tr>
<td>min-publicationDate=2017-07-01</td>
<td>Apply a filter whereby the publication date must be 2017-07-01 or later</td>
</tr>
<tr>
<td>properties=titleOfInvention,publicationDate</td>
<td>Display the title and date of the publication</td>
</tr>
<tr>
<td>sort=publicationNumber</td>
<td>Sort by publication number</td>
</tr>
</tbody>
</table>

Depending on the requested format (e.g. .html, .xml, .json, .ttl), the data is either displayed directly in your browser or made available by the browser for download.

There is more information about these parameters in the document "API overview" (see section 3.4). The API conforms to the [Linked Data API Specification](https://www.w3.org/TR/ld-api/).
2.4. Querying with SPARQL

Open the web pages of *Linked open EP data* (see section 1.5 "Accessing *Linked open EP data*") and click **SPARQL query** in the black menu bar. The SPARQL query form will open.

You can select sample queries from the drop-down list in the top left corner. The query will be pasted in the form below the section with the lines starting with **prefix**, which you can ignore for the moment.

After clicking the **Perform query** button, the query result will be displayed below this query form.

You can also copy and paste an existing query like this one into the query form and run it:

The query ...

```
SELECT DISTINCT ?broaderCPC ?title {
 ?CPC a ?CPCsymbol.
 ?CPCsymbol rdfs:subClassOf cpc:Classification.
 ?CPC rdfs:label "A44B 11/2523".
 ?CPC skos:broader* ?broaderCPC.
 ?broaderCPC dcterms:title ?title
}
ORDER BY ASC(?broaderCPC)
LIMIT 20
```
... will return all the CPC symbols and their titles which are in the CPC hierarchy "above" the CPC symbol "A44B 11/2523".

```sparql
SELECT DISTINCT ?broaderCPC ?title {
  ?CPC a ?CPCsymbol.
  ?CPC rdfs:subClassOf cpc:Classification.
  ?CPC rdfs:label "A44B 11/2523".
  ?CPC skos:broader* ?broaderCPC.
  ?broaderCPC dcterms:title ?title
  ORDER BY ASC(?broaderCPC)
  LIMIT 20
}
```

Figure 7: SPARQL query and result thereof
3. **Linked open EP data web pages**

This section describes the various parts of the web pages which come with the product. Section 1.5, "Accessing Linked open EP data" explains how to access the web pages.

Each of the following sections covers one of the items in the main menu bar.

![Linked open EP data](image)

Figure 8: The main menu bar

### 3.1. About

This part gives a very short overview of this product.

### 3.2. API (application programming interface)

The API is a powerful web interface that can be used by developers to explore, browse and query patent data and retrieve it in a format of their choice.

**Note:** This is a general purpose browser for linked-data data sets, which allows you to explore data by simple pointing and clicking and saves you from having to write short scripts. It is not a professional tool for patent searching.

Sections 2.1-2.3 give examples of how to use this API.

This interface conforms to the [Linked Data API Specification](link). See the specification for more information about its capabilities.

![API functions](image)

Figure 9: Functions of the API

The API has some intuitive functionality:

- **go to:**
  This function is only available if a list of items is displayed.
  It allows you to jump to one of the items currently displayed to another.

- **page:**
  This function is only available if a list of items is displayed.
  You can specify the size of a "page": a page can contain 1, 10, 25 or 50 items.
  You can jump to the first, previous or next page.
• **data format:**
  You can specify how the result should be displayed. There is a choice between HTML (rendered in a type of table), JSON, RDF, text (same as JSON), TTL or XML. Depending on the format you choose and the browser, the data is either directly displayed in the browser or can be downloaded via the browser.

• **view:**
  Here you can specify how much data concerning the retrieved items is displayed and how it is displayed. The available views depend on the (list of) items retrieved. The individual views are described in more detail in the document "API overview" (see section 3.4)

At the right side of each value, there is a triangle which opens a drop-down menu:

![Drop-down menu for a value](image)

The retrieved data can be displayed in one of three views:
- only those items which have the indicated value
- sorted in ascending order
- sorted in descending order.
3.3. SPARQL query

SPARQL is the language used to query linked data sets. You can retrieve the W3C specification here.

Developers can send a query directly to the database at https://data.epo.org/linked-data/query (note: httpS !) or use the SPARQL query web form accessible via this menu item.

Section 2.4 gives an example of how to use this API.

**Note:** This endpoint has a timeout of about one minute. Queries running any longer than this will abort.

---

**Figure 11: SPARQL query web form**

1. You can select one of a few pre-defined queries to get you started. These will be copied into the query form where they can be executed as-is or modified according to your needs.

2. This function enables you to include / exclude the most common prefixes relevant for this data set in the query form. The button on the right-hand side allows you to add further prefixes. In most cases, there is no need to change anything here.

3. This is the SPARQL query editor. Typically the first few lines are used for prefixes, followed by the SELECT clause.

4. You can choose whether the result of a query should be displayed as a table, plain text, JSON or XML.

5. This button starts the query execution.
The query result is displayed at the bottom of the page.

3.4. Documentation

This page contains a set of documentation:

- Data coverage
- API overview
- API reference
- Patent ontology overview
- Patent ontology reference
- CPC ontology overview
- CPC ontology reference

3.5. Download

This page allows you to download the linked data set in bulk. In the case of heavy usage or in a production environment, you should download the data to your own triple store or quad store.

You can download:

- The full data set:
  This is a complete data set in N-Triples format, including patent data, CPC data and vocabulary files. The full data set is large and updated weekly.

- CPC data:
  This is the latest linked data representation of the CPC taxonomy. Typically the CPC taxonomy is updated three-to-five times a year.

- Vocabulary files:
  These are relatively small files containing the ontologies and vocabularies defined by the EPO. Changes to these files are relatively infrequent.

All files are compressed using GZIP and are accompanied by a file containing its SHA 512 checksum.
Downloads

This page provides links to the linked data representation of the EPO patent data to allow the data to be downloaded. This data is in standardized form and can be loaded into an RDF triple store.

There are three kinds of downloads:

- **full dataset** is a complete dataset, including patent data, CPC data and vocabulary files. The full dataset is large and updated weekly.
- **cpc** is the latest linked data representation of the CPC taxonomy.
- **vocabularies** are relatively small files containing the ontologies and vocabularies defined by EPO.

**Full Dataset**

The initial data contains complete and up to date patent data, CPC taxonomy and vocabulary files. This data is updated weekly.

The initial data is provided in **N-TRIPLES** format. It is divided into multiple files to make it easier to download. Each download file is a single valid N-TRIPLES file compressed using GZIP. For each file, there is another file containing its SHA 512 checksum.

**Files**

This data was last updated on 2017-11-05.

- full_part01.nt.gz (908M, checksum)  full_part02.nt.gz (860M, checksum)  full_part03.nt.gz (136M, checksum)
- full_part04.nt.gz (166M, checksum)  full_part05.nt.gz (163M, checksum)  full_part06.nt.gz (165M, checksum)
- full_part07.nt.gz (92M, checksum)  full_part08.nt.gz (81M, checksum)  full_part09.nt.gz (78M, checksum)

---

Figure 12: Download page