European Patent Office

fiveIPoffices
European Patent Office /// Japan Patent Office ///
Korean Intellectual Property Office /// State Intellectual
Property Office of the People's Republic of China ///
United States Patent and Trademark Office

Common Documentation

TECHNICAL SPECIFICATION AUTHORITY FILE V1.0

1. INTRODUCTION

This technical specification defines the minimum data elements required for an Authority File of patent documents and the proposed file formats for an Authority File.

An Authority Files provides *a list of all patent document publication numbers issued by a patent office or a regional organisation* with the primary purpose *to allow other patent offices (and potentially other users) to assess the completeness of the available patent documentation*. In order to allow consistency checks the list should provide all publication numbers for patent documents of which the **numbers** were published at least once. This therefore also includes numbers for which no published document is available (e.g. as is the case for USPTO applications withdrawn late, for destroyed documents) or for which the publication concerned only the publication of bibliographic data.

At the IP5 WG2 Meeting November 29 to December 2, 2011 in Beijing all IP5 Offices agreed on the definition of an Authority File which should primarily include the publication number, the kind code, and the publication date. Following the creation of an initial Authority File in 2012, further discussions on update frequency, quality issues, back-file coverage, and further information fields will be conducted among the IP5 Offices.

2. DEFINITIONS

For the purposes of this Specification:

(a) the term “patent documents” includes patents for inventions, plant patents, design patents, inventors’ certificates, utility certificates, utility models, patents of addition, inventors’ certificates of addition, utility certificates of addition, and published applications therefor. “Documents” means patent documents, unless otherwise stated;

(b) the terms “publication” and “published” are used in the sense of making available:

(i) a patent document to the public for inspection or supplying a copy on request;
(ii) multiple copies of a patent document produced on, or by, any medium (e.g., paper, film, magnetic tape or disc, optical disc, online database, computer network, etc.).

3. REFERENCES

References to the following Standards are of relevance to this Recommendation:

WIPO Standard ST.1	Recommendation Concerning the Minimum Data Elements Required to Uniquely Identify a Patent Document
WIPO Standard ST.2	Standard Manner for Designating Calendar Dates by Using the Gregorian Calendar
WIPO Standard ST.3	Recommended Standard on Two–Letter Codes for the Representation of States, Other Entities and Intergovernmental Organizations
WIPO Standard ST.6	Recommendation for the Numbering of Published Patent Documents
OPD Spec V1.0.0	IP5 Document: One Portal Dossier Specifications, Ver.1.0.0, January 27, 2012
WIPO Standard ST.16	Recommended Standard Code for the Identification of Different Kinds of Patent Documents
WIPO Standard ST.36	Recommendation for the processing of patent information using XML (eXtensible Markup Language)
USPTO Authority File Internet site	http://www.uspto.gov/patents/process/search/authority/index.jsp
EPO Publication Server Data Coverage Files	https://data.epo.org/publication-server/data-coverage

4. BACKGROUND

Within the Common Documentation Foundation Project the IP5 Offices strive to agree on a common set of documentation to be available to all IP5 Offices. Besides a definition on which data sources are to be included in the IP5 Common Documentation it was also considered essential to identify clearly all documents published by an IP5 Office in order to allow the other IP5 Offices to assess the completeness of the available patent documentation.

5. RECOMMENDATIONS

The following paragraphs list the technical recommendations provided for the generation of an Authority File. It should be noted that where alternatives are provided these should be carefully considered in view of usability and evaluation of their advantages and

disadvantages; ideally the IP5 Offices should agree a standardised specification for an Authority File so that the file from each IP5 Office can be used in a similar manner.

5.1. RECOMMENDED MINIMUM SET OF DATA

The Authority File for an office lists all patent documents that have been issued/published by a patent office or regional organisation from the first publication onwards per type and level of publication (kind code A, B, etc.).

The Authority File should contain:

- publication number with the following three elements as originally published by the Offices:
    - the country code (see ST.3)
    - the number of publication
    - kind code (see also ST.16)
- publication date
- optionally, a publication indicator (to indicate withdrawn or missing documents)
- optionally, the priority number(s) and date(s) of said publication.

The provision of the optional data indicated above remains within the discretion of the office generating the Authority File.

This single list shall be sorted first by publication number, secondly by type of document (kind code), thirdly by publication date and (optionally) fourthly by publication indicator and fifthly by priority number(s).
If for a publication no kind code was allocated the corresponding data field should remain empty. In case the kind code or the publication date of a publication are not known any more to the publishing patent office the indicator '*unknown*' should be present in the corresponding data field.

5.2. FIELD FORMATTING

5.2.1. Publication Number

Preferably the publication numbers should be provided in the format "as originally printed" by the Offices preferably with the removal of any non-alphanumeric characters such as dots, commas, dashes, slashes, etc.
The ST.3 two character code for the country or region of the office and the number of the publication can be recorded in a single data field but might also be separated into two separate data fields (see e.g. the XML examples listed below). The kind code is always recorded in a separate field (see below).

Optionally and for consistency reasons the same publication number formatting defined for the One Portal Dossier may also be applied to the Authority File; it should then preferably be explicitly stated in an accompanying file or at another suitable location that this alternative format has been used. In the context of the One Portal Dossier Specification the IP5 offices discussed and defined the publication number format to be used. This format is described in the document *'One Portal Dossier Specification, V.1.0.0, January 27 2012'* in paragraph 3.1.6 *'The docdb publication number format'*. Table 3.1-5 provides

definitions for the three elements (country code CC, kind code KC, number) of the publication number defined above.

5.2.2. Publication date

The publication date shall be formatted in accordance with the definition provided in ST.2 paragraph 1(a) as a single numeric data string comprising eight numerals with 4 digits for the year, two digits for the month and two digits for the day i.e. CCYYMMDD (CC for the century, YY for the years in a century, MM for the month and DD for the day), e.g. *19980315* for 15 March 1998. If deemed necessary year, month, and day might be separated by a hyphen e.g. 1998-03-15.

5.2.3. Optional Data Fields

5.2.3.1. Publication Indicator

A single character code to indicate whether a document was withdrawn ('W') or is missing ('M'). An office may provide in this indicator field other indication characters but should clearly document the meaning for each of the characters in the definition file for the Authority File.

5.2.3.2. Priority Number(s) and Date(s)

Priority numbers shall be indicated in accordance with the provisions of *'One Portal Dossier Specification, V1.0.0, January 27, 2012'* under paragraph 3.1.5 " 3.1.5. '*The docdb application and priority number format'*. Note: The definition of priority numbers includes also the use of country codes and kind codes as defined in table 3.1-4 of the One Portal Dossier Specification.
Priority dates may be included as a further optional criteria to enable cross-checking of the integrity of a priority number. If priority dates are included these should follow immediately the corresponding priority number separated by a single space character. The dates should be presented as a single numeric data string comprising eight numerals with 4 digits for the year, two digits for the month and two digits for the day, e.g. 19980315 for 15 March 1998 (same format as defined for publication date).
As with the country code for publication numbers above the country indicator and the kind code for a priority number may be provided in a separate data field or (in a tsv- or csv-file) separated form the priority number by a single space character.

5.3. RECOMMENDED FILE STRUCTURES

For both of the file structures defined below UTF-8 should be used as encoding standard. It is preferred to provide a single file for all publication numbers; however if this proves impractical due to the resulting file size (c.f. USPTO Authority Files), in which case a single file per year or for a period of years might be considered. To improve file handling a separation of the Authority File into multiple files covering different periods in time can be envisaged, e.g. a dynamic file including data for the present and the last calendar year and a static file including all older data is possible. If deemed necessary separate Authority Files may be provided based on the following criteria:
- post grant vs. pre grant documents

- invention patents, utility models, design patents

5.3.1. Preferred File Structure

A single text coded list with the data fields as defined above. To separate the data field entries in the structured text file predefined unique separator characters must be used e.g. a tsv- (tab-separated values) or csv- (comma-separated values) format could be used, i.e. allowable separators are tab, comma, or semicolon characters. Each record comprises at least the three mandatory fields publication number, kind code, and publication date. Complete list records are separated from each other by the CRLF character. An example for such a list would be:

EP2540531, B1,20040129,, <CRLF>
EP2540532, B1,20040130,, <CRLF>
EP2540632, B1,20040215,W, <CRLF>

For the inclusion of multiple priority numbers in this file structure the individual priority numbers are separated by a single space character.

Advantages of this file structure are:
- the simplicity of the structure with little file overhead allows to achieve a small file size even for large amounts of data;
- both tsv and csv are formats which can be easily supported in any programming environment.

Disadvantages are:
- the file structure cannot easily cope with complex field contents and thus cannot easily be extended with further fields.

5.3.2. Alternative File Structure

As an alternative to the text coded list defined above is to provide an eXtensible Markup Language (XML) coded file.

To assure the correct coding the file should contain an XML declaration in the first line defining UTF-8 coding, e.g.:
<?xml version="1.0" encoding="UTF-8"?>

An ST.36 like XML structure is suggested including list entries which are defined as elements called *document-id* with the following further elements as defined above

- *country*
- *doc-number*
- *kind*
- *date*
- *pub-ind* (optional)
- *prio-claims* (optional) with an indication of the priority *date* (further optional)

The content of the elements should be formatted as defined above. If the country code and the number of publication are to be provided in a single field they are provided as a single entry in the doc-number element in which case the country element is not necessary.

An example for a single entry with only the mandatory information and the long publication number coding would be:

```
<document-id>
  <country>EP</country>
  <doc-number>2416641</doc-number>
  <kind>A1</kind>
  <date>20120215</date>
</document-id>
```

An example for a single entry with the complete information and the publication number coded as printed on the publication would be:

```
<document-id>
    <country>EP</country>
    <doc-number>0840422</doc-number>
    <kind>A2</kind>
    <date>19980506</date>
    </pub-ind>
    <priority-claims>
        <priority-claim>
        <country>IT</country>
        <doc-number>MI9601637</doc-number>
          <kind>A</kind>
        <date>19960731</date>
        </priority-claim>
        <priority-claim>
        <country>IT</country>
        <doc-number>MI9602247</doc-number>
          <kind>A</kind>
        <date>19961029</date>
        </priority-claim>
        <priority-claim>
        <country>US</country>
        <doc-number>08903097</doc-number>
        <date>19970730</date>
        </priority-claim>
    </priority-claims>
</document-id>
```

A DTD-file should be provided together with the Authority File by each office to document the selected structure.

Advantages of this file structure are:
- clear and easily readable file structure;
- flexibility to adapt the file structure to more complex extensions.

Disadvantages of this file structure are:
- due to the format overhead the file size will be significantly larger than for a tsv or csv structured file; this disadvantage can however be reduced by using a single file per year or split files as outlined in 5.3 or file compression tools.

6. DOCUMENTATION AND ADMINSTRATION OF THE AUTHORITY FILE

The technical specification provided above includes a number of options which should allow each office to chose those options allowing the most efficient and reliable Authority File generation within its documentation environment. This flexibility requires however that the Authority File of each office must be accompanied with a *definition file* listing the options chosen by an office and indicating the data coverage of the files. For each Authority File generated a date of production and the coverage of data must be documented. Preferably this information should be coded in the file name.

The primary purpose of the Authority File is to assess the completeness of patent documentation hence the update frequency for the Authority File should be aligned with the performance of these checks. It is proposed to aim for an at least annual update of the files.

Presently the USPTO[1] and the EPO [2]provide public access to Authority Files or data coverage files by allowing to download the files over the internet. For the IP5 Authority File each office should provide a similar public access or distribute the file directly to the other IP5 Offices.

[1] http://www.uspto.gov/patents/process/search/authority/index.jsp
[2] https://data.epo.org/publication-server/data-coverage