Ddoc:

A rule-based approach to Applicant's citations extraction

P.O. Jonsson (po@jonases.se)

Abstract
Ddoc is a rule-based text-mining program originally intended for extracting patent prior art references in patent applications. Ddoc is available as a free service at https://ddoc.eu and can be used on any text to extract patent *publication* or *application* references.

For patent *publication* references, Ddoc uses a word-by-word detection. During the processing, a vector with indicators is populated. The indicators are country code (cc), number (num) and kind code (kc). The retrieved citations are checked against a database and the publication date is added to the output. Ddoc can detect references from 173 different national or regional offices (i.e. for all publications available in Espacenet/OPS and beyond). So-called *bundles* of documents (a single country code followed by a list of numbers) are detected for DE, EP, FR, GB, JP, US, WO and TW patent publication references.

For patent *application* references, Ddoc uses regular expressions. Patent applications from 11 national or regional offices (CN, DE, FR, GB, IT, JP, KR, TW, US; PCT and EP) are detected.

The Open Patent Service (OPS) at the European Patent Office is used for checking extracted numbers. For the most common citation countries, a local database is used to speed up the checking of patent numbers. The local database is 100-150 times faster than the OPS and contains bibliographic information for around 75% of the documents present in Espacenet/OPS. Ddoc has been shown to retrieve at least 95% of the cited documents (recall) with a precision of 95%.

1. Introduction

1.1 Objective
Ddoc is a program used for extracting patent prior art references referred to in any text document. It is implemented in Open Object Rexx 5 and is based on an earlier program with the same name used at the European Patent office 2010-2024 for extracting patent prior art references in patent applications.

The intention of this article is trifold:
(i) To show that rule-based extraction models are still competitive compared to other, more recent machine learning models;
(ii) To give some firm practical advice to anyone working with extraction of patents or patent applications from any kinds of texts;
(iii) To produce some statistics on citations of prior art in patent applications.

1.2 Implementation
Ddoc is written in Open Object Rexx. Ddoc was originally developed 1999/2000 as a simple data entry script in classic Rexx and has since then been ported from OS/2 over Windows 2000 to Windows XP to Windows 7 and Windows 10. Over its lifetime, the program has been migrated

from classic Rexx to Object-Rexx to Open Object Rexx. In 2018 the program was migrated to macOS and has been developed under macOS using ooRexx since.

ooRexx is an interpreted open source language but it might be misleading to think of it as a simple scripting language. In reality it is a fully-fledged object-oriented language, implementing most concepts needed by programmers. ooRexx is implemented in C/C++ and executes very fast, so the disadvantage of having to interpret the program is compensated by the fast execution of all functions and methods.

It might also be misleading to think of Ddoc as a *script*. Although it all started as a simple script simplifying manual data entry, a major rewrite was made in 2007 to make use of OCRed data. In 2018 the program contained more than 14 000 lines of code including all the data extraction routines developed over the years. After 2018 Ddoc has been rewritten to use the Open Patent Service (OPS) at the European Patent Office for checking the extracted numbers. During 2023 a local database has been added to speed up the checking of patent numbers. The local database is 100-150 times faster than the OPS and contains around 90% of the documents making up the minimum PCT documentation and around 75% of the Espacenet total coverage. The less common countries missing in the local database are thus checked using the OPS. Ddoc is currently counting close to 22 000 lines of code after the rewrite since 2018.

1.3 Historic development, raison d'être

As already mentioned, Ddoc started in 1999/2000 as a simple script written by a patent examiner for his own use and for use by his colleagues. The name Ddoc was chosen to reflect the way these documents are denoted in search reports (with a 'D,A' or exceptionally 'D,X' / 'D,Y'), so called D-DOCuments in the lingua of patent examiners.

In the end of 2005, high quality OCRed versions of patent applications became available. In 2007 the first version of the program using text extraction was launched and several improvements have been included since then. Ddoc can be used on any text containing patent application or publication numbers. At the moment it is implemented as a service at https://ddoc.eu.

2. Definitions

Tweak
A *tweak* is the reformatting of text to better fit the following processing steps. A large number of *tweaks* are made for Japanese citations to make the extraction easier, and some well-known OCR errors are also handled this way. See below for details.

Bundle
A *bundle* is a series of patent (or application) citations where the country code is indicated only once at the beginning. These kinds of citations are quite common for US applications but they exist for other country codes as well.

CC
The *Country Code* is the short form for indicating a patent reference, like WO, EP, US etc. The complete list of detected country codes can be found below.

Num
Num denotes the *number part* of a patent reference. In most cases this is a pure number but in some cases, it may also include letters. Examples are TW Invention (TWI) or utility Model (TWM) type publications, where an ‘I’ or ‘M’ is a part of the number.

Kc
The *kind code* establishes the type of a document, i.e. if it is an A1, B2, U etc. type of publication. The kind code can be a single character or a character and a digit. Depending on the reference data system they are mandatory or not. For Ddoc the documents are presented using the Espacenet representation.

Patent application citations
Application citations are references to patent applications or US preliminary/serial applications. Also references to priority documents, or in fact to the priority documents of the application itself, are covered by this definition.

Patent publication citations
Publication citations are "normal" applicant's citations in the form of prior art references to patents or published patent applications. These are the "D-documents" that gave Ddoc its name.

Non-Patent Literature citations
Non-Patent literature references, or NPL for short, are articles, brochures, reports or any other type of references to published prior art that are NOT patents or published patent applications. NPL are not extracted by Ddoc.

Internet citations
These kinds of citations are not considered in this article.

Espacenet
Espacenet is a free database provided by the European Patent Office (EPO), containing the minimum PCT documentation and some further patent documents from further countries. It claims to contain over 150 million patent documents (April 2024). The local database of Ddoc covers the most common patent publication countries and covers close to 112 million unique documents or 75% of the minimum PCT documentation (April 2024).

Docket numbers
Docket numbers are in-house references used in the place of a real patent application number or patent publication number. A docket number is mostly used by US applicants and is a reference that can be used to identify a patent application at the US Patent Office USPTO.

3. Overview

For extraction of patent *publication* citations, Ddoc can be roughly divided in 3 major parts: (i) pre-processing, (ii) main detection and (iii) post-processing. The extraction of patent *application* citations follows a similar pattern. Each part has its own specific purpose which will be described below in more detail. Additionally, there is some "wrapping" programming not related to the extraction per se, which is ignored in this article.

For historical reasons the implementations of the main data extraction routines for patent *applications* and patent *publications* are quite different. Originally only patent *publications* were considered; after all, only publications are D-documents in the true meaning of the word (i.e. useful as citeable prior art). At the time when the main data extraction was devised, the only available option was a word-by-word based detection using relatively complex rules. At one stage it was requested that also *applications* be considered (they are applicant's citations in a more general meaning). In particular it was found that references to US serial applications provided a rich source of information.

During the creation of the patent application extraction algorithms, it was discovered that the pre-processing needed for application citations is not the same as for publication citations. For this reason, the pre-processing was split into two parts, one general part aimed at both types of citations, and the remaining pre-processing tuned to each specific extraction process. It was discovered, that whereas for publication citations the tweaking needed was fairly similar for all citations, it differed quite substantially for different application formats. Each part will be discussed in more detail below.

In Ddoc the *applications* are extracted before the *publications* but this division/order arose for evolutionary reasons only; it does not provide any further advantage than the (relatively small) time saving in re-using one part for all extractions and, in some cases, to make sure that detected applications are not detected again as publications. This is for some countries a non-trivial decision and needs consideration. For a different architecture it could well be that a complete parallelization could be obtained by treating each part, i.e. each application format and the publication formats in parallel. In such a scenario each extraction part could have its own dedicated pre-processing.

The main processing steps are thus:

a. Pre-processing 1 common to applications and publications;
b. Pre-processing 2 for application citations;
c. Extraction of application citations;
d. Post-processing of application citations;
e. Pre-processing 2 for publication citations;
f. Main extraction routine for publication citations;
g. Post-processing for publication citations.
h. Presentation and storing of the results

4. Extraction of application citations

4.1 Pre-processing 1

The initial tweak serves to make detection of country codes easier in the following processing steps. Furthermore, a large number of words and expressions are removed from the initial text so as to make the following detection easier. At the end, also some of the most common OCR errors are mended. Here only a short summary of the steps is listed.

The initial tweak is in its turn split in two major parts, Tweak_JP and Tweak_description_1. The reason for this split is that the tweaking of Japanese citations must be done as early as possible to catch all the “Buzzphrases” that can, with the proper rules, be converted into useful information for the later detection steps. Some elementary linearization steps are however needed before this rather exhaustive tweak. Some of the processing steps below are code page and operating system dependent.

In Tweak_JP:

1. Remove all kind of (trilingual!) diacritics in text, (replace ä and ö with a and o etc.);
2. Replace all whitespace characters (cr, lf, tab) with space (20 hex);
3. Detach all parentheses before parsing;
4. Normalise the text to one space character in all positions;
5. Tweak Emperor Heisei, Showa etc to JP;
6. Align Japanese Emperor Year references;
7. Add JP kind codes wherever possible.

Steps 1-4 are related to the overall processing whereas steps 5-7 are related to tweaks necessary for detection of Japanese *publication* citations. These are discussed in more detail below.

In Tweak_description_1:

8. Remove XML tags;
9. Remove page numbering tags;
10. Remove words that can be confused with country codes (FIG mistaken for Finnish "G" type citation etc);
11. Remove some commonly used words that disturbs the later parsing;
12. Replace some common names of countries with their corresponding country codes;
13. Try to mend some obvious OCR errors;
14. Normalise the text to one space character in all positions;

In the normal case the input to Ddoc is the description of a patent application in the form of text downloaded from for instance the Espacenet at the EPO. This data is normally well formatted and free of anomalies, but since the input to Ddoc can also be text from a variety of sources (HTML code copied from the internet; XML files including formatting tags etc.; other text passages from virtually any source) all these pre-processing steps are needed. Please keep in mind that many more tweaks may be added, but that such tweaks may have unexpected side effects and need to be checked carefully. The tweaks used here are the results of a large number of tests. Please also note that in order to be complete, at least English, French, and German language tweaks need to be considered.

One important pre-processing block (12) tweaks the text and replaces a large number of country names with their respective country codes. One example: 'British' or 'Britain' is replaced throughout the text with GB since it is not uncommon to refer to 'British patent application number...' without mentioning the country code. There is a small risk in doing so (i.e. when British is mentioned in connection with a number that is not a patent reference) but the following post processing will remove most of the false positives. As was already stated earlier, the tweaks used in Ddoc emanate from a large amount of feedback from users and have been found to work well. Depending on the context, one has to strike a balance that is meaningful.

In block 13 there are also a number of tweaks made to try to "mend" OCR errors. One example is to replace W0 (W zero) with WO. Other common OCR errors that have been noticed are that the characters i, l, 1, (one) and the slashes / | \ are often mistaken for each other, like when Kokai is detected by the OCR engine as Kokal. In the same manner U, O and 0 are very often incorrectly detected or mistaken for each other.

4.2 Pre-processing 2 for application citations

During the introduction of the application detection algorithms it became clear that each country has very specific requirements for patent applications and that the lingua used is very country specific. The three different US application formats are the exception to this rule, they need almost no further tweaking to be extracted. They were also the first application formats included.

4.3 Extraction of application citations

All the routines for detecting application numbers follow a similar pattern. Using regular expressions, a number of application formats are detected one by one. Thereafter each citation is checked for existence in a local database or in the OPS.

Currently 13 different types of application formats are extracted in Ddoc. They have been added one by one starting with the most common/easiest ones: US serial applications, US provisional applications, US Design applications, PCT, EP, JP, DE, FR, GB, IT, TW, CN and KR. The

detection uses regular expressions to capture as many different formats as possible. The list is not complete for obvious reasons but a statistical study showed that adding one further application country code would amount to less than 1 % of the application formats available in EP and PCT applications at the EPO. Different national/regional patent offices may have completely different results here. Ddoc was originally aimed at the needs of patent examiners at the EPO and each further format will provide very small improvement to the overall coverage. Depending on the increasing Asian activity in the patent field some further Asian patent application formats were included (CN, KR, in addition to the already existing JP and TW).

4.4 Post-processing of application citations

As a general rule the post processing for applications is less demanding than for publications. One reason is that the regular expressions used to detect applications have already performed a part of the syntax check. Another reason is that applications have, with some exceptions, a more homogenous syntax than publications. Further, applicants tend to cite applications and US Serial application numbers in a less varied way than how they cite publications. The major part of the post-processing for application numbers is the check made in the local database or in the OPS to establish if the number exists as an application or priority number. If this is the case it is assumed that the citation is correctly detected.

There is an additional problem with cited applications in that the applications referred to are not always, at the time of citation extraction, publicly available. Such references will not yield any results since the check is only done for *published* application numbers.

It has been decided to include in the list of extracted application numbers only those applications that can be verified in the way above, even if the extraction algorithms have shown to be very reliable without this check (US serials are already better than 99% correct without the check). The theoretical recall is in this way somewhat lower than it could be, but from a practical point of view, a document that is correctly extracted but not present in a public database is of little use, so this decision has a small practical consequence.

For all applications, a mapping is made to the EPOs standardised format. In the Espacenet, a patent application number consists of: The Country Code (two letters), the year of filing (four digits), and a serial number (variable, maximum seven digits), occasionally with a letter appended at the end. There are exceptions to this rule, so for completeness reference is made to the EPOs definition:

https://worldwide.espacenet.com/help?locale=en_EP&method=handleHelpTopic&topic=applicationnumber

For US applications in particular this mapping can become quite complex. Reference is made here to a document published by the USPTO on a regular basis and titled: "Filing Years and Patent Application Serial Numbers Since 1882":

https://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm

As a further complication for US serials, the filing date indicated in Espacenet/OPS may be different compared to the date indicated at the USPTO depending on the source of the first US filing. I quote here from the document linked to above:

"*The following table allows a user to determine the approximate year when a patent application was filed based on its application series code and serial number. This table is intended to be used*

only as a rough guide and some application serial numbers (e.g., those based on patent filings under the Patent Cooperation Treaty (PCT)) have filing dates that may not fall within the time periods indicated by this table."

It is outside of the scope of this article to explain this in detail but Ddoc identifies and maps all US Serials correctly to the format used in Espacenet/OPS.

5. Extraction of publication citations

5.1 Pre-processing 1 for publication citations

In a dedicated processing step (tweak_jp, steps 5-7 above) an effort is made to detect and include the kind codes for citations of Japanese prior art. Japanese citations in patent applications provide a challenge to every citation extractor for several reasons: (i) a large number of coexisting patent numbering systems exists; (ii) a large number of different publication types and kind codes exists or have existed in the past; (iii) there is an overlap between application numbers and publication numbers (the same number may actually refer to several different documents); (iv) a very varying language when Japanese prior art is cited, possibly due to some extent to machine translations, but more often due to the fact that non-Japanese applicants citing Japanese prior art are not using a harmonised approach. From looking at citations in original Japanese patent applications it appears that a much more coherent vocabulary and syntax is used for Japanese national filings. In particular, Japanese national filings have a specific section where prior art is to be mentioned**.** As for Ddoc, a large effort was made in mid-2009 to try to cope with these variations and from a rough count it is estimated that the retrieval ratio was increased by as much as 10-15% due to this single effort.

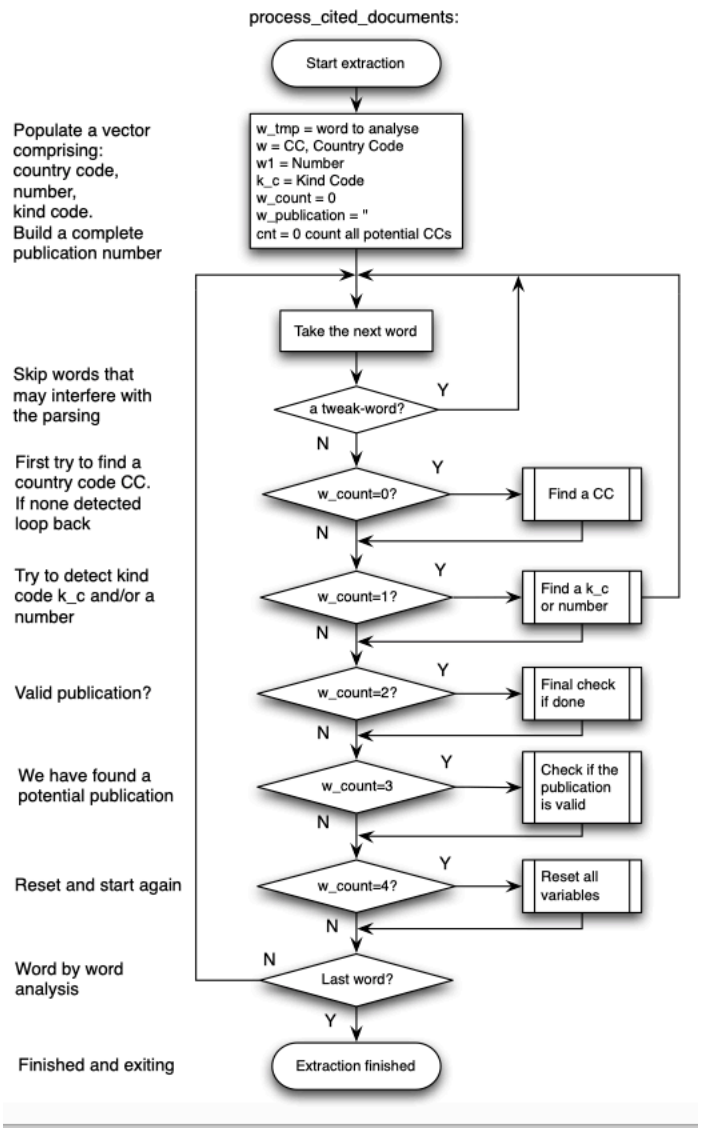
5.2 Pre-processing 2 for publication citations

One of the further pre-processing steps for publications only (fix_bundle) is a very important one, in that it takes care of so-called *bundles* of citations. For US applicants it is quite common to refer to a whole range of citations by referring once to the country code (US) and thereafter list the individual numbers, see for instance EP2275034 (US bundles) or EP2108961 (JP bundle). For a rule-based system it is very difficult to implement this kind of processing in a reliable manner since there is no well-defined separator between the individual numbers and assumptions on the length of each individual number must be made. Consequently, what may work well for one type of bundles may fail for another type merely because the numbers are differently arranged. In Ddoc a number of country codes that are known to have *bundles* have been considered (currently US, JP, GB, EP, WO, FR, DE, TW). Bundles are processed and converted into individual citations before the main data extraction takes place.

The length of most patent applications is normally below 35 pages and mostly 10-20 pages, but the exceptional cases create problems in the processing. In the first benchmarking set of 98 documents tested, one application comprised some 350 pages of text, 10 times more than the average, and applications with more than 4000 pages have been filed at the EPO. The processing of these verbose applications presents a challenge to Ddoc and the processing time becomes prohibitive. Ddoc has a parameter that allows it to ignore any text of the input document after having listed a certain number of words or characters. There are also watchdog items that abort the extraction according to specific criteria.

5.3 Main extraction routine for publication citations

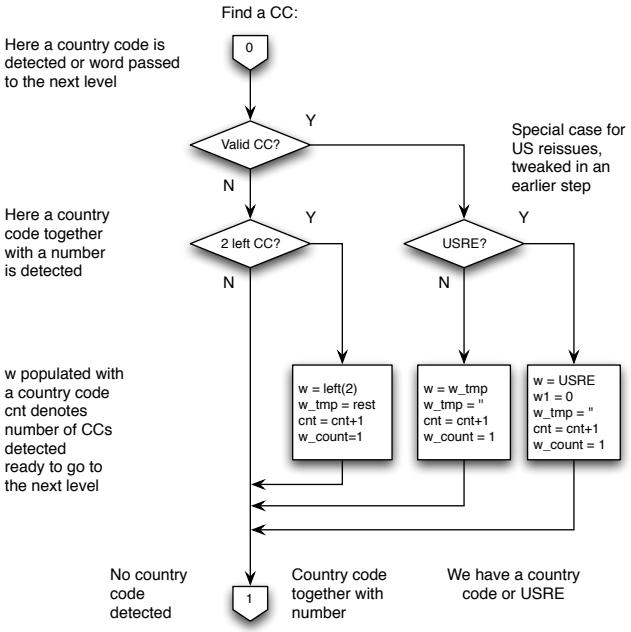
During the processing of patent publications, a vector with indicators is populated. The indicators are country code (CC) number (num) and kind code (kc). In principle the following steps are performed:



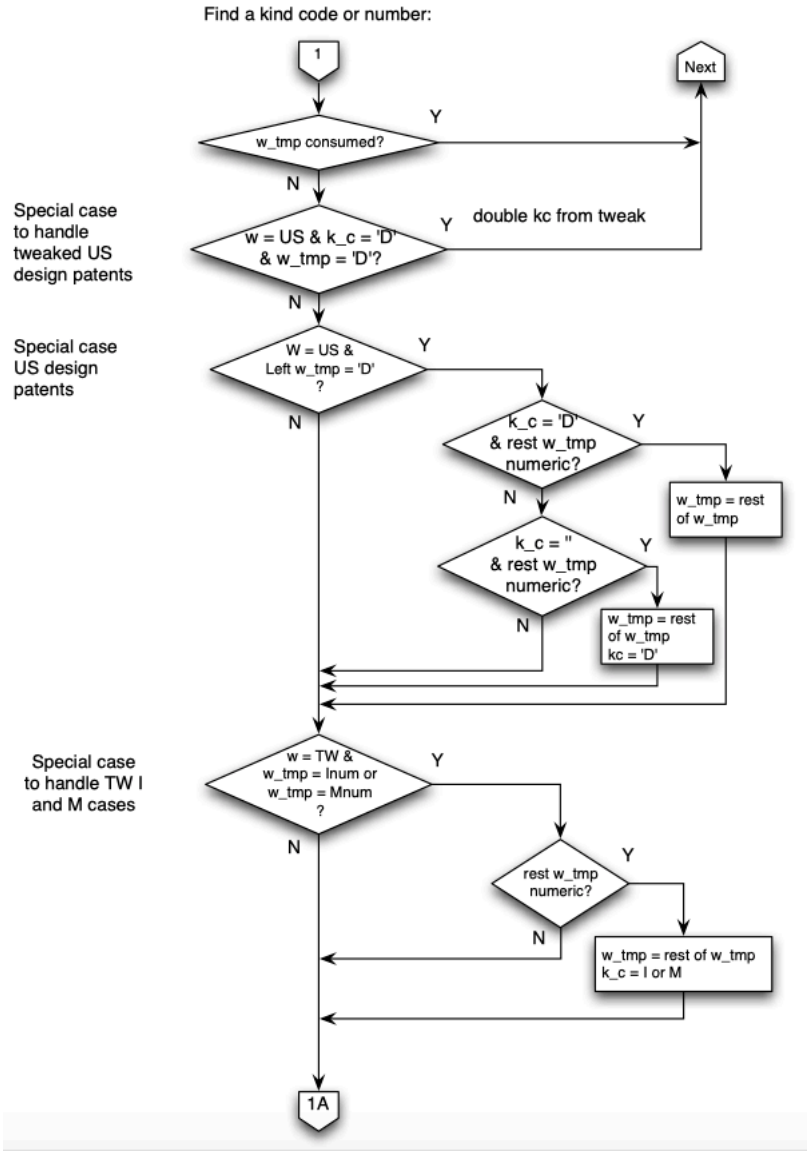
The processing of patent publication references is made word by word, in contrast to the patent application detection where regular expressions are used. The reason for this is that at the time when Ddoc was rewritten to extract the applicant's citations (2007), regular expressions were not available in the Rexx version used at the EPO. Also, even if word-by-word extraction is slow in comparison, it has proven to be very robust and has worked very well over the years. Since the first version written in 2007, the main extraction routine has remained in principle untouched, i.e. it has stayed the same for almost 20 years!

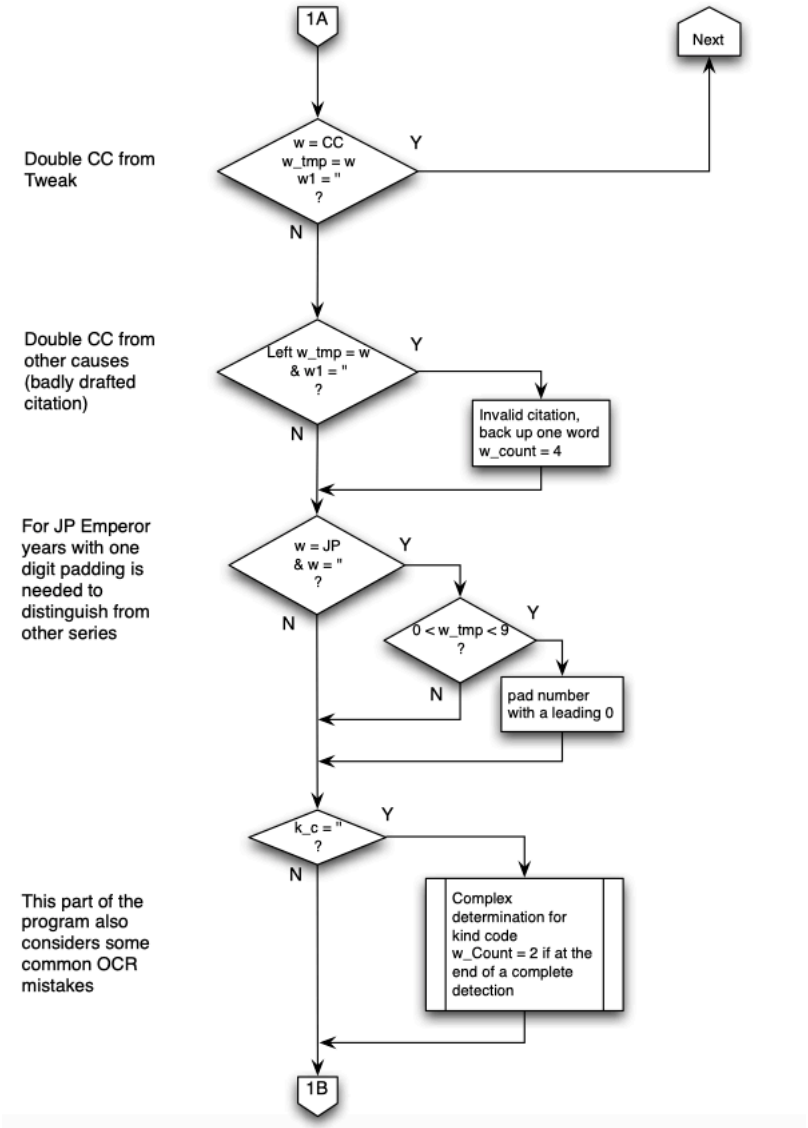
The central data extraction is divided in five different parts that will be explained below by the means of flow diagrams.

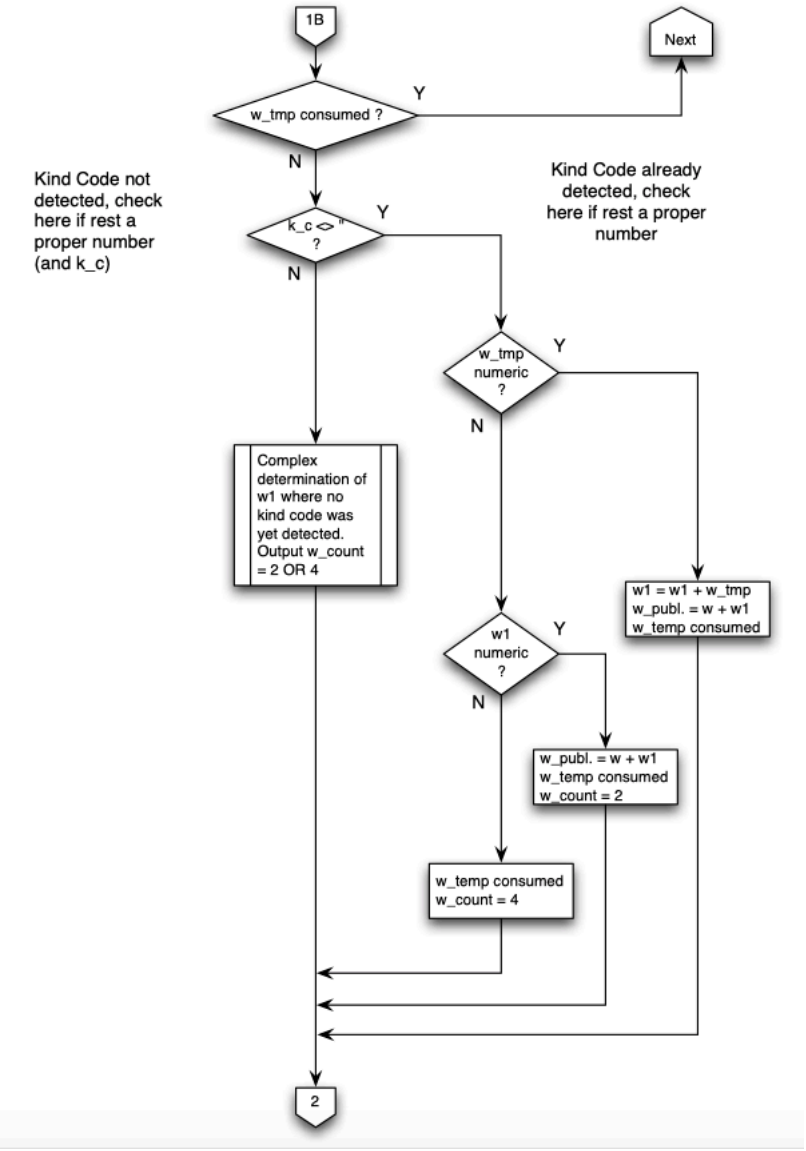
The first part is aimed at detecting a country code. The underlying idea is that the country code (or the mention of a country) will initialize the further processing. Until a country code is found, the words will simply be cast away. Some special cases are taken care of here that could not be taken care of otherwise, like the US reissues.



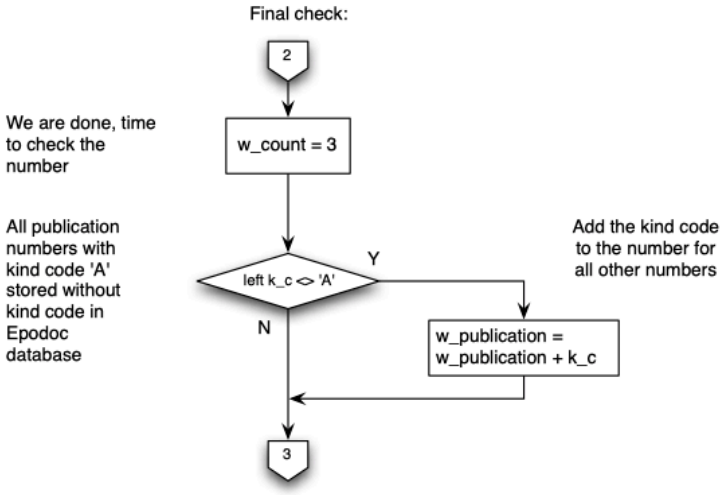
Once a country code has been detected, the next step is to find a kind code or (a part of) a number. In this phase the citation is built up piece by piece until what appears to be a complete citation has been detected. This is the most complex part of the program since numbers or parts of numbers and kind codes may come in an almost random fashion and the vector must be built assuming almost no specific format. Once a complete vector with CC number and, potentially, a kind code, have been found and the next word is not a further number or a kind code, the number is assumed to have been extracted.


Find a kind code or number:
1
Next
w_tmp consumed?
Y
N
Special case to handle tweaked US design patents
w = US & k_c = 'D' & w_tmp = 'D'?
Y
double kc from tweak
N
Special case US design patents
W = US & Left w_tmp = 'D' ?
Y
N
k_c = 'D' & rest w_tmp numeric?
Y
N
w_tmp = rest of w_tmp
k_c = '' & rest w_tmp numeric?
Y
N
w_tmp = rest of w_tmp kc = 'D'
Special case to handle TW I and M cases
w = TW & w_tmp = Inum or w_tmp = Mnum ?
Y
N
rest w_tmp numeric?
Y
N
w_tmp = rest of w_tmp k_c = I or M
1A

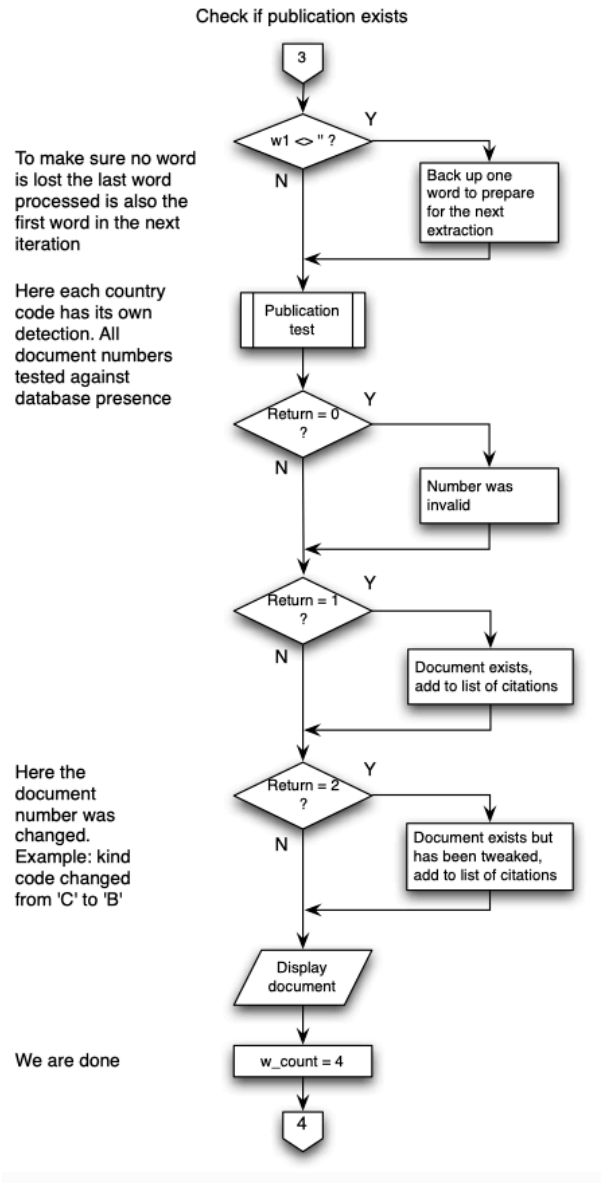

1A
Next
Double CC from Tweak
w = CC
w_tmp = w
w1 = ''
?
Y
N
Double CC from other causes (badly drafted citation)
Left w_tmp = w
& w1 = ''
?
Y
Invalid citation, back up one word
w_count = 4
N
For JP Emperor years with one digit padding is needed to distinguish from other series
w = JP
& w = ''
?
Y
N
0 < w_tmp < 9
?
Y
N
pad number with a leading 0
This part of the program also considers some common OCR mistakes
k_c = ''
?
Y
N
Complex determination for kind code
w_Count = 2 if at the end of a complete detection
1B



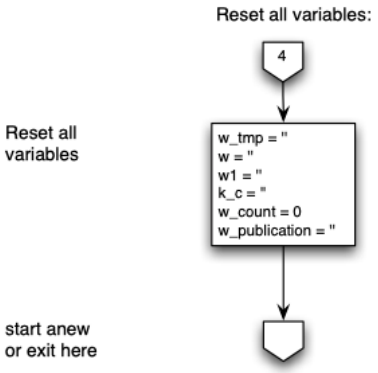
The next step is to see if the number is actually complete, and if this is the case pass the control on to the next level.



If it is positively decided that a potential publication has been extracted completely, a test is made to see if the number is available in the databases (local database or OPS). This step is explained in detail below (post processing).



Once a citation has been validated (or discarded), all temporary variables are reset and the program loops back to where it was again, starting to look for the next potential country code.



5.4 Post-processing for publication citations

The different post processing routines for publications have various grades of complexity, depending on how common the specific country codes are and how intricate the number systems are. Countries with a long history of patents, like Germany or Japan require a relatively complex checking whereas other countries have less complex patenting schemes. EP, WO and CN for instance are relatively straightforward to post-process.

In all cases a check is made in a local database or in the OPS to see if the number exists as a patent publication number. If this is the case it is assumed that the citation is correctly detected. This provides for a high precision on the expense of recall, since there might be perfectly correct publication numbers that are not available in the OPS. One example is older Asian documents. The check for existence is furthermore not 100% fool proof since number combinations exist (Gene sequences or certain standards for instance) that overlap with patent publication numbers. During the check some further tweaking may take place to correct the length of the number or padding it with zeros etc. For some complex patent numbering schemes like the Japanese, the document number may need a “special code” to avoid conflicting numbering (example: JPH used instead of JP for certain ranges of documents). Another problem that may arise is that patent *publications* can have the same numbers as other, unrelated patent *applications*. This is the case for instance for some Japanese numbering schemes. In the post-processing it is also possible to tweak the kind code or to add a kind code if it was wrongly detected from the extraction or simply missing. In this way a JP document may first be tested as a “B” type publication and thereafter as a “C”, “U” or “Y” until a number is found. Without this kind of post-processing the retrieval rate would be much lower for certain numbering schemes.

One further problem is the presence of common words in uppercase like IN, AT etc. that are also country codes (here: India and Austria). In principle a rule-based system may, without further checking, accept such cases as correct extractions. Consider the phrase "IN 100-110 degrees Celsius" or "AT 50-150 bar". Also, if only uppercase letters are considered (most examples like the ones above will be in lowercase) a sufficient number of examples exist where the detection incorrectly detects such phrases as patent numbers. Certain gene sequences fall into this category of problems as well. See for instance EP2302079.

Ddoc currently detects 105 country codes, all countries present in the Espacenet/OPS. The post processing takes place in one of three groups of filters:

Thorough check:
For a number of country codes, a dedicated post-processing is made. The filters have been added one by one according to the need. Some are more elaborate than others but in general these filters can be seen as expert rules in the sense that they implement a lot of knowledge that has been collected over the years. The following country codes have a dedicated post-processing:
AT, AU, BE, CA, CH, DE, EP, FR, GB, JP, KR, RU, TW, US, USRE and WO. Typical post processing includes padding the number with 0s, adding or removing a kind code etc. The reference here is the Espacenet or the OPS database.

Decent check:
For a number of country code with a similar structure a less stringent check is made. Again, some padding is made and country code added or removed as the case may be. The following country codes share a decent post-processing: CN, DK, FI, LT, NL.

Rudimentary check:
A rudimentary check is made for all candidates that are not in any of the above two groups of country codes. Currently some 85 countries belong to this category. Basically, only a plausibility

check is made to see if the number is of an acceptable length and that the country code and kind code exists.

All-in-all the following country codes are checked:
AM, AP, AR, BA, BG, BR, AM, AP, AR, AT, AU, BA, BE, BG, BR, BY, CA, CG, CH, CL, CN, CO, CR, CS, CU, CY, CZ, DD, DE, DK, DO, DZ, EA, EC, EE, EG, EM, EP, ES, FI, FR, GB, GC, GE, GR, GT, HK, HN, HR, HU, ID, IE, IL, IN, IS, IT, JO, JP, KE, KG, KR, KZ, LT, LU, LV, MA, MC, MD, ME, MN, MO, MT, MW, MX, MY, NI, NL, NO, NZ, OA, PA, PE, PH, PL, PT, RO, RS, RU, SA, SE, SG, SI, SK, SM, SU, SV, TH, TJ, TN, TR, TT, TW, UA, US, UY, UZ, VN, WO, YU, ZA, ZM

Kind code:
In the post-processing, the kind code (if detected) is also checked. The list below shows all kind codes that exist in the Espacenet database at the EPO. In principle a kind code can be a single character (A,B,C etc) or a combination of a character and a digit (A3,B2 etc). New kind codes are also added at times when a country decides to amend their patent publication schemes. In Ddoc a fixed list of actual existing kind codes is used, but in general it could be sufficient to accept a single character [A-Z] or a single character and a single digit [A-Z][0-9]. These are all kind codes detected in Ddoc: A A1 A2 A3 A4 A5 A6 A7 A8 A9 A0 B B1 B2 B3 B4 B5 B6 B7 B8 B9 B0 C C1 C2 C3 C4 C5 C6 C7 C8 C9 C0 D D1 D2 E E1 E2 F F1 F2 F3 G G1 G2 H H1 H2 H9 I I1 I2 I3 I4 I5 K K1 L M P P1 P2 P3 Q R R1 R2 R3 S S1 T T1 T2 T3 T4 T5 T6 T7 T8 T9 T0 U U1 U2 U3 U4 U5 U6 U7 U8 U9 U0 W W1 W2 Y Y1 Y2 Y3 Y4 Y5 Y6 Z Z1 Z2.

6 Final processing steps

After the main extraction is finished and the entire description parsed, Ddoc stores the extracted citations in a file, together with information on the application or publication date.

7 Further aspects, practical considerations

Exotic country codes
Some specific areas of technology may have a preference for prior art from a specific country or region. One example can be production and refinement of alcohol from sugar canes or bio-waste, where Brazilian and Swedish prior art may compete. In other cases, new emerging technologies arise in countries that have a short or non-existing tradition of patents. Examples here may be India or China. In such cases the absolute number of citations may be relatively low but increasing, and, from a technical point of view, more important than a large bulk of technologically outdated citations.

An extraction level of 70-80% (precision and/or recall) can quite easily be obtained with also the most rudimentary extraction, for instance by detecting the US publications alone, but the most interesting documents may be the odd ones. Here a rule-based system will prevail over systems needing training, since it is difficult to find reliable training material for rare formats.

In order to be successful in the extraction of citations, one should consequently study carefully the context where the citations are to be used and maybe in some cases put more weight to such areas that may a priori look less relevant in terms of numbers.

During the continuous improvement of Ddoc, using evaluation of feedback from users (mostly cases where Ddoc failed the recognition of a specific prior art), a large collection of realistic "difficult" citations has been collected. Some of the items have been mentioned here.

DD citations

DD citations are patents or patent applications from the former German Democratic Republic, DDR, or East Germany. They are rare but we have collected a number of applications where such citations are made, see for instance DE102006058465, WO2007056470 or EP2085622. Whereas it is quite simple to devise a filter for these citations in a rule-based system like Ddoc, it may be very difficult to do the same for a system relying on training.

Docket numbers

A docket number is sometimes used in the place of a real patent application or publication number, in particular by US applicants. Even if such a number can theoretically be searched at the USPTO to locate a specific prior art (after it has been made available to the public) it involves a lot of manual effort to locate such a number. These are some applications that contain references to docket numbers: EP1770140, EP1717719, EP1621166, WO2010151857, WO2009158583, WO2009158510.

8 Benchmarking

8.1 Results

A number of corpora have been used to check the retrieval ratio of Ddoc. For a random selection of European patent applications, it has been found that Ddoc can retrieve at least 95% of the cited documents (recall) with a precision of 95%. This is at par with manual extraction by patent experts.

In the view of the author it is very difficult to reach higher values for larger sets of documents. The reasons for this are manifold but the following are the most common error sources:

- Documents that are incorrectly OCRed or incorrectly translated lead to around 1-2 % of missed citations;
- Documents with incorrect citations (from typos, reversed numbers, incomplete or wrong citations) amount to at least a further 1%;
- Documents that are not publicly available, like US docket numbers, amount to at least 1%;
- Documents that are theoretically publicly available but not available in Espacenet/OPS, where the check is made. These documents amount to around 1% as well.

The following definitions have been used below:

Precision:	relevant citations retrieved / all citations retrieved
Recall:	relevant citations retrieved / actually relevant citations
F-Score	F = 2*P*R / (P+R) (= Harmonic mean)

Source: Wikipedia:

https://en.wikipedia.org/wiki/Precision_and_recall

https://en.wikipedia.org/wiki/F-score

The corpora used for checking Ddoc will be made available for benchmarking elsewhere. At the time of creating these corpora around 2010, the text used was taken from in-house databases at the EPO. Those original texts have been replaced with the corresponding fulltext documents made

available by the EPO at the Espacenet/OPS. Listed here are the results from three recent runs. Originally these corpora contained 100 items each, but some of the items never became public and so the corpora have been shrunk so as to contain only publicly available documents.
The performance is measured here in terms of recall and precision but the performance has also been measured using trec_eval-9.0.7, a tool provided here:

https://trec.nist.gov/trec_eval/index.html

The corpora will be made available in a format suitable for the trec_eval tool.

8.2 Corpus: 92_Gold

This is a manually checked corpus containing 92 EP applications published 2008-2011. It is believed to be completely correct, hence "Gold". Some EP applications arriving to the EPO via the PCT route have the PCT application text.

Result of comparison between Benchmark and Ddoc

13 Dec 2023 12:54:31 (c) P.O. Jonsson 2023

Ref. Data: 223 Applications and 625 Publications

Total: 848 citations in 82 AP files and 90 PN files

Ddoc: 214 (208 correct) AP and 614 (607 correct) PN

Total: 828 (815 correct) citations in 82 AP files and 90 PN files

Ddoc:	Precision	Recall	F-Score
AP	0.972	0.933	0.952
PN	0.989	0.971	0.980
ALL	0.984	0.961	**0.973**

8.3 Corpus: EP_Direct_99

EP_Direct_99 is a corpus containing 99 EP Applications published 2010-2011 and filed directly at the EPO. They have been extracted using the "Cited Documents" field in Espacenet, with some manual corrections. This is considered a "Silver" corpus since it has not been completely checked manually.

Result of comparison between Benchmark and Ddoc

24 Dec 2023 16:25:36 (c) P.O. Jonsson 2023

Ref. Data: 394 Applications and 2395 Publications

Total: 2789 citations in 94 AP files and 98 PN files

Ddoc: 370 (361 correct) AP and 2377 (2344 correct) PN

Total: 2747 (2705 correct) citations in 94 AP files and 98 PN files

Ddoc:	Precision	Recall	F-Score
AP	0.976	0.916	0.945
PN	0.986	0.979	0.982
ALL	0.985	0.970	**0.977**

8.4 Corpus: EP_Direct_92

EP_Direct_92 is a corpus containing 92 EP Applications published 2007-2015 that have been extracted using the "Cited Documents" field in Espacenet, with some manual corrections. This is also considered a "Silver" corpus since it has not been completely checked manually. Some applications arriving to the EPO via the PCT Route have the PCT application text.

Result of comparison between Benchmark and Ddoc

11 Dec 2023 22:37:00 (c) P.O. Jonsson 2023

Ref. Data: 22 Applications and 419 Publications

Total: 441 citations in 12 AP files and 91 PN files

Ddoc: 21 (20 correct) AP and 425 (408 correct) PN

Total: 446 (428 correct) citations in 12 AP files and 91 PN files

Ddoc:	Precision	Recall	F-Score
AP	0.952	0.909	0.930
PN	0.960	0.974	0.967
ALL	0.960	0.971	**0.965**

9 Citation Statistics

From a batch processing of some European patent applications in Oct-Dec 2010 some statistics can be given. Unfortunately, the data is rather old. The applications are a mix of EP, PCT and French applications filed with the EPO. From the total number of applications approximately 65% had citations (application and/or publication citations) that were found by Ddoc. As can be seen from the distribution below, the number of documents cited drops rapidly after the first few application/publication formats. A similar analysis today (2024) would probably disclose many more CN citations.

Total number of files processed	49707	
Total number of EP ANs	29924	60,20%
Total number of PCT ANs	13231	26,62%
Total number of FA ANs	3160	6,36%
Total number of other	3392	6,82%

All files extracted	43989	
Files with citations	28713	65,27%
Files with no citations	15209	34,57%

Files with bundles	4217	
Files with long description	2117	

Number of application citations

ALL APs	20785	100,00%	
US Ser	16142	77,66%	1
PCT APs	1848	8,89%	2
EP APs	924	4,45%	3
JP APs	822	3,95%	4
GB APs	459	2,21%	5
DE APs	334	1,61%	6
FR APs	173	0,83%	7
IT APs	83	0,40%	8

Number of publication citations

ALL PNs	190975	100,00%	
US PNs	92661	48,52%	1
WO PNs	51837	27,14%	2
EP PNs	20856	10,92%	3
DE PNs	10428	5,46%	4
JP PNs	9108	4,77%	5
FR PNs	2610	1,37%	6
GB PNs	1695	0,89%	7
CN PNs	323	0,17%	8
KR PNs	265	0,14%	9

10 Final Remarks

Since patent numbering schemes are based on fixed rules, it is safe to assume that a rule-based extraction system will be very stable once these rules have been defined. Defining rules is however very labour intensive compared to machine learning approaches.

Further, even if rule-based extraction systems are mostly static, some maintenance is necessary:

(i) Number formats in the databases sometimes need to be changed;
(ii) US Serial ranges are only rough estimates until the end of the year.

Examples of (i):

- The introduction in 2012 of an additional “H” in certain Japanese documents in Espacenet (JP replaced with JPH due to a clash between some recent Emperor numbers with other JP numbers);
- The introduction at the EPO of the Unitary Patent required the detection of EP-C documents in addition to EP-A and EP-B documents.

Example of (ii):

- The activation of a new US series of applications (“serials”).

11 References

When this article was originally written in 2010 it contained references to a number of scientific articles, since the intention was to publish it. For different reasons this never took place and the article is now meant more as a documentation of the rule-based text mining tool Ddoc and all external references have been excised.