SPIF - Patent Identification Format

A Modest Proposal to Bring Sanity to Basic Patent Identification Information

Version: 0.1.1

Contents

Reference Materials

Overview

Our goal is to define a minimum viable product for the clear exchange of patent asset lists as typically used in the patent market and communicating lists of patent assets. Our hope is that the format will be useful in additional contexts and spur the development of future versions of the standard that may include alternative formats, e.g., XML/JSON/etc. The initial release will focus on specific guidance for the “big-5” offices (US, CN, EP, JP, KR) as well as PCT applications.

The goal is to publish a simple, free, human-readable, open format for use in communicating patent identifying information.

Change History

20210129 - First revision

- Updated mandatory column names to include “ - SPIF” to avoid collisions with column names from prior export formats
- Addressed the question of how additional countries and formats are handled
- Additional minor cleanups

Context

Asset lists provided by sellers and brokers in the (i) M&A, (ii) licensing, and (iii) patent sales process are surprisingly hard to work with.[1] This causes in-house teams, outside counsel, and more to spend significant work multiple times over cleaning a list and making it suitable for use across tools (e.g., Cipher, Derwent, Questel, Innography, Unified Portal, etc.). Some samples of problematic entries:

Sample Input	**Likely Meaning**	**Type (app or pub)**	**Comments**
CNZL201480022610.X	CN201480022610	app	Extra check digit and prefix
ZL03827150.8	CN03827150	app	Missing country code
GB3123328	EP3123328A1	pub	Wrong country code (there's no UK patent with that number)
CH,2420637	EP2420637A2	pub	Wrong country code (there's no Swiss patent with that number)
ZA 2015/000715	ZA201500715	app	Extra zero

[1] There are multiple root causes including the lack of standardized data interchange formats, inconsistent practices by national PTOs, as well as inconsistent practices by docketing and analytics tool vendors.

WO002/001258	WO02002/001258	app	Missing year digits
GB2405228,319405.7	GB2405228B8	pub	Extra stuff, app number?
US2014214418	US20140214418A1	pub	Missing zero
US7123456BB	US7123456B2	pub	Made up kind code “BB”
KR1341015B1	KR101341015B1	pub	Missing 10 prefix
20067013095,Korea	KR20067013095	app	Country name spelled out as suffix
US2017163019A1	WO02017163019A1	app	There is a US app with that number, but they meant WIPO
GB2568035	GB2568035B	pub	Ambiguous
GB2568035	EP2568035B1	pub	
US10229419	US10/229,419	app	Ambiguous
US10229419	US10229419B2	pub	

Significant time, money, and resources could be better focused on analyzing the patents in the package rather than sorting out the list.

This is not a solution for patent characteristic identification, priority dates, live dead status. This is a simple asset identification and disambiguation solution.

Initial Solution and Minimum Requirements

We have chosen Excel as the initial format because it is a broadly available tool and easily human-readable. We are defining specific columns and the basic format of the rows to reduce variability and increase the chances that the assets listed in the file can be identified successfully. We plan to promulgate this solution to docketing and analytics tool vendors as well as the patent buying and selling community.

General Goals

1. A simple solution (“KISS”)
    a. Analogy: what is the street address of a property I want to look at?
    b. Out of scope: what are the characteristics of that property?
2. Easily parsed asset list for commonly used systems
    a. A handful of patent analytics tools (e.g. IFI Claims (Cipher, Google Patents, Unified); Derwent; Innography) are widely used.
    b. The format is designed with the expectation that the input format should be as easy as possible for tools to unambiguously identify the correct matching patent numbers for listed patent assets.

3. Human checkable data - e.g., titles and filing dates are recommended to enable a person to scan and see potential mistakes
4. Provide some redundant information to help disambiguate the columns and increase the chance of correctly identifying the assets
5. Utility patents and utility models only
    a. Not plant patents or design patents
    b. Not trademarks or copyrights
6. Use existing research/development from WIPO/EPO etc. where available
    a. Our brief investigation found that these solutions are overkill for this project (see "Reference Materials")
7. Most importantly, we can call the checking tool "Spiffy."

Coverage

1. Asset types
    a. Utility patent applications and issued patents
    b. Utility models
    c. PCT applications
2. Countries/Regional Offices
    a. Initial countries
        i. US
        ii. EP
        iii. JP
        iv. KR
        v. CN
        vi. PCT
    b. Additional countries
        i. At the present time, other countries ("unsupported countries") can be expressed using a standardized convention that will remove ambiguities
3. Timeframe
    a. Because patent offices have changed formats over time, we are focused on supporting assets with priority dates after 1 Jan 2000

EP Assets: List Only Once or List with Designated States Separately

For EP patents that are issued (or pending), one of two options is permitted. In no event should the asset be listed in formats like DE2551856B1, GB2551856B1, EP(CH)2551856B1, or other similar variants that impair understanding of the asset list.

Option 1: Only one row for the EP with the EP designator should be provided. Consider, e.g., EP2551856B1 as of June 2, 2020, it is:

- granted in: CH, DE, LI
- expired in: AL, AT, BE, BG, CY, CZ, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR

We recommend Option 1 if referring to the entire set of EP nationalizations and designated states.

Option 2: We are aware that it is important to know which countries an EP asset is validated for docketing purposes and annuity payments. For that reason, the same EP asset can be listed in multiple rows with the EP patent number (e.g., EP2551856B1) with another column for which country the asset is validated in.

We **recommend** calling that column “Validation Country - SPIF”.

We anticipate that future versions of the standard that support additional countries/treaties will adopt this approach for other treaties, e.g., ARIPO.

Defined Columns in the Master Data Sheet

The required columns must be named exactly as shown. The formats of the numbers for the supported patent offices and the PCT are shown in the Country Specific Guidance Section. These formats have been selected for broad compatibility with existing tools.

Column Name	**Description**	**Examples**	**Priority**
Application Number - SPIF	The patent-office assigned application/serial # including the country code and omitting spaces and check digits. There should be nothing else in the field (e.g., no leading/trailing whitespace) Reasons to include country code: Need it to look up regardless. Eliminates confusion between the actual serial number and designated country for EP. It eliminates the need to merge columns for matching and also solves issues with Excel reformatting the fields into numbers. Check digits should be eliminated	US13624395 EP11759439 KR1020127027195 CN201180015433 JP2010549365 WO2011JP056984	Required where there is no Publication Number Recommended otherwise but can be blank
Publication Number - SPIF	The patent office assigned patent	US9123456B2 EP2551856B1	Required when issued or published, otherwise

	number (when available), or publication number (when available). including the country code and the kind code. There should be no spaces, punctuation, or other characters. There should be nothing else in the field (e.g. no leading/trailing whitespace) Check digits should be eliminated.	KR101487211B1 CN102822907B JP4879373B2 US20130014973A1 EP2551856A1 KR1020127027195A CN102822907A JP2011118054A1 WO2011118054A1 (Blank is ok if no publication number is available)	blank

Additional Fields

Additional fields may be provided and parsers supporting SPIF are not required to evaluate these columns when performing matching. However, if these column names are present the data in the columns **must** conform to the provided descriptions.

Column Name	**Description**	**Examples**	**Priority**
Title - SPIF	The title of the patent. This makes human verification of a file easier	High frequency cable, high frequency coil, and method for manufacturing high frequency cable	Recommended
Filing Date - SPIF	The filing date of the patent. This makes human verification and some machine verification simpler. Excel date (not text), set Excel Date format to: "`yyyy-mm-dd`" (ISO-8601)	2012-09-21	Recommended
Country - SPIF	Two-digit country code. This should be present in the numbers already but may be provided as a separate column. This is *not* the validation country for EP assets.	EP US DE GB WO	Optional

Family identifiers - Multiple potential column names are permitted as shown at right.	It is often helpful to be able to realize that multiple assets are all in the same family. Family identifiers are not mandatory; however, if they are provided the columns must be named according to the following pattern: “Family - <Type>” where <Type> is replaced with: INPADOC, DocDB, Internal, or a product-defined string, e.g. “Family - XYZTool”. This can be used to improve matching and/or spot common problems with the data.	Family - Internal 2011-01 Family - INPADOC 20110929WO2011118054A1 Family - DocDB 44672642	Recommend that at least one family column be provided; multiple columns are permitted if properly named

File Format Requirements and Formatting

1. Microsoft Excel 2007+/OOXML (e.g. “.xlsx” file format)
    a. Not CSV, not anything else
    b. Not “classic” Excel, e.g. “.xls”
    c. Note, a goal is to have the format be human-readable and machine-readable. This will help build trust in the results.
        i. We recognize the problem of doing this (people will screw it up). The alternative is people can’t check and correct their files so they will end up with a CSV version and an Excel version and ...
2. The sheet containing the data is called “Master Data - SPIF”
3. The first row has the names of the columns only (Row 1 in Excel) and starts in Column A
4. One row per asset only (Rows 2 and up)
    a. To the extent practical, each asset should only appear once, e.g. do not list both the publication and the patent as two rows.
5. No merged cells anywhere in the Master Data Sheet
6. Column order recommendation, any order is allowed provided the mandatory columns are named exactly:
    a. Application Number - SPIF
    b. Publication Number - SPIF

Checking Tools

The authors have built an open-source checking tool in the style of "lint" called Spiffy that is available at https://github.com/erikoliver/spiffy.

Country-Specific Guidance

General nomenclature:

- Application: the identifier assigned to a patent application, pre-publication
- Publication: the identifier assigned to a published application document
- Patent: the identifier assigned to a published grant

Practices:

- We provide templates describing the allowed format for the currently supported countries and patent asset types
- The following placeholders are used in the templates:
    - CC = Two character ISO-3166 country code, e.g. US, DE, GB, WO, EP, etc.
    - # = 1-digit (0-9)
    - YYYY = Four-digit year, e.g. 2000 or 2019
    - YY = 2 digit year suffix, e.g. for 2001 it would be 01
    - KK = Kind code, an uppercase letter, optionally followed by a digit (e.g. "A" or "B1" or "B2")
- Numbers should be zero-padded to the given number of digits, e.g. a 7-digit template "US#######" would represent the number 654321 with zero-padding, "US**0**654321"
- Spaces should be omitted and there should be no whitespace on either side of numbers, e.g. " US7123456B2 " is wrong because of the extra spaces on either side of the patent number.
- Check digits (e.g. .3, .X suffix) should be omitted
- Punctuation should be omitted (e.g. commas, dashes, periods, slashes)

1. USPTO

a. Application #s: US######## (always 8 digits, no slashes or commas. First two digits are the series number, not the year suffix)
b. Publication #s: USYYYY#######KK (pad the publication # to 7 digits)
c. Patent #s: US#######KK / US########KK (either 7 or 8-digit US patent numbers are accepted)
d. Reissue
    i. Pre 2001: USRE#####E (pad the patent # to 5 digits)
    ii. Post 2001: USRE#####E# (pad the patent # to 5 digits)
e. Other items - not handled

2. EPO

a. Application #s: EP######## (pad to 8 digits)
b. Publication #s: EP#######KK (pad to 7 digits)

c. Patent #s: EP#######KK (pad to 7 digits)
d. Reminder - Do not list EPO examined patents that have been nationalized with other country codes.

3. CNIPA

a. Application #s
   i. Pre August 2007: CNYYYY1###### / CNYYYY2###### / CNYYYY8###### / CNYYYY9######
   ii. Post August 2007: CNYYYY1####### / CNYYYY2####### / CNYYYY8####### / CNYYYY9#######
b. Publication #s
   i. Pre August 2007: CN1######KK / CN2######KK
   ii. Post August 2007: CN1########KK / CN2########KK
c. Patent #s
   i. Pre August 2007: CN1######KK / CN2######KK
   ii. Post August 2007: CN1########KK / CN2########KK

4. KIPO

a. Application #s: KR10YYYY####### / KR20YYYY####### (utility model)
b. Publication #s
   i. Pre 2004: KRYYYY#######A / KRYYYY#######U (utility model)
   ii. Post 2004: KR10YYYY#######A / KR20YYYY#######U (utility model)
c. Patent #s: KR10#######B# / KR20#######Y# (utility model)

5. JPO

a. Application #s: JPYYYY###### (6 digits)
b. Publication #s: JPYYYY######KK (pad to 6 digits)
c. Patent #s: JPYYYY######KK (pad to 6 digits) / JP#######KK
d. Utility model #s: JPYYYY######U (pad to 6 digits) / JP#######U

6. WIPO / PCT

a. Application #s
   i. Pre 2004: WOYYYYCC##### (pad to 5 digits)
   ii. Post 2004: WOYYYYCC###### (6 digits)
b. Publication #s: WOYYYY######KK (6 digits)
c. (No issued patents)

7. All other Countries (Unsupported Countries)

a. At this time, only general formatting conventions are provided for other countries. Their interpretation will remain tool dependent.
b. Applications #s: CC<application number>KK - no spaces, punctuation, or check digits should be included between the CC and the KK, KK may be omitted if there is none
c. Patent #s: CC<patent number>KK - no spaces, punctuation, or check digits should be included between the CC and the KK

Tool Handling - Import

Because the “Application Number - SPIF” and “Patent Number - SPIF” column can contain countries for which standardization has not been agreed, compliant tools are **required** to provide a summary after the import of a SPIF-compliant file to help users understand what was read in.

The summary should indicate:

- Total number of records provided
- Total number of SPIF compliant items
- Total number of SPIF items from supported countries and # of matched items
- Total number of SPIF items from unsupported countries and # of matched items

We **recommend** that the tool provide an export or download showing the matching in detail to enable users to diagnose issues—primarily with unsupported countries.

Tool Handling - Export

Tools should include a suitable application number and publication number in every row of an export in the SPIF columns (“Application Number - SPIF” and “Publication Number - SPIF”). If your export supports it, color coding **may** be used to distinguish between entries for fully supported countries and unsupported countries. If done, use the Excel “good” style for entries in the SPIF columns that are from supported countries and the “neutral” style for entries in the SPIF columns that are from unsupported countries. An example is shown in Figure 1.

FIGURE 1: Sample showing the use of Excel styles to optionally distinguish supported vs. unsupported countries in the export.

A	B
Application Number - SPIF	**Publication Number - SPIF**
US12044448	US8238590B2
AU2011201800A1	AU2011201800B2

For reference, the RGB color codes for the three “good” and “neutral” styles in Excel are given below.

Style	**Font color**	**Background color**
Good	#006100	#C6EFCE
Neutral	#9C6500	#FFEB9C

Open Issues

General

1. Assets without a serial number, e.g. a patent were just filed and the serial number is not available or not in the docketing system.
2. Pre-publication assets (e.g. <18 months) so there is a valid serial # but it is not yet matchable in analytics tools.
    a. More specifically, how can an analytics tool communicate that an asset may be valid, but just not displayable versus an outright error in formatting the #?
3. Data quality
4. Other countries - Assets outside of the 5 supported countries + WIPO may be listed and their handling is tool-dependent. It is expected that this SPIF standard will be revised over time to handle additional countries.

Country-Specific Concerns

1. No open issues at this time

Future Directions

Items to consider for future versions of the format.

1. Extra worksheet in Excel with an explanation of the data, e.g.
    a. Compliance with SPIF v0.1.1
    b. Creation date of the datasheet
    c. Author of the datasheet
    d. The tool used to create the datasheet
    e. Note implementers are recommended to ignore any sheets not named “Master Data - SPIF” when working with the OOXML file and thus it is permissible to add other data in the current version of SPIF, it is just not specified.
2. Roadmap for other countries
3. Other formats, e.g. XML, JSON, etc.

Reference Materials

1. WIPO Standard ST.36 - https://www.wipo.int/export/sites/www/standards/en/pdf/03-36-01.pdf
2. DocDB - http://documents.epo.org/projects/babylon/eponet.nsf/0/6266D96FAA2D3E6BC1257F1B00398241/$File/T09.01_ST36_User_Documentation_vs_2.5.8.1_en.pdf
3. Derwent country/region compatibility notes https://clarivate.com/derwent/dwpi-reference-center/dwpi-coverage/
4. EPO format documentation for:

    a. China https://www.epo.org/searching-for-patents/helpful-resources/asian/china/numbering.html
    b. Japan https://www.epo.org/searching-for-patents/helpful-resources/asian/japan/numbering.html
    c. Korea https://www.epo.org/searching-for-patents/helpful-resources/asian/korea/numbering.html
5. "Turn of the Century" Wired Magazine, January 1, 2002
    a. "...no organized attempt has as yet been made to establish any system, each manufacturer having adopted whatever his judgment may have dictated as the best, or as most convenient for himself." At the time, American screws, nuts, and bolts were custom-made by machinists, and there was no guarantee that bolts made by shops on different streets, let alone in different cities, would be the same. "So radical a defect should exist no longer," Sellers proclaimed."