EPO Sequence Listings

Product documentation

Date	Authors
19 February 2019	Processes and statistics: Adrian Stoica Development Support, Dir. 2833 Search and Knowledge Sequence data format: Stéphane Nauche Team Cancer Immunology, Dir. 1111 Distribution and packaging: Davide Lingua, Sonia Kaufmann Patent Data Services, Dir. 5413 Publications

1. INTRODUCTION

In the context of this document the sequence listing publication in text format means the process by which the EPO makes available to the public sequence listing data in a searchable format (more specifically, EMBL format). Subject to this process are PCT and EP sequence listings which have been filed at the EPO for search purposes.

1.1. PROCESSES

Nowadays the EPO (like most patent offices) requires applicants to file sequence listings in a standardised electronic format (currently WIPO ST. 25) in order to make it easier to provide information for different publication and search processes. However, in the past sequence listings have been submitted in multiple formats, including on paper.

The EPO built through the years an internal database including a collection of sequence listings in text format going back as far as 1989. This database does not contain strings of data that have been embedded in the text of the specification but only those filed as such in a separate machine-readable document. The data as filed by the applicant has been re-formatted and marked up: for this reason it is not the original version.

The sequence listing publication in text format occurs once per week and is triggered by the legal publication of a patent application.

An integrated software solution (so called ‘sequence publication process’) is used to execute the weekly sequence listing publication in text format. This process is in place since April 2012 and the extracted sequences are made available on a weekly basis via the EPO Download Area *(link? contact patentdata@epo.org for further details).*

Additionally, a back-file extraction (‘backlog process’) took place from February to November 2018 to make the data from 1989 to March 2012 available.

File name structure in the back-file

The back-file data is available as a bulk data product.

The back-file is delivered with data from two sources 1) the weekly files from the sequence publication process and 2) the database extractions originating from the backlog process.

1) The weekly files from the sequence publication process as of 2012 to date are sorted by year (each year is one directory) and are a collection of weekly packages:

Weekly package name: SEQL_yyyyww.zip, containing

yyyyww_SEQL_EPDirect.txt	EP filings
yyyyww_SEQL_PCT.txt	PCT filings where EPO is the Searching Authority
yyyyww_SEQL_EPRegio.txt	PCT filings entering the EP regional phase

2) The database extractions originating from the backlog process for the period 1989 to 2012, arranged by week of extraction, are included in the directory “Years 1989 to 2012” and contain files of the kind:

SEQL_BACKLOG_201812_2403_0200.txt
201812 is the year and week number and 2403 is the day and month when the extraction took place

File name structure in the front-file

The weekly package name has the following format: SEQL_yyyyww.zip, and it will contain the following files:

yyyyww_SEQL_EPDirect.txt	EP filings
yyyyww_SEQL_PCT.txt	PCT filings where EPO is the Searching Authority
yyyyww_SEQL_EPRegio.txt	PCT filings entering the EP regional phase

1.2. DATA

The data included in this data set originates from the EPO internal database and contains the publication number, legal publication date and the number of sequences for that publication.

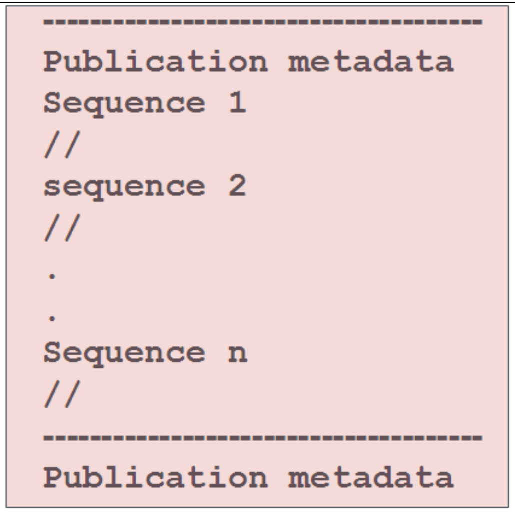
Data format

The entries in the datasets are structured so as to be usable by human readers as well as by computer programs. The structure is systematic enough to allow computer programs easily to read, identify, and manipulate the various types of data included.

Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data which make up the entry.

The two exceptions to this are the sequence data lines and the feature table lines, for which a fixed format was felt to offer significant advantages to the user. Users who write programs to process the database entries should not make any assumptions about the column placement of items on lines other than these two: all other line types are free-format.

Data for each sequence listing starts with the publication metadata. The publication is only given above the first sequence of a given sequence listing and is not repeated for each sequence



```
--------------------------------------
Publication metadata
Sequence 1
//
sequence 2
//
.
.
Sequence n
//
--------------------------------------
Publication metadata
```

Structure of a SEQL text file

Publication meta-data

Sequence listing separator

The Publication meta-data and sequence(s) information for a given application publication are preceded by a line containing 38 "-"

Example

```
--------------------------------------
```

The RT line :

The RT (Reference Title) lines give the title of the patent publication. It is followed by two spaces, then by the actual data

Example :

```
RT  Cripto blocking molecules and therapeutic uses thereof
```

The RA line

The RA (Reference Author) lines list the inventors of the patent application. It is followed by two spaces, then by the actual data

Example:

```
RA  MINCHIOTTI G., RUVO M., DE FALCO S., MARASCO D., LONARDO E., PARISH C., ARENAS E.;
```

The RL line :

The RL (Reference Location) lines should contain information on the patent application publication:

- The first RL line discloses the publication number, kind code and the patent publication application date
- The second RL lines discloses the application number, kind code and the patent application filing date
- The subsequent RL line contains the patent priority number and date
- The next RL line applicant name

The words "Patent publication number" are to be followed by the patent's application number and the patent's application date

Example :

```
RL  Patent publication number EP2280022-A1; 02-Feb-2011
RL  Patent application number EP20090166967; 31-Jul-2009
RL  Patent earliest priority EP20090166967; 31-Jul-2009
RL  CONSIGLIO NAZIONALE RICERCHE [IT];
```

Sequence data

The sequence information lines:

The ID (IDentification) line is always the first line of a sequence entry. The format of the ID line is:
ID   <1>; SV <2>; <3>; <4>; <5>; <6>; <7>
The tokens represent:

1. Primary accession number, which is the application number followed by a "_" and the sequence number
2. Sequence version number
3. Topology: 'circular' or 'linear'
4. Molecule type (see note below)
5. Data class which is PAT for Patents
6. Taxonomic division, according to EMBL rules

7. Sequence length

Note: Molecule type: this represents the type of molecule as stored and can be any value from the list of current values for the mandatory mol_type source qualifier. This item should be the same as the value in the mol_type qualifier(s) in a given entry.

Example:

```
ID   EP20090166967_1; SV 1; linear; Other DNA; PAT; UNC; 24 BP.
```

The AC Line

The AC (ACcession number) line lists the accession numbers associated with the entry. This number is identical to the first element of the ID line. The AC number will be changed by the EBI (European Bioinformatics Institute) upon the incorporation in the public patent sequence repositories

Example

```
AC   EP20090166967_1;
```

The OS Line

The OS (Organism Species) line specifies the preferred scientific name of the organism which was the source of the stored sequence. In most cases this is done by giving the Latin genus and species designations known.
Alternatively the English common name is given. In case of Artificial sequence or synthetic construct, those words will be given in the OS line

Example:

```
OS   synthetic construct
```

The RN Line

The RN (Reference Number) line gives a unique number to each reference Citation within an entry. This number is used to designate the reference in comments and in the feature table. The format of the RN line is: The reference number is always enclosed in square brackets.
The subsequence RN line normally refers to the sequence number or SEQ ID NO of the sequence listing

Example:

```
RN   [1]
RN   Sequence ID NO: 1
```

The FH Line

The FH (Feature Header) lines are present only to improve readability of the Feature information. The lines contain no data and may be ignored by computer programs. The format of these lines is always the same:

The first line provides column headings for the feature table, and the second line serves as a spacer. If an entry contains no feature table (i.e. no FT lines - see below), the FH lines will not appear.

Example:

```
FH   Key             Location/Qualifiers
FH
```

The FT Line

The FT (Feature Table) lines provide a mechanism for the annotation of the sequence data. Regions or sites in the sequence which are of interest are listed in the table. In general, the features in the feature table represent signals or other characteristics reported in the cited references. In some cases, ambiguities or features noted in the course of data preparation have been included. The feature table is subject to expansion or change as more becomes known about a given sequence.

For more information on nucleotide Feature Table Definition Document:

WebFeat:

A complete list of feature table key and qualifier definitions, providing full explanations of their use.
URL: http://www.ebi.ac.uk/embl/WebFeat/index.html

EMBL-Bank Annotation Examples.

A selection of EMBL-Bank approved feature table annotations for some common biological sequences (i.e., ribosomal RNA, mitochondrial genome).
URL: http://www.ebi.ac.uk/embl/Standards/web/index.html

For more information on amino acid Feature Table Definition Document:

Unitprot User manual:

A complete list of features and qualifiers
URL : http://expasy.org/sprot/userman.html

Example:

```
FT   source          1..24
FT                   /mol_type="Other DNA"
FT                   /organism="synthetic construct"
FT                   /note="synthetic primer"
```

The SQ Line

The SQ (SeQuence header) line marks the beginning of the sequence data and gives a summary of its content.

Nucleotide sequences

The line contains the length of the sequence in base pairs followed by its base composition. Bases other than A, C, G and T are grouped together as "other". (Note that "BP" is also used for single stranded RNA sequences, which is not strictly accurate, but has been used for consistency of format.) This information can be used as a check on accuracy or for statistical purposes. The word "Sequence" is present solely as a marker for readability.

Example:

```
SQ   Sequence 24 BP; 7 A; 8 C; 5 G; 4 T; 0 other;
```

Amino acid sequences

The line contains the length of the sequence

Example:

```
SQ   SEQUENCE   497 AA;
```

The Sequence Data Line

The sequence data line has a line code consisting of 5 blanks.
Nucleotide sequences

The sequence is written 60 bases per line, in groups of 10 bases separated by a blank character, beginning at position 6 of the line. The direction listed is always 5' to 3', and wherever possible the non-coding strand (homologous to the message) has been stored. Columns 73-80 of each sequence line contain base numbers for easier reading and quick location of regions of interest. The numbers are right justified and indicate the number of the last base on each line.

Example:

```
SQ   Sequence 786 BP; 193 A; 173 C; 183 G; 237 T; 0 other;
     ggcccagccg gccatggctg aggttaaatt gatggaatcc ggtggtggtt tggttcaacc        60
     aggtggatct atgaagttgt cctgtgttgc ttctggtttt actttttcca actactggat       120
     gaactgggtt agacaatcac cagaaaaagg attggaatgg gttgctgaga ttagagttaa       180
     atccaacaat tacgctactc actacgctga atctgttaga ggaagattca ctacctccag       240
     agatgactcc aagtcttccg tttacttgca aatgaacaat ttgagaggtg aggatactgg       300
     aatctactac tgcagtagag tttactacta tggtcacgac tacgctatgg attactgggg       360
     tcaaggtacc tccgttactg tctcgagtgg ttccacttct ggttctggaa agccaggatc       420
     aggagagggt tctaccaagg gatccgctgt tgacattgtt ttgactcaat ctccagctat       480
     tatgtctact tcattgggtg aaagagttac tatgacttgt actgcttcat ctccagtctc       540
     atctacctat ttgcactggt accaacaaaa gcctggttct tctcctaagt tgtggatcta       600
     ctccacctct aagttggctt ccggtgttcc tgatagattt tctggatctg gttccggaac       660
     ttcatactca ttgactattt cttccatgga agctgaggat gctgctacct actactgtca       720
     ccagtaccac agatcaccaa gaaccttcgg tggtggtacc aaattggaga ttaaaagagc       780
     ggccgc                                                                  786
```

Amino acid sequences

The sequence is written 60 amino acids per line, in groups of 10 residues separated by a blank character, beginning at position 6 of the line. The direction listed is always N terminal to C terminal.
Example

```
SQ   SEQUENCE   499 AA;
     MGARASVLSG GKLDAWEKIR LRPGGKKKYR IKHLVWASRE LERFALNPGL LETTEGCQQI
     LEQLQPTLRT GTEEIKSLYN AXATLYCVHQ RIEVKDTKEA LEEVEKIQKK SQKKTQQAAM
     GEGNSSQVSQ NYPIVQNAQG QMVHQPLSPR TLNAWVKVVE EKAFNPEVIP MFSALSEGAT
     PQDLNTMLNT VGGHQAAMQM LKDTINEEAA EWDRTHPPQA GPIPPGQIRE PRGSDIAGTT
     SNLQEQIRWM TSNPPIPVGE IYKRWIILGL NKIVRMYSPV SILDIRQGPK EPFRDYVDRF
     FKTLRAEQAT QEVKNWMTDT LLVQHANPDC KTILRALGPG ATLEEMMTAC QGVGGPGHKA
     RVLAEAMSQA SNSAAAIMMQ KGNFKGPRRI KCFNCGKEGH LARNCRAPRK KGCWKCGKEG
     HQMKDCTERQ ANFLGKIWPS NKGRPGNFLQ NRPEPTAPPA ESFGFGEEIA PSPKQAPKQE
     PKGELYPLAS LKSLFGNDP
```

The sequence separator:

The sequence terminates with a line containing two "/"

Example:

```
//
```

Example:

```
RT  Cripto blocking molecules and therapeutic uses thereof
RA  MINCHIOTTI G., RUVO M., DE FALCO S., MARASCO D., LONARDO E., PARISH C., ARENAS E.;
RL  Patent publication number EP2280022-A1; 02-Feb-2011
RL  Patent application number EP20090166967; 31-Jul-2009
RL  Patent earliest priority EP20090166967; 31-Jul-2009
RL  CONSIGLIO NAZIONALE RICERCHE [IT];
ID   EP20090166967_1; SV 1; linear; Other DNA; PAT; UNC; 24 BP.
XX
AC   EP20090166967_1;
XX
OS   synthetic construct
XX
RN   [1]
RN   Sequence ID NO: 1
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..24
FT                   /mol_type="Other DNA"
FT                   /organism="synthetic construct"
FT                   /note="synthetic primer"
XX
SQ   Sequence 24 BP; 7 A; 8 C; 5 G; 4 T; 0 other;
     cagacctgaa ggagacctat tccc                                            24
//
ID   EP20090166967_2; SV 1; linear; Other DNA; PAT; UNC; 24 BP.
XX
AC   EP20090166967_2;
XX
OS   synthetic construct
XX
RN   [1]
RN   Sequence ID NO: 2
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..24
FT                   /mol_type="Other DNA"
FT                   /organism="synthetic construct"
FT                   /note="synthetic primer"
XX
SQ   Sequence 24 BP; 6 A; 7 C; 5 G; 6 T; 0 other;
     gtcagcgtaa acagttgctc tacc                                            24
//
--------------------------------------
RT  VIRAL VECTORS ENCODING A DNA REPAIR MATRIX AND CONTAINING A VIRION-ASSOCIATED SITE
SPECIFIC MEGANUCLEASE FOR GENE TARGETING
RA  DANOS O., IZMIRYAN A., BOURDEL A.;
RL  Patent publication number WO2011007193-A1; 20-Jan-2011
RL  Patent application number WO2009IB06689; 17-Jul-2009
RL  Patent earliest priority WO2009IB06689; 17-Jul-2009
RL  CELLECTIS [FR]; DANOS OLIVIER [FR]; IZMIRYAN ARAKSYA [FR]; BOURDEL ALIX [FR];
ID   WO2009IB06689_1; SV 1; linear; Unassigned Protein; PAT; UNC; 497 AA.
XX
AC   WO2009IB06689_1;
XX
OS   Human immunodeficiency virus type 1
XX
RN   [1]
RN   Sequence ID NO: 1
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..497
FT                   /mol_type="Unassigned Protein"
FT                   /organism="Human immunodeficiency virus type 1"
XX
SQ   SEQUENCE   497 AA;
     MGARASVLSG GKLEAWEKIR LRPGGKKKYR MKHLVWASRE LERFALNPGL LETAEGCQQI
     IEQLQPTLKT GSEELKSLFN TVATLWCVHQ KVDVKDTKEA LDKIEEVQNE SQQKTQQAAA
     GTGSSSKVSQ NYPIVQNAQG QMVHQPLSPR TLNAWVKVVE EKGFKPEVIP MFSALSEGAT
     PQDLNMMLNI VGGHQAAMQM LKETINEEAA EWDRVHPVHA GPIPPGQMRE PRGSDIAGTT
     STLQEQIGWM TGNPAIPVGD IYKRWIILGL NKIVRMYSPA SILDIRQGPK EPFRDYVDRF
     YKTLRAEQAT QEVKNWMTET LLVQNANPDC KSILRALGPG ATLEEMMTAC QGVGGPSHKA
     RVLAEAMSQA QHTNILMQRG NFKGQKRIKC FNCGKEGHLA RNCKAPRKKG CWKCGKEGHQ
     MKDCTERQAN FLGKIWPSNK GRPGNFPQNR LEPTAPPAEN WERGEEMTPL PKQEQKNKDP
     PPLVSLKSLF GNDPLSQ
//
```

2. STATISTICS UP TO 2018

2.1. NUMBER OF PCT AND EP APPLICATIONS PER YEAR BETWEEN 1989 AND 2018

Taken from the back-file statistics “publication_dossier3_extended.csv”

Year	PCT files	EP files	Total
1989	1	0	1
1993	2	1	3
1994	251	101	352
1994	703	355	1058
1995	898	473	1371
1996	1121	561	1682
1997	1386	783	2169
1998	2084	956	3040
1999	2502	995	3497
2000	2889	1384	4273
2001	4066	1577	5643
2002	3812	1543	5355
2003	2278	1500	3778
2004	2427	2742	5169
2005	2996	3049	6045
2006	2998	2660	5658
2007	2842	2586	5428
2008	3197	2895	6092
2009	3558	2732	6290
2010	3067	3158	6225
2011	2688	5128	7816
2012	2804	3807	6611
2013	2858	3245	6103
2014	2980	2951	5931
2015	2876	2968	5844
2016	3200	3480	6680
2017	3418	3592	7010
2018	3296	4137	7433
Total	**67.198**	**59.359**	**126.557**

2.2. NUMBER OF SEQUENCES PUBLISHED PER YEAR BETWEEN 1989 AND 2018

Year	Total	No. of sequences
1989	1	6
1992	3	173
1993	352	8229
1994	1058	23762
1995	1371	49043
1996	1682	42022
1997	2169	61286
1998	3040	98063
1999	3497	147010
2000	4273	286336
2001	5643	957044
2002	5355	669724
2003	3778	297400
2004	5169	683543
2005	6045	2294643
2006	5658	1044872
2007	5428	1553505
2008	6092	1187783
2009	6290	1789741
2010	6225	1771724
2011	7816	6734348
2012	6611	1622880
2013	6103	1607965
2014	5931	2483375
2015	5844	4307287
2016	6680	4546135
2017	7010	2052427
2018	7433	3711037
Total	**126.557**	**40.031.363**