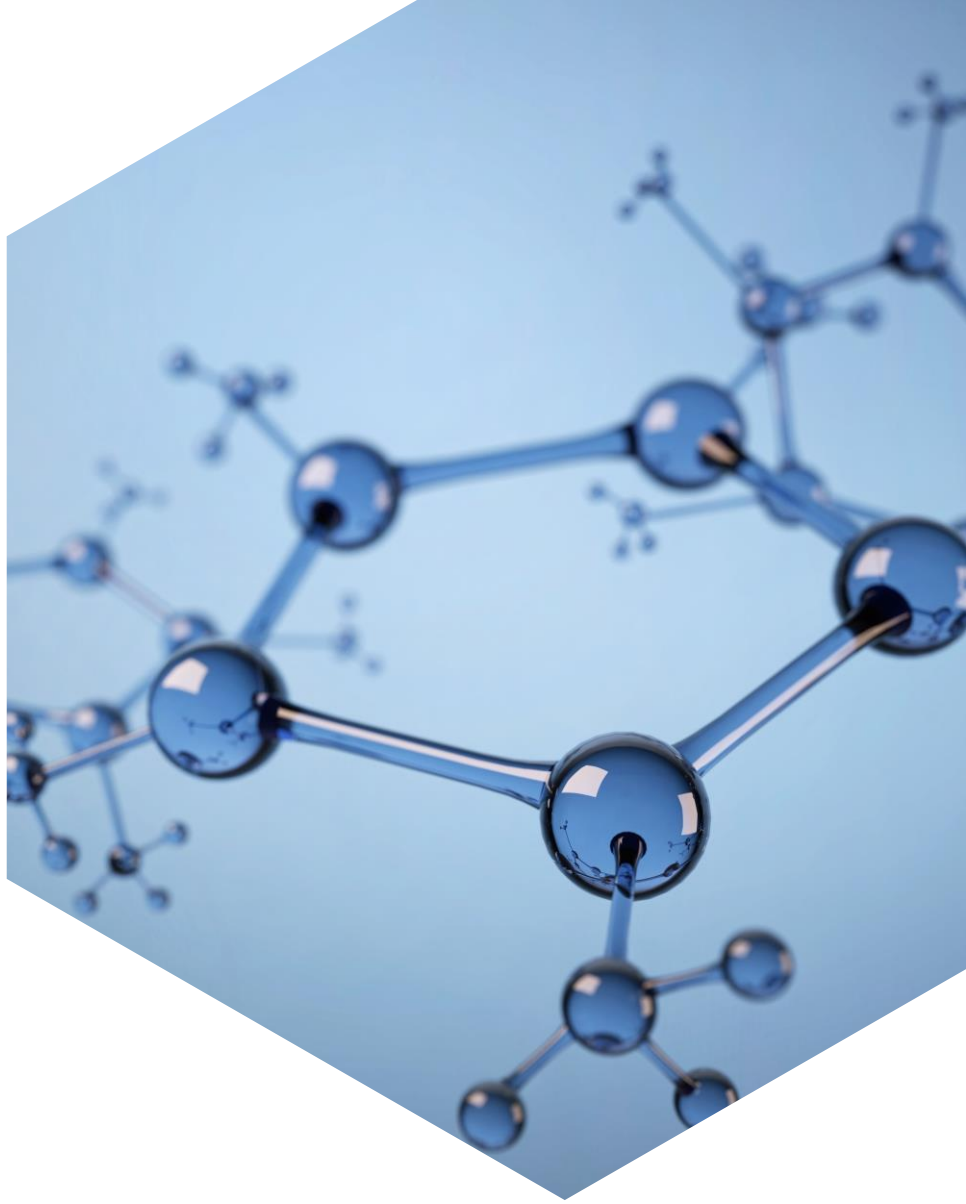
TIPS & TRICKS FOR SEARCHING FULL-TEXT DATABASES IN CAS STNEXT®

Sarah W. Stokes, Senior Product Specialist, IP, FIZ Karlsruhe

CAS
A division of the
American Chemical Society

AGENDA

- Full-text patent data on CAS STNext
  - Data standardization and normalization
- Best practices for searching in full-text files
  - Key Terms
  - Claims search fields
  - Numeric Property Searching



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Full-text and value-added patent data on CAS STNext

- CAS STNext provides access to international patent information from a large number of patent authorities around the world
- Many databases of patent information on CAS STNext contain added value – indexing, enhancement, structure searchability, and other aspects
- CAS STNext also provides access to a number of full-text, primary source data for national and international patent issuing authorities
- Use both value-added files and full-text files together for more comprehensive results and sophisticated search techniques, including NPS

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure
CAS
A division of the American Chemical Society

Full-text patent databases on CAS STNext

- Primary source patent databases
  - Includes full text of the original document, including claims
- Many full-text patent databases available on STNext from around the world, including English translations for all
- Full-text databases are enhanced with Numeric Property Searching (NPS) capabilities as well as data standardization and Key Terms
- Full-text file cluster can be used by entering FIL FULLTEXT (cluster includes some files without patent documents)

Data standardization at FIZ Karlsruhe

US 2019/0274269 A1

US 10,470,428 B2

200710187643. 1

19173119.9

JP 2020-4368 A

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

US 20190274269

US 10470428

CN 2007-10187643

EP 2019-173119

JP 2020004368

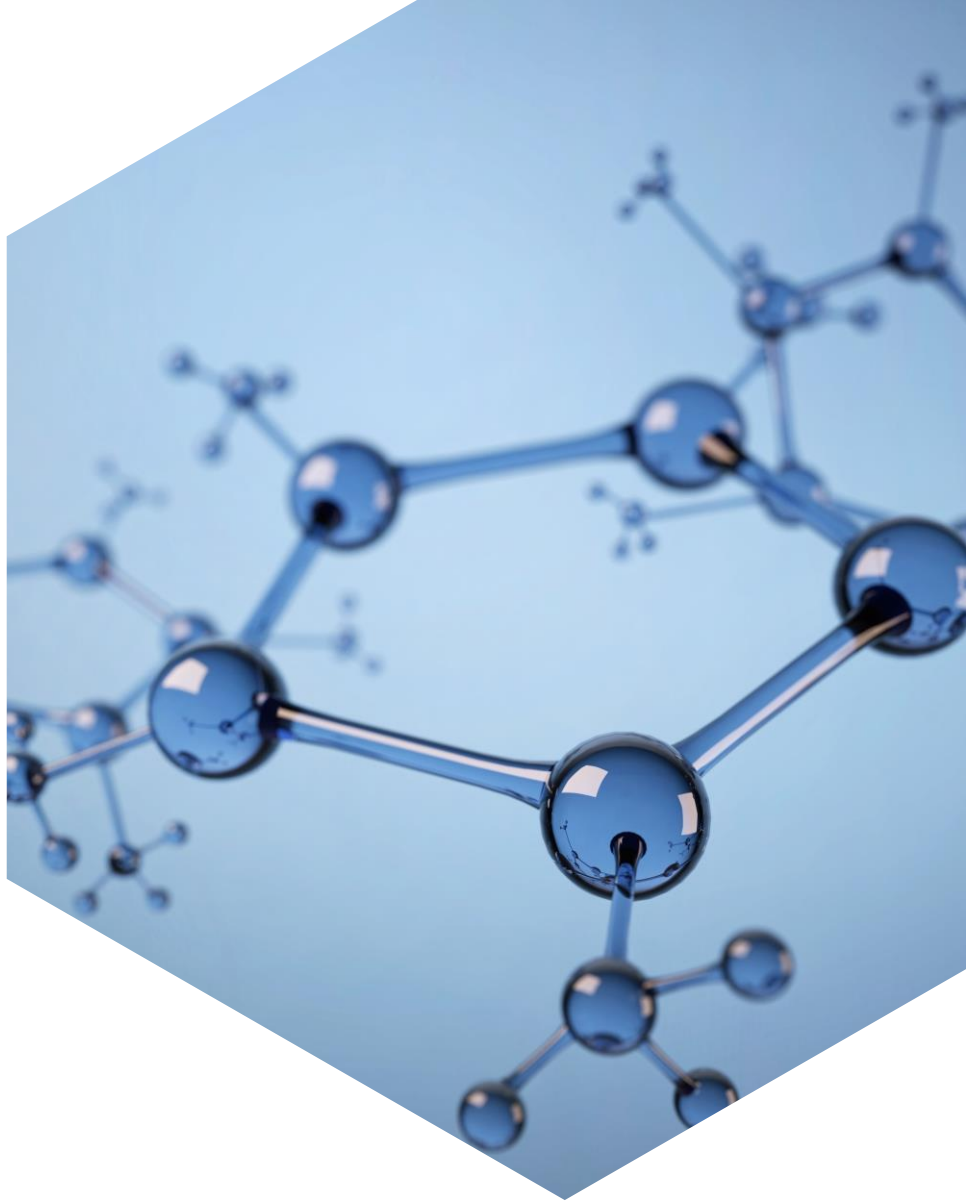
Data standardization at FIZ Karlsruhe

- Validation and standardization of numbers according to the **standardization table**:
    - >100 patent authorities
    - >2.200 numbering formats for patent and application numbers

- **Data harmonization is essential for many processes:**
    - Deduplication
    - Patent family sorting
    - Transfer between databases

AGENDA

- Full-text patent data on CAS STNext
  - Data standardization and normalization
- Best practices for searching in full-text files
  - Key Terms
  - Claims search fields
  - Numeric Property Searching



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Challenges of searching in full-text

- Large amounts of text make it difficult to isolate specific relevant details – lots of noise
- Full-text documents lack standardized indexing, which, making searching less efficient
- Inconsistency/ambiguity about chemical naming or information in primary source documents
- Needle-in-a-haystack problem – so much text, with important details appearing in small useful sections

CAS STNext has multiple enhancements and features that make these challenges less intimidating!

Key Terms make full text searching more efficient

- Noun phrases (1-5 words) representing important aspects of a patent:
  - Extracted from English full-text records using linguistic and statistical methods
  - Sorted according to relevance
- Key Terms (/KT) broaden search scope more precisely than Basic Index (/BI) searches
- Analyze documents to find additional concepts as well as not yet perceived aspects
- Key Terms are a good summary, covering the entire text for relevance checking in full-text files, which can sometime be difficult

Use Key Terms to retrieve relevant results from full text

```
=> S L50 AND (CFTR OR CYSTIC FIBROSIS)/KT

TOTAL FOR ALL FILES
L71          541 L50 AND (CFTR OR CYSTIC FIBROSIS)/KT
```

Files with available Key Terms return 541 results.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Key Terms for query building and data analysis

```
=> FIL JPFULL

=> S CHIMER? (1W) ANTIGEN? (1W) RECEPTOR AND KT/FA
   L1   512

=> ANA KT
   L2   ANALYZE L1 1- KT:   8664 TERMS

=> D 1-50

TERM #   # OCC  # DOC  % DOC KT
------ ------- ------ ------ ---------------
     1     161    161  31.45 AMINO ACID SEQUENCE/KT
     2      97     97  18.95 CHIMERIC ANTIGEN RECEPTOR/KT
     3      82     82  16.02 CANCER CELL/KT
     4      81     81  15.82 CELL RECEPTOR/KT
     5      71     71  13.87 TARGET CELL/KT
     6      70     70  13.67 HOST CELL/KT
     7      67     67  13.09 IMMUNE CELL/KT
     8      67     67  13.09 TUMOR CELL/KT
     9      63     63  12.30 TRANSMEMBRANE DOMAIN/KT
    10      62     62  12.11 HEAVY CHAIN VARIABLE REGION/KT
```

Key Terms in an EPFULL record

L41 ANSWER 1 OF 235 EPFULL COPYRIGHT 2022 LNU on STN.

KT

cftr potentiator; **cftr** corrector; **cystic** **fibrosis**; human **cftr** mutation; solid spray dispersion ingredient; spray dry drying condition; dry spray dry dispersion; buchi mini spray dryer; buchi mini spray drier; combined organic phase; **cftr** genetic mutation; selected spray drying apparatus; non-volatile solvent; patient possess; tablet core weight; rotary drum drying; biconical vacuum; tablet dispersion; augmenting activity; **cftr** inducing; cf associated gene exhibit; bulking agent; dry spray process parameter; wetting agent; solid dispersion range; anti-inflammatory agent; mucolytic agent;. . .

PI **EP 3925607** A4 **20211222**

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Databases with claim search fields

Full text files

AUPATFULL
CANPATFULL
CNFULL
DEFULL
EPFULL
FRFULL
GBFULL
IFIALL
IFICLS
INFULL
JPFULL
KRFULL
PCTFULL
USPATFULL
USPATOLD
USPAT2

Patent family databases

CAplus – Claims for the basic patent
WPINDEX / WPIDS / WPIX – All claims for CN and KR, main claim for 12 authorities

Further files

USGENE - Main claim searchable; all claims can be displayed

Full text files: Stay within one claim with (S) operator

AN 2017007542 CANPATFULL
...
CLMEN **1.** Method for production of a thermoelectric component or at least a
...
2. Method according to Claim 1, characterized in that a plurality of substrates with inserted green bodies are combined to form a stack, wherein the substrates extend within the stack in parallel to one another and in each case a substantially planar separation plate is laid between two substrates that are adjacent within the stack, which separation plate extends in parallel to the substrates and which produces an electrically conductive and force-transmitting connection between the green bodies of the adjacent substrates, and in that the entire stack is arranged between the two electrodes.
3. Method according to Claim 1 or 2, characterized in that a plurality of substrates having inserted green bodies are arranged individually or stacked in a plane between the two electrodes.
4. Method according to Claim 1 or according to one of Claims 2 to 3, wherein the electrodes and/or the separation plates consist of graphite, characterized in that the green bodies, for contacting, are exposed to a
...

(S)

(S)

(S)

(S)

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

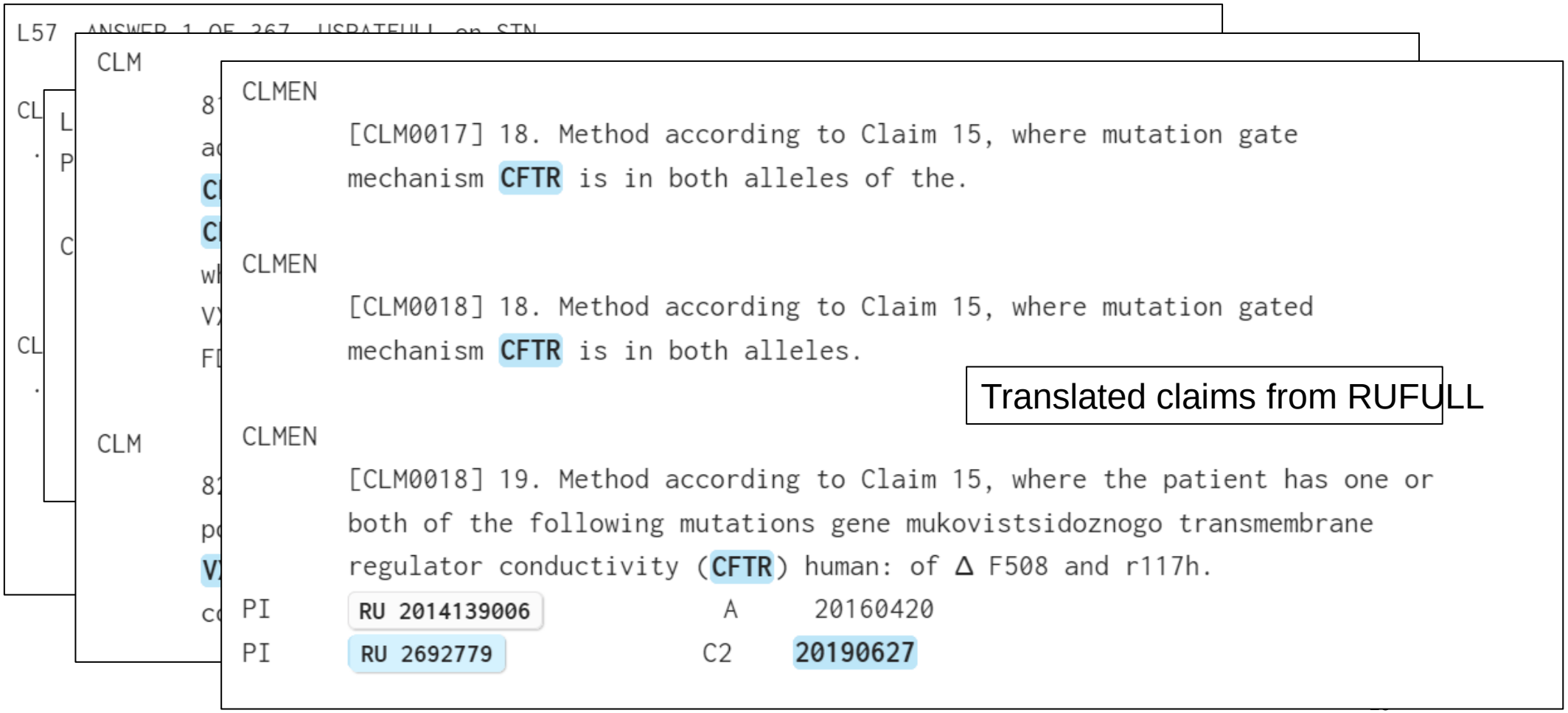
Search within claim text fields in full text files to find specific information within national publication claims

```
=> S L71 AND (CFTR? OR (CYSTIC? (3A) ?MEMBRAN?) OR IVACAFTOR OR KALYDECO OR VX770 OR VX
770)/CLM,CLM.EX,CLMEN

TOTAL FOR ALL FILES
L92        435 L71 AND (CFTR? OR (CYSTIC? (3A) ?MEMBRAN?) OR IVACAFTOR OR KALYD
               ECO OR VX770 OR VX 770)/CLM,CLM.EX,CLMEN
To see the field codes for search terms in an L-number, enter "DISPLAY
QUERY" followed by the L-number at an arrow prompt.  To see the field
codes for search terms in a saved query, enter "ACTIVATE" and the
query name, followed by '/Q' at an arrow prompt.
```

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Examples of relevant claim text from full text files



Translated claims from RUFULL

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Reload of patent fulltext files

JPFULL, CNFULL, AUPATFULL, CANPATFULL, DEFULL, GBFULL, FRFULL

Content expansion:

- **JPFULL**: 12 million additional family records (1913-1999) added; 19.55 million family records in total
- **CNFULL:** 3 million additional design patents; > 45 million family records in total

New search options:

- New search fields 1 and display options
- The Locarno classification (/LCL) is available for design patents.
- **Numeric property search** now available, see HELP NPS for details
- **Interactive Claims Viewer** integration for JPFULL
- For JPFULL also the FI classification (/FCL) and the F-term classification (/FTERM) are available for Japanese patents and utility models back to 1960.

1

/CLM.IC	independent claims
/CLM.CG	claim group
/DETDEN	detailed description in English
/PAS and /PAN	standardized/normalized patent assignees
/KT	Key Terms
/UO and /UOS	ultimate owner /standardized

More info in Database summary sheet
or HELP CHANGE

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Claim Group allows for more relevant claim searching

Available in several patent fulltext files

- Available in several patent fulltext files, among them PCTFULL, EPFULL + previously mentioned files
- Search index comprised of claims belonging to the same group of an independent claim and its dependent claims
- Adds comprehensiveness and precision to your search results: Less noise compared to a search in all claims, more relevant results compared to a limited search in just a single claim
- Check the availability via **CLM.CG/FA**
- Claim Group search field: **/CLM.CG**

Capture patents that do not have claim text available in CAplus and search their claims in full text files

```
=> S L2 AND P/DT NOT CLM/FA
      15904610 P/DT
       8007165 CLM/FA
L7          709 L2 AND P/DT NOT CLM/FA

=> FILE JPFUL KRFUL EPFUL DEFUL FRFUL PCTFUL
=> TRANSFER L7 1- PN.B /PN
L8           TRANSFER L7 1- PN.B :     706 TERMS
L9           71 FILE JPFULL
L10         128 FILE KRFULL
L11          67 FILE EPFULL
L12          51 FILE DEFULL
L13          31 FILE FRFULL
L14          27 FILE PCTFULL

TOTAL FOR ALL FILES
L15         375 L8
L16             QUE  TERMS FROM L8 WITH NO HITS:   358 TERMS
```

Review those records for which no patent claim text is available
(709 out of 8419)

Transfer those patent numbers over to a selection of the patent full text files

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Numeric Property Searching can be used to refine answer sets

- Numeric values play a key role in patents which use chemical and physical properties to set the scope of patent protection
  - Exact values are typically expressed in broad ranges
- Searching numeric data in patents is challenging
  - Standard text searching tools are too limited
  - Great variety of properties and units and a great variation of representation
- Full text and some value-added files allow searching for numeric data extracted from patents

Extracting, normalizing and indexing numeric data

- Numbers and their units within the English-language text are identified
    - About 1,800 property unit variants are identified
    - Numbers are considered from exact values, closed ranges and open ranges
    - Numerals are detected as well
- Identified original data are normalized to base units and indexed for searching
- 59 numeric property search fields with their respective base units are available
- Additional units (CGS-units, imperial and US customary units) are converted to base units

STNext databases that offer Numeric Property Search

```
=> D CLUSTER NPS

CLUSTER NAME      CLUSTER DEFINITION
----------------- ----------------------------------------------------
NPS               1MOBILITY 2MOBILITY AGRICOLA ANABSTR AUPATFULL CABA
                  CANPATFULL CEABA CNFULL COMPENDEX DEFULL ENCOMPPAT
                  ENCOMPPAT2 EPFULL FRFULL FSTA GBFULL INFULL INSPEC
                  JPFULL KRFULL NTIS PCTFULL PQSCITECH RAPRA
                  RDISCLOSURE RUFULL TEMA TULSA TULSA2 USPATFULL USPAT2
                  WPIDS WPINDEX WPIX
                  Numeric Property Searching
```

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Use NPS feature in full-text files to retrieve specific information for dosage amounts

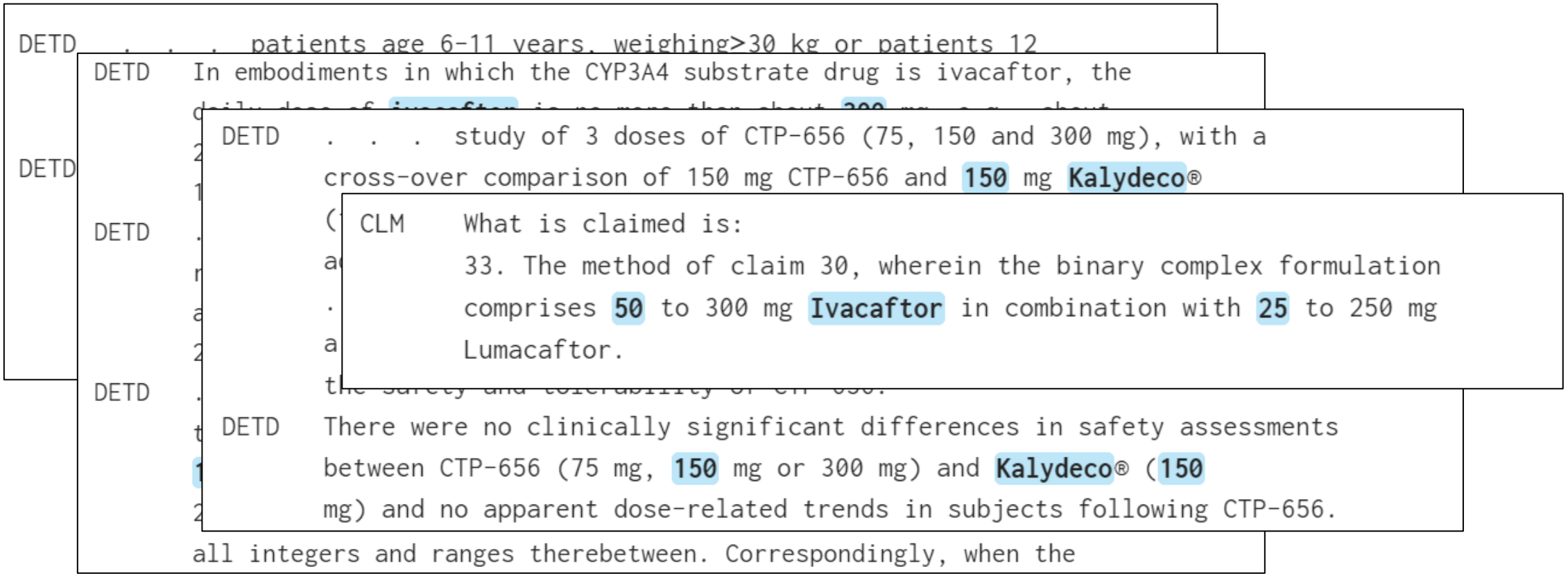
```
=> S L50 AND (IVACAFTOR OR KALYDECO OR VX770 OR VX 770) (5A) 125-175 MG/M

    TOTAL FOR ALL FILES
    L113        108 L50 AND (IVACAFTOR OR KALYDECO OR VX770 OR VX 770) (5A) 125-175
                    MG/M
```

For additional training materials on Numeric Property Searching syntax and best practices, check our training website

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Examples of hit records from full text NPS search

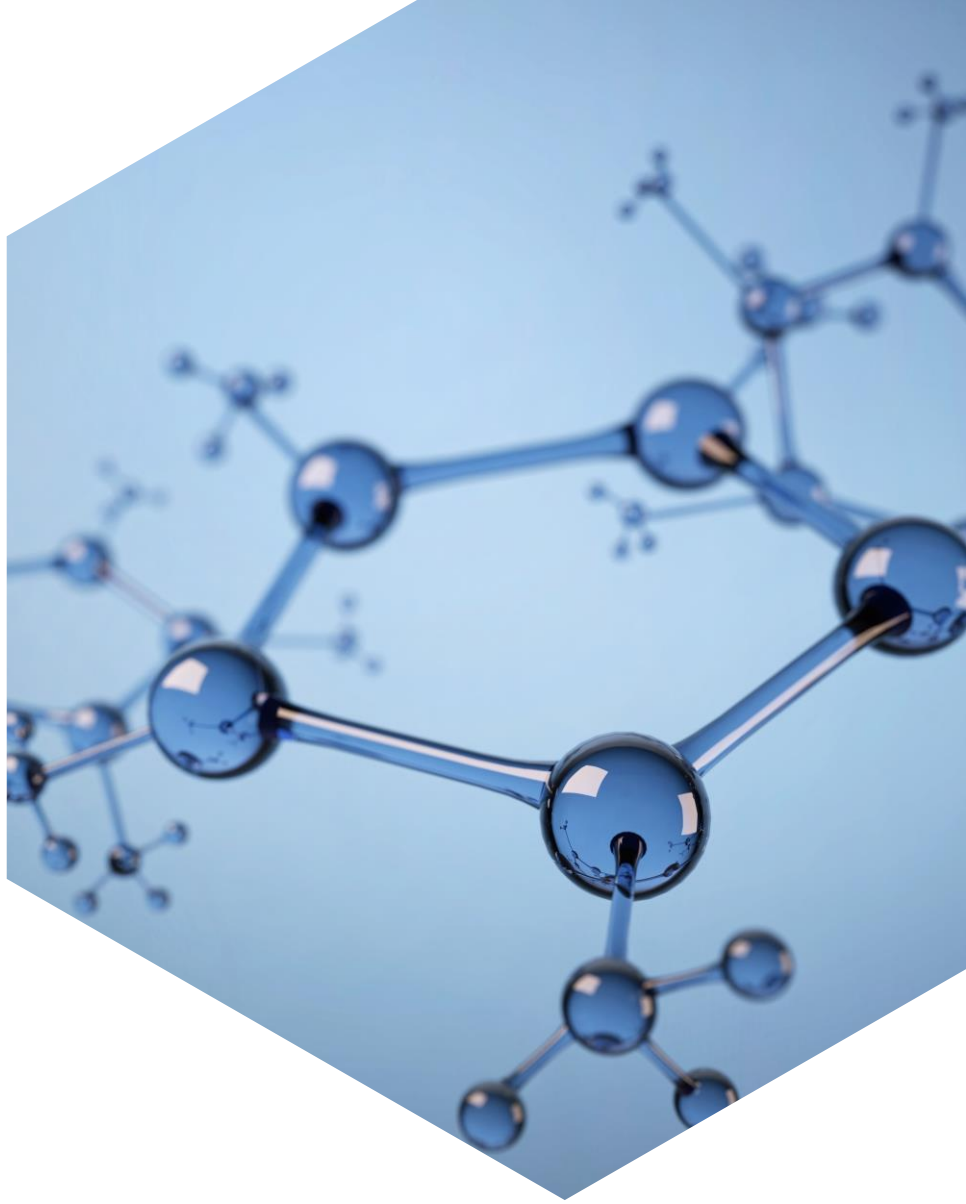


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Summary

- CAS STNext contains a vast amount of patent data in full-text, primary source format
- Searching in full-text files has been historically challenging due to lack of indexing and noise from parts of the patent less relevant to searches – background, etc.
- CAS STNext includes many features that simplify and improve effectiveness when searching this data
  - Key Terms, Claim text fields, numeric property searching, and more!



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Thank you

Sarah W. Stokes
Senior Product Specialist, IP
sstokes@cas.org

CONTACT

CAS
help@cas.org
cas.org

EMEA Help
EMEAhelp@cas.org

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society