The TLS215_CITN_CATEG Table

In this notebook, we analyse the `TLS215_CITN_CATEG` table. The `TLS215_CITN_CATEG` table in PATSTAT provides detailed information about the categories assigned to citations during patent searches. These categories are crucial for assessing the relevance and potential impact of cited documents on patent applications.

Keys and Constraints

- **Primary Key**: The primary key is a composite of `PAT_PUBLN_ID`, `CITN_REPLENISHED`, `CITN_ID`, `CITN_CATEG`, and RELEVANT_CLAIM, ensuring each record is unique and traceable.
- **Foreign Key**: PAT_PUBLN_ID, CITN_REPLENISHED, and CITN_ID serve as a foreign key referencing the TLS212_CITATION table. This relationship links each citation category back to its original citation record in TLS212.

Here's an overview of each attribute in the table:

- `PAT_PUBLN_ID` : Identifies the specific patent publication associated with the citation.
- `CITN_REPLENISHED` : A flag indicating whether the citation has been replenished or updated.
- `CITN_ID` : A unique sequence number for each citation made by a particular document.
- `CITN_CATEG` : The citation category assigned in the search report, typically denoted by codes.
- `RELEVANT_CLAIM` : Specifies the claim number(s) in the patent application to which the citation applies, showing a direct link between the cited document and particular claims of the patent.

Usage and Interpretation

In the patent examination process, these categories help determine:

1. **Novelty and Inventive Step**: Categories like X and I are crucial for identifying prior art that may directly affect the novelty or inventive step.
2. **Contextual Relevance**: Categories such as A, L, and D provide background information or procedural context without impacting patentability directly.

This classification system provides insights into the depth and type of impact cited documents have on a patent's claims, contributing significantly to the patentability analysis conducted by examiners.

In [2]:
```
from epo.tipdata.patstat import PatstatClient
from epo.tipdata.patstat.database.models import (
    TLS201_APPLN,
    TLS212_CITATION,
    TLS211_PAT_PUBLN,
    TLS214_NPL_PUBLN,
    TLS215_CITN_CATEG
)
from sqlalchemy import and_, case, func, select

# Initialise the PATSTAT client
patstat = PatstatClient(env="TEST")

# Access ORM
db = patstat.orm()
```

Key Fields in the TLS215_CITN_CATEG Table

PAT_PUBLN_ID

Identifies the specific patent publication associated with the citation.This field represents the patent **publication making** the citation, in other words it refers to the **citing** publication. It identifies the patent document that includes a citation to another patent or non-patent literature. `PAT_PUBLN_ID` identifies a specific patent publication. Each patent publication has a unique `PAT_PUBLN_ID` in the `TLS215_CITN_CATEG` table, which is used in `TLS212_CITATION` to associate citations with that publication.

CITN_REPLENISHED

A flag indicating whether the citation has been replenished or updated. The `CITN_REPLENISHED` attribute in the PATSTAT database refers to a special type of citation that is "replenished" or copied from one patent publication to another, for example in the context of European Patent Office (EPO) and international (PCT) applications. It is meant to fill in the citation information that might be missing from European publications but is present in the corresponding international publication. When a European patent application (Euro-PCT) is based on an international (PCT) application, the EPO typically does not repeat the citations from the international search report. However, these citations are still relevant to understanding the Euro-PCT application, so PATSTAT "replenishes" the citation list by adding citations from the corresponding international PCT application. It applies to any patent publication where citations are carried over from previous publications, whether they are from the same authority or from an international stage (such as PCT).

CITN_ID

A unique sequence number for each citation made by a particular document. It ensures that each citation, whether it's a patent citation or non-patent literature (NPL) citation, is uniquely identifiable within the context of a single publication. This ID allows multiple citations to be associated with one publication, avoiding duplication and keeping the records organized. The `CITN_ID` is assigned sequentially to distinguish each citation listed in a patent publication, starting from 0. For example, the first citation in a document gets `CITN_ID = 1`, the second gets `CITN_ID = 2`, and so on.

The number is purely a running number and has no special meaning beyond helping to distinguish citations within that particular citing publication.

CITN_CATEG

The `CITN_CATEG` field in the `TLS215_CITN_CATEG` table represents the **categories assigned to patent citations in official search reports**, mainly by search examiners. Each category indicates the relevance of the cited document to the claims of the patent application being examined, as defined by the search authority. This field provides critical information about how cited documents impact the novelty, inventive step, or general background of the application.. Each category signifies a level of relevance and impact on the patent application's claims:

The `TLS215_CITN_CATEG` table captures essential citation categories from patent search reports. These categories, defined by **Annex XIV** of the **DOCDB**, indicate how cited documents impact the patentability of claims. Here is an overview of each category and its role in the patent examination process:

- X: Indicates that the cited document alone challenges the novelty or inventive step of the claimed invention, possibly rendering it unpatentable.
- Y: Suggests that the cited document, when combined with others, may impact the inventive step or novelty.
- A: Refers to documents that provide general background or technical information relevant to the patent but do not directly challenge its claims.
- O: Non-written disclosures, like oral or display presentations.
- P: Intermediate documents with relevant information.
- T: Theory or principle underlying the invention.
- E: Earlier patent applications published after the filing date, often with potentially conflicting claims.
- D: Documents cited within the application itself.
- L: Documents cited for additional reasons not directly affecting patentability.
- &: Family-related documents, where the citation belongs to the same patent family.

Special Cases and Additional Categories

- I: Introduced for publications after April 2011, this category is a refinement of X specifically for prejudicing inventive step.
- R: Found in Chinese publications, it refers to applications or utility models filed on the same day related to the same invention.

In patent data, citations are often categorized to indicate the importance or relevance of a document to a patent's claims. However, not all citation data is structured in the same way.

1. ***Structured Citations ("Rich" Citations)***:

In structured citations, each citation category (e.g., **X, Y, A**) is assigned individually to specific claims within the patent. This rich structure allows for precise associations between a citation category and the claim it affects, and is often used in regions or databases where high granularity is maintained in the citation data. For instance, X might apply to claims 1-3, Y to claims 4-5, etc.

2. ***Unstructured or "Poor" Citations***:

In "poor" citations, categories are combined in a single field as a string (e.g., YAX, XPI, etc.), with all relevant categories listed together. This structure lacks the finer detail of indicating specific claims for each category and is common when data sources aggregate citation information to simplify or where citation data is provided in a less detailed format.

For example, a citation labeled as YAX implies that the cited document is relevant for different reasons: Y: Relevant in combination with other documents. A: Defines the state of the art. X: Highly relevant, potentially prejudicing novelty on its own.

RELEVANT_CLAIM

Specifies the claim number(s) in the patent application to which the citation applies, showing a direct link between the cited document and particular claims of the patent.

In [ ]:
```
sample_citn_records = (
    db.query(
        TLS215_CITN_CATEG.pat_publn_id,
        TLS215_CITN_CATEG.citn_replenished,
        TLS215_CITN_CATEG.citn_id,
        TLS215_CITN_CATEG.citn_categ,
        TLS215_CITN_CATEG.relevant_claim
    )
    .order_by(TLS215_CITN_CATEG.pat_publn_id, TLS215_CITN_CATEG.c
itn_id)
)

sample_citn_res = patstat.df(sample_citn_records)

sample_citn_res
```

In [4]:
```
category_counts = (
    db.query(
        TLS215_CITN_CATEG.citn_categ,
        func.count(TLS215_CITN_CATEG.citn_categ).label("count")
    )
    .group_by(TLS215_CITN_CATEG.citn_categ)
    .order_by(func.count(TLS215_CITN_CATEG.citn_categ).label("cou
nt"))
)


category_counts_df = patstat.df(category_counts)

category_counts_df
```

Out[4]:

	citn_categ	count
0	None	0
1	YAX	1
2	AX	1
3	DXA	1
4	PXA	1
...	...	...
85	R	56195
86	I	169383
87	X	743307
88	Y	1101373
89	A	5026355

90 rows × 2 columns

In [5]:
```
rich_citations_query = (
    db.query(
        TLS215_CITN_CATEG.pat_publn_id,
        TLS215_CITN_CATEG.citn_replenished,
        TLS215_CITN_CATEG.citn_id,
        TLS215_CITN_CATEG.citn_categ,
        TLS215_CITN_CATEG.relevant_claim
    )
    .filter(
        TLS215_CITN_CATEG.citn_categ.notilike('%[^A-Z]%')  # Assu
ming rich categories are single letters.
    )
    .order_by(TLS215_CITN_CATEG.pat_publn_id)
)

rich_citations_res = patstat.df(rich_citations_query)
rich_citations_res
```

Out[5]:

	pat_publn_id	citn_replenished	citn_id	citn_categ	relevant_claim
0	1043	0	1	Y	6
1	1043	0	4	A	1
2	1043	0	2	Y	13
3	1043	0	2	Y	5
4	1043	0	1	Y	7
...	...	...	...	...	...
7243111	606449486	0	1	X	2
7243112	606449486	0	3	X	10
7243113	606449486	0	1	I	7
7243114	606449486	0	5	X	10
7243115	606449486	0	4	X	2

7243116 rows × 5 columns

In [12]:
```
poor_citations_query = (
    db.query(
        TLS215_CITN_CATEG.pat_publn_id,
        TLS215_CITN_CATEG.citn_replenished,
        TLS215_CITN_CATEG.citn_id,
        TLS215_CITN_CATEG.citn_categ,
        TLS215_CITN_CATEG.relevant_claim
    )
    .filter(
        func.length(TLS215_CITN_CATEG.citn_categ) > 1  # Only sel
ects categories with more than one character.
    )
    .order_by(TLS215_CITN_CATEG.citn_categ)
)

poor_citations_res = patstat.df(poor_citations_query)
poor_citations_res
```

Out[12]:

	pat_publn_id	citn_replenished	citn_id	citn_categ	relevant_claim
0	595857354	571823806	4	&P	2
1	595857354	571823806	4	&P	5
2	571823806	0	4	&P	6
3	595857354	571823806	4	&P	7
4	595857354	571823806	4	&P	8
...	...	...	...	...	...
124182	519582558	496043128	1	px	4
124183	496043128	0	1	px	1
124184	519582558	496043128	1	px	1
124185	519582558	496043128	1	px	2
124186	496043128	0	1	px	3

124187 rows × 5 columns