Leveraging On-Site Citation and Family Functionality

Introduction

With the deployment of the cumulative patch `alexandria-sql-patch-alpa-3636-20191101`, on-site CLAIMS Direct installations now have the ability to utilize family and citation functionality in-house. What was previously only available using the remote CLAIMS Direct shared API is now possible internally with simple SQL functions. The following post will outline the steps required to prepare the data tables as well as presenting a brief walk-through of the functionality.

Preparation

Note: If your initial load occurred after May 13, 2020, these tables were included in the initial load and you do not need to follow the preparation instructions here. However, to maintain the supplemental tables required for on-site family and citation functionality, you must add `--api,` e.g., `CD_OTHER_OPTIONS='--api'` to the `apgupd` configuration file.

Sequential Loading

In order to leverage the new functionality, the necessary lookup tables need to be populated on-site. The functions available to achieve this are:

- `cdws.f_update_cited_documents` – serving forward and backward citations
- `cdws.f_update_priority_documents` – serving simple and extended family

Each function takes as input parameter a modified-load-id. Therefore, to populate each table, each modified-load-id from the `xml.t_patent_document_values` table needs to be processed. The following pseudo-SQL will serve as an example.

```
DECLARE v_load_id integer;
BEGIN
  FOR v_load_id IN
      SELECT modified_load_id
      FROM xml.t_patent_document_values
      GROUP BY modified_load_id
      ORDER by modified_load_id LOOP

    PERFORM cdws.f_update_cited_documents( v_load_id );
    PERFORM cdws.f_update_priority_documents( v_load_id );

  END LOOP;
END;
```

Although the above SQL exemplifies the fundamental logic involved, it isn't the most efficient method. To that end, we are offering a supplemental patch that installs bulk loading functions to populate these tables: `alexandria-sql-patch-alpa-3636-x-20191215`. In addition to adding bulk loading functions, this patch expands some of the citation and family functionality not included in the original `alexandria-sql-patch-alpa-3636-20191101` patch, namely:

- FUNCTION `cdws.f_family_citations_backward(v_family_id integer)`
  Returns backward citations for an entire family
- FUNCTION `cdws.f_family_citations_forward(v_family_id integer)`
  Returns forward citations for an entire family
- FUNCTION `cdws.f_family_meta(integer)`
  Returns bibliographic data for an entire family

For a full enhancement listing, please see the README included in the package.

Bulk Loading

After the supplemental patch is installed, 2 new functions for bulk loading are exposed.

- FUNCTION `cdws.f_update_cited_documents_bf()`

- FUNCTION `cdws.f_update_priority_documents_bf()`

These new functions bulk-load the auxiliary tables and are called simply as:

```
SELECT cdws.f_update_cited_documents_bf();
SELECT cdws.f_update_priority_documents_bf();
```

Each function requires at least 24 hours to complete and may take up to 7 days, depending on the environment. They can be run in parallel if desired. Because of the duration required to complete the functions, it's important that the SQL runs uninterrupted by server restarts or terminal disconnects.

Important

In order to expedite the completion of these functions, it is **required** that `apgupd` be paused. Optionally upon completion, executing an `ANALYZE` on the auxiliary tables is recommended:

```
ANALYZE VERBOSE cdws.t_cited_documents;
ANALYZE VERBOSE cdws.t_applications;
ANALYZE VERBOSE cdws.t_priority_documents;
```

Critical

In order to automatically populate these auxiliary tables every new load-id, you must use the `--api` flag to `apgupd`. If this parameter isn't available with your version of `Alexandria-Client-Tools`, you will need to update to the latest version. In versions 2.5.3 and above, this parameter is set in `/etc/alexandria/apgupd.conf`, as a `CD_OTHER_OPTIONS`. Please see Client Tools documentation for configuration details.

To confirm that the citation and family functionality have completed loading, locate the log files (`citation.log` and `family.log`). The last line of each log should show an entry in the following format:

```
INFO:  added N new citations records from the work table (time)
INFO:  added N new priority records from the work table (time)
```

To check whether the calculations are still running, use one of the two SQL queries below while in the database or using pgAdmin.

```
SELECT datname, query FROM pg_stat_activity;
SELECT * FROM pg_stat_activity;
```

Usage

Once the auxiliary table loading is complete, a wide variety of citation and family functionality is available directly from the on-site CLAIMS Direct instance.

Citations

Bringing the CLAIMS Direct API methods `citations/forward` and `citations/backward` on-site is achieved using the new functions:

- FUNCTION `cdws.f_citations_backward`
- FUNCTION `cdws.f_citations_forward`

Each function takes as input a `publication_id` and returns 2 columns: `ucid` and `source`.

```
SELECT * FROM cdws.f_citations_backward( xml.f_ucid2id('US-5551212-A') );
     ucid      | source
---------------+--------
 US-4905451-A  | SEA
 US-2043419-A  | SEA
 US-4631900-A  | SEA
 DE-8002686-U1 | SEA
 EP-0225665-A1 | SEA
 US-3187480-A  | SEA
 US-4828110-A  | SEA
 DE-3138439-A1 | SEA
 EP-0061805-A1 | SEA
 DE-3606826-A1 | SEA
 US-4546875-A  | SEA
 US-4534151-A  | SEA
 EP-0313721-A2 | SEA
```

Of course, `GROUP BY` and `ORDER BY` are available, e.g., using `substring`, group forward citations by country:

```
SELECT substring(ucid, 1, 2) AS country, count(*) AS ncites
FROM cdws.f_citations_forward( xml.f_ucid2id('US-5000000-A') )
GROUP BY country
ORDER BY ncites DESC;
 country | ncites
---------+--------
 US      |    264
 WO      |     31
 EP      |     27
 AU      |      1
 KR      |      1
```

Family

Bringing CLAIMS Direct API methods `family/simple` and `family/extended` functionality to the on-site install is accomplished with the functions:

- FUNCTION `cdws.f_family_simple`
- FUNCTION `cdws.f_family_extended`

Each of these functions, as with the citation functions above, take a `publication_id` as input. The return columns are `ucid` and `family-id`. The extended function returns additional `publication_id` and `level` columns.

```
SELECT * FROM cdws.f_family_simple( xml.f_ucid2id('US-5000000-A') );
       ucid       | family_id
------------------+-----------
 AR-242634-A1     |  26932266
 CA-1335430-C     |  26932266
 EP-0431047-A1    |  26932266
 HU-T60328-A      |  26932266
 JP-H05502366-A   |  26932266
 KR-900702042-A   |  26932266
 WO-1990002193-A1 |  26932266

SELECT * FROM cdws.f_family_extended( xml.f_ucid2id('US-5551212-A') );
 publication_id |      ucid      | level
----------------+----------------+-------
       71329085 | US-5551212-A   |     0
         302369 | AT-96388-T     |     1
        1220699 | AU-636238-B2   |     1
        2293384 | AU-8174591-A   |     1
        5137694 | CA-2050271-A1  |     1
        5137693 | CA-2050271-C   |     1
       14883151 | DE-4027762-C1  |     1
       17156127 | DE-59100535-D1 |     1
       93003830 | DK-0477487-T3  |     1
       21052692 | EP-0477487-A1  |     1
       21052691 | EP-0477487-B1  |     1
       25818932 | ES-2046827-T3  |     1
      128970015 | JP-H054613-A   |     1
      131336800 | JP-H0825524-B2 |     1
       59117993 | NO-913030-D0   |     1
      177975213 | NO-913030-L    |     1
       62462393 | TR-25401-A     |     1
      159577685 | ZA-9106183-B   |     1
```

A second optional argument in the `cdws.f_family_extended` function, `v_include_apps Boolean`, can be used to include application filing data from the patent office. The default setting of this argument is `false` in order to maintain consistency with the EPO. In some cases, however, application filing data is not included in the priority claims data. Setting this argument to `true` will include the application filing data, as in the following example.

```
SELECT ucid
FROM cdws.f_family_extended(xml.f_ucid2id('AU-2012201497-B2'), true ) as f
WHERE f.ucid = 'WO-2008136115-A1';
       ucid
------------------
WO-2008136115-A1
(1 row)
```

Additional Functions

In addition to the above, there are other functions in the supplemental patch worth exploring, namely:

- FUNCTION `cdws.f_family_meta( v_family_id integer )`
  Returns a variety of bibliographic and legal status information
- FUNCTION `cdws.f_family_citations_backward(v_family_id integer)`
  Returns backward citations for an entire family
- FUNCTION `cdws.f_family_citations_forward(v_family_id integer)`
  Returns forward citations for an entire family

Family Meta

The `cdws.f_family_meta` function brings a variety of information together for a simple family. The following columns are returned:

- family-id - DOCDB-assigned simple family integer identifier
- ucid – unique character identifier (publication)
- published – publication date
- anucid – unique character identifier (application)
- filed – application filing date
- title – title of document
- ifi_status – IFI Snapshot document status
- ifi_anticipated_expiration - IFI anticipated expiration date

- ifi_adjusted_expiration - IFI adjusted expiration date
- epo_legal_status_xml – XML representation of EPO legal status

An abbreviated example:

```
SELECT ucid, published, anucid, filed, title FROM cdws.f_family_meta( 26932266 );
       ucid       | published |     anucid     |  filed   |
title
------------------+-----------+----------------+----------
+-----------------------------------------------------------------------------------------
 AR-242634-A1     |  19930430 | AR-31479889-A  | 19890830 | ETHANOL PRODUCTION BY GENETICALLY ENGINEERED
ESCHERICHIA COLI STRAINS
 CA-1335430-C     |  19950502 | CA-609829-A    | 19890830 | ETHANOL PRODUCTION BY ENGINEERED MICROBES
 EP-0431047-A1    |  19910612 | EP-89909966-A  | 19890830 | ETHANOL PRODUCTION BY GENETICALLY ENGINEERED
ESCHERICHIA COLI STRAINS
 HU-T60328-A      |  19920828 | HU-577189-A    | 19890830 | PROCESS FOR PRODUCING ETHANOL WITH TRANSFORMED
MICROORGANISMS
 KR-900702042-A   |  19901205 | KR-900700901-A | 19890830 |
 US-5000000-A     |  19910319 | US-35206289-A  | 19890830 | Ethanol production by Escherichia coli strains co-
expressing Zymomonas PDC and ADH genes
 WO-1990002193-A1 |  19900308 | US-8903753-W   | 19890830 | ETHANOL PRODUCTION BY GENETICALLY ENGINEERED
ESCHERICHIA COLI STRAINS
```

Family Citations

Forward and backward citations for an entire family are also available directly inside CLAIMS Direct.

```
SELECT * FROM cdws.f_family_citations_backward( 26932266 );
       ucid
------------------
 EP-0047641-A2
 WO-1986001833-A1
 WO-1986004357-A1
```

Again, `GROUP BY` and `ORDER BY` are available.

```
SELECT substring(ucid, 1, 2) AS country, count(*) AS ncites
FROM cdws.f_family_citations_forward( 26932266 )
GROUP BY country
ORDER BY ncites DESC;
 country | ncites
---------+--------
 US      |    290
 WO      |     36
 EP      |     35
 JP      |      2
 FR      |      1
 KR      |      1
 AU      |      1
```