**Office Title Normalization**

Problem: Office titles in CBDB are unstructured strings.

Purpose: Code CBDB office title strings.

**Step 1: Ontological structure of office title.**

A full office title should contain four ontological parts: dynasty [D], classification(s) [C], office(s) [O], title [T]. For example: 清朝[D]文職[C]中央[C]吏部[O]文選清吏司[O]主事[T]. The ontological structure of office title is: D+C+O+T, and:

1. C and O can have multiple values;
2. C does not have hierarchy, but O has. That is for C, C1C2=C2C1; but for O, O1O2≠O2O1;
3. The order of DCOT should be retained for building the tree of office title.

**Step 2: Find a collection of elements for each ontological part, build tables accordingly.**

Reference sources for each part:

* D part: use CBDB’s DYNASTIES table.
* C part: UCI’s 明代職官中英辭典 and IHP’s 清代職官資料庫, need more work, e.g., 地方軍事與治安機構類 contains two elements “地方” and “軍事與治安機構”. Absorb CBDB’s OFFICE\_CATEGORIES table.
* O and T parts: UCI’s 明代職官中英辭典 and IHP’s 清代職官資料庫.

Table columns for each part (names for new tables and columns are normalized, i.e., starting with C\_OT, see Table and Column Name Normalization):

D part: c\_dy / c\_dynasty / c\_dynasty\_chn / c\_start / c\_end / c\_sort (existing columns in DYNASTIES table).

C part, table name C\_OT\_CLS:

* c\_ot\_cls\_id: office title, classification ID.
* c\_ot\_cls\_chinm: office title, classification Chinese name.
* c\_ot\_cls\_engnm: office title, classification English name.
* c\_ot\_cls\_desc: office title, classification description.
* c\_ot\_cls\_start: office title, classification starting date.
* c\_ot\_cls\_end: office title, classification ending date.

O part, table name C\_OT\_OFC:

* c\_ot\_ofc\_id: office title, office ID.
* c\_ot\_ofc\_chinm: office title, office Chinese name.
* c\_ot\_ofc\_engnm: office title, office English name.
* c\_ot\_ofc\_desc: office title, office description.
* c\_ot\_ofc\_start: office title, office starting date.
* c\_ot\_ofc\_end: office title, office ending date.

T part, table name C\_OT\_TIT:

* c\_ot\_tit\_id: office title, title ID.
* c\_ot\_tit\_chinm: office title, title Chinese name.
* c\_ot\_tit\_engnm: office title, title English name.
* c\_ot\_tit\_desc: office title, title description.
* c\_ot\_tit\_start: office title, title starting date.
* c\_ot\_tit\_end: office title, title ending date.

**Step 3: Add columns to existing CBDB office title table (OFFICE\_CODES).**

Columns to be added:

* c\_ot\_raw: office title raw values.
  + Data type: string, single value.
* c\_ot\_coding: office title coding values (corrected values).
  + Data type: string, single value.
* c\_ot\_d\_coding: office title dynasty.
  + Data type: integer, can have multiple values.
  + Linked to: DYNASTIES.c\_dy
* c\_ot\_c\_coding: office title classification.
  + Data type: integer, can have multiple values.
  + Linked to: C\_OT\_CLS.c\_ot\_cls\_id
* c\_ot\_o\_coding: office title office.
  + Data type: integer, can have multiple values.
  + Linked to: C\_OT\_OFC.c\_ot\_ofc\_id
* c\_ot\_t\_coding: office title title.
  + Data type: integer, can have multiple values.
  + Linked to: C\_OT\_TIT.c\_ot\_tit\_id

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| c\_ot\_raw | c\_ot\_coding | c\_ot\_d\_coding | c\_ot\_c\_coding | c\_ot\_o\_coding | c\_ot\_t\_coding |
| 廣西司員外郎 | 清朝中央刑部廣西司員外郎 | 20 | 101 | 304\_2003 | 10001 |

Assuming: 清朝 = 20；中央 = 101；刑部 = 304；廣西司 = 2003；員外郎 = 10001

**Step 4: Build office tree.**

Two patterns of search are used in building office tree: flat search, which generates parallel categories from ontological parts (D, C, and T parts); and in-depth search, which generates hierarchical trees from ontological parts (O part).

Example:

清朝[D]武職[C]中央[C] 禁旅八旗[O] 鑾輿衛[O] 指揮使[T]

清朝[D]文職[C]中央[C]吏部[O]文選清吏司[O]主事[T]

清朝[D] 文職[C]地方[C]行政[C]浙江[O]巡撫[T]

清朝[D] 文職[C]地方[C]行政[C] 臺北府分防 [O] 通判[T]

明朝[D]武職[C]中央[C] 禁旅八旗[O] 鑾輿衛[O] 指揮使[T]

明朝[D]武職[C]中央[C] 禁旅八旗[O] 鑾輿衛[O] 指揮使[T]

宋朝[D]武職[C]中央[C] 禁旅八旗[O] 鑾輿衛[O] 指揮使[T]

Build layer 1 Dynasty [D]: Flat search on D, shows: 清朝 / 明朝 / 宋朝

|- 清朝

|- 明朝

|- 宋朝

Build layer 2 Classification [C]: Click 清朝, flat search on C, shows: 文職 / 中央 / 武職

|-- 清朝

|-- 武職

|-- 文職

|-- 中央

|-- 地方

|-- 行政

Build layer 3 Office [O]:

Click 文職, flat search within the rest of O, shows: 中央 / 地方 / 行政

|- 清朝

| |-- 武職

| | |-- 中央

| | | |-- 禁旅八旗

| | | |-- 鑾輿衛

| | | |-- 指揮使

| |-- 文職

| | |-- 中央

| | | |-- 吏部

| | | | |-- 文選清吏司

| | | | | |-- 主事

| | |-- 地方

| | | |-- 行政

| | | | |-- 浙江

| | | | | |-- 巡抚

| | | | |-- 臺北府分防

| | | | | |-- 通判

| | |-- 行政

| | | |-- 地方

| | | | |-- 浙江

| | | | | |-- 巡抚

| | | | |-- 臺北府分防

| | | | | |-- 通判

| |-- 中央

| |-- 地方

| |-- 行政

**Future Work: Table and column name normalization.**

Example:

C\_OT\_CLS\_ID:

* C\_: CBDB
* OT\_: Office title table series.
* CLS\_: Classification
* C\_OT\_CLS: Office title classification table.
* \_ID: column name.