6月—8月工作总结

共进行了10次Meeting，主要包含以下内容：

## 1.制定工作计划，并根据实际情况进行调整

| 时间段 | 任务 |
| --- | --- |
| 06月07日--08月11日 | 1.熟悉CBDB数据表格特点  2.构建繁体简体转换字典  3.对ALTNAME\_DATA进行数据处理，并总结错误类型 4.对KIN\_DATA进行数据处理，并总结错误类型  5.对ENTRY\_DATA进行数据处理，并总结错误类型  6.对CBDB数据统计并进行绘制图表进行可视化 |
| 08月12日--08月31日 | 1.对BIOG\_TEXT\_DATA进行数据处理，并总结错误类型 2.对ASSOC\_DATA进行数据处理，并总结错误类型 3.对其他剩余数据表格进行数据处理，并总结错误类型 |
| 09月01日--10月31日 | 1.总结所有表格的错误规律 2.制定修改方案 3.进行系统设计和交互功能设计  4.系统工程实现 |
| 11月01日--11月30日 | 1.加入大模型，构建新的数据处理工作流  2.制定绘画数据加入CBDB数据库的方案 |

## 2.繁简体转换

* 用OPENCC 将 GBK字符表 转换为简化字
* 对比转之前和转之后的不同的字
* 将这些字生成简化字字典 **sim2trad\_dict\_opencc**
* 对CBDB表中的书名、人名、notes不做检查
* 下一步工作：主要对官职、地点做检查

## 3. ALTNAME\_DATA

### ALTNAME\_DATA Table

| 数据字段 | 错误类型 | 输出 | 处理方案 |
| --- | --- | --- | --- |
| c\_alt\_name | 方框、汉字、全角空格、数字、韩文、[n/a] | ALTNAME\_DATA\_anomaly\_altname\_chn.csv |  |
| c\_alt\_name\_chn | 半角和全角空格、方框、英文A\_Z（大小写）、数字、[n/a] | ALTNAME\_DATA\_anomaly\_altname\_chn |  |

### 新建 ALTNAME\_DATA\_anomaly\_altname\_chn

| 数据字段 | | | | error\_types | 处理方案 |
| --- | --- | --- | --- | --- | --- |
| c\_alt\_name | c\_alt\_name\_chn | c\_notes | c\_pages |  |  |
| [n/a] | [n/a] | 无 |  | Delete | 直接删除 |
| □ | □ | 无 |  |
| Yuyan | [n/a] | 无 | 无 |
| 良仲 | Liangzhong |  |  | Interconvert | 互换，C符合D的规则，D符合C的规则 |
| Ke□ | 克□ |  |  | Somebody | 将方框变成 X某，拼写首字母大写 |
| □Shiqi | □十七 |  |  |
| lei yuan xuan |  |  |  | Blank | 在Interconvert的基础上， C列，两个空格全部删除，字首字尾的空格都删掉  D列，有一个空格删除 |
|  | 臞 山 |  |  |
| yi leyao | Yi Leyao |  |  |  | 人来处理 |

### ALTNAME\_DATA Table

| 数据字段 | | | | | | 错误类型 |
| --- | --- | --- | --- | --- | --- | --- |
| c\_personid | c\_alt\_name | c\_alt\_name\_chn | c\_alt\_name\_type\_code | c\_source | c\_pages | error\_types |
| 相同 | 相同 | 相同 | 0 / X |  |  | Delete |
| 相同 | 相同 | 相同 | 3 / 4 |  |  | Check |
| 相同 | 相同 | 相同 |  | 无 | 无 | Delete |
| 相同 | 相同 | 相同 |  | 有 | 无 | Merge  (Remove c\_pages, and then move c\_source to the first record) |
| 相同 | 相同 | 相同 |  | 无 | 有 |
| 相同 | 相同 | 相同 |  | 有 | 无 | Supplement |
| 相同 | 相同 | 相同 |  | 有 | 有 |

## 4. KIN\_DATA Table

### 新建 KIN\_DATA\_NEW

| c\_personid | c\_personid\_name\_chn | c\_kin\_id | c\_kin\_id\_name\_chn | c\_kin\_code | c\_kin\_id\_female | c\_kinrel\_chn | c\_kinrel |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 93146 | 楊廷玉 | 93663 | 武曌（武則天） | 580 | 1 | 表姑 | K-1(female) |

### 数据 KIN\_DATA\_NEW\_kin\_code\_stat

| c\_kin\_code | c\_kinrel\_chn | c\_kin\_code\_num |
| --- | --- | --- |
| 135 | 妻子 | 23752 |

### 新建 KIN\_DATA\_NEW\_kin\_female\_stat\_true

* + 统计KIN\_DATA\_NEW 中 c\_kin\_code 的类别数量
  + 亲属关系与性别是否吻合

| c\_kin\_code | c\_kinrel\_chn | c\_kin\_code\_num | c\_kin\_code\_num\_female | c\_kin\_code\_num\_male | ratio | c\_kinrel\_female |
| --- | --- | --- | --- | --- | --- | --- |
| 135 | 妻子 | 23751 | 23512 | 239 | 0.989937266 | 1 |

### 新建 KIN\_DATA\_NEW\_error\_type

* + A是B的父亲，则一定存在B是A的女儿 或 儿子，如果没有，则错误类型为缺失
  + A是B的父亲，如出现 B是A的母亲，结合Kin\_ship\_code 表一起做判断，错误类型为矛盾

| c\_personid | c\_personid\_name\_chn | c\_kin\_id | c\_kin\_id\_name\_chn | c\_kin\_code | c\_kin\_id\_female | c\_kinrel\_chn | c\_kinrel | error\_type |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 陳安期 | 3047 | 陳安定 | 125 | 0 | 兄 | B+ | Missing |

## 5. ENTRY\_DATA Table

| c\_entry\_code | c\_pages | c\_notes | error\_types |
| --- | --- | --- | --- |
| 0 | 0 | 0 | EmptyEntry |

| c\_age | | | error\_types |
| --- | --- | --- | --- |
| 小于0 | | | WrongAge |
| 大于100 | | | WrongAge |

| c\_entry\_nh\_year | | | error\_types |
| --- | --- | --- | --- |
| 大于61 | | | HugeNHYear |

| c\_nianhao\_id | | | error\_types |
| --- | --- | --- | --- |
| 0 | | | 不算错误 |
| 是不是在 NIAN\_HAO 这张表里 | | | NoNianHao |

| c\_entry\_code | | | error\_types |
| --- | --- | --- | --- |
| 是不是在ENTRY\_CODES 这张表里 | | | NoEntryCode |

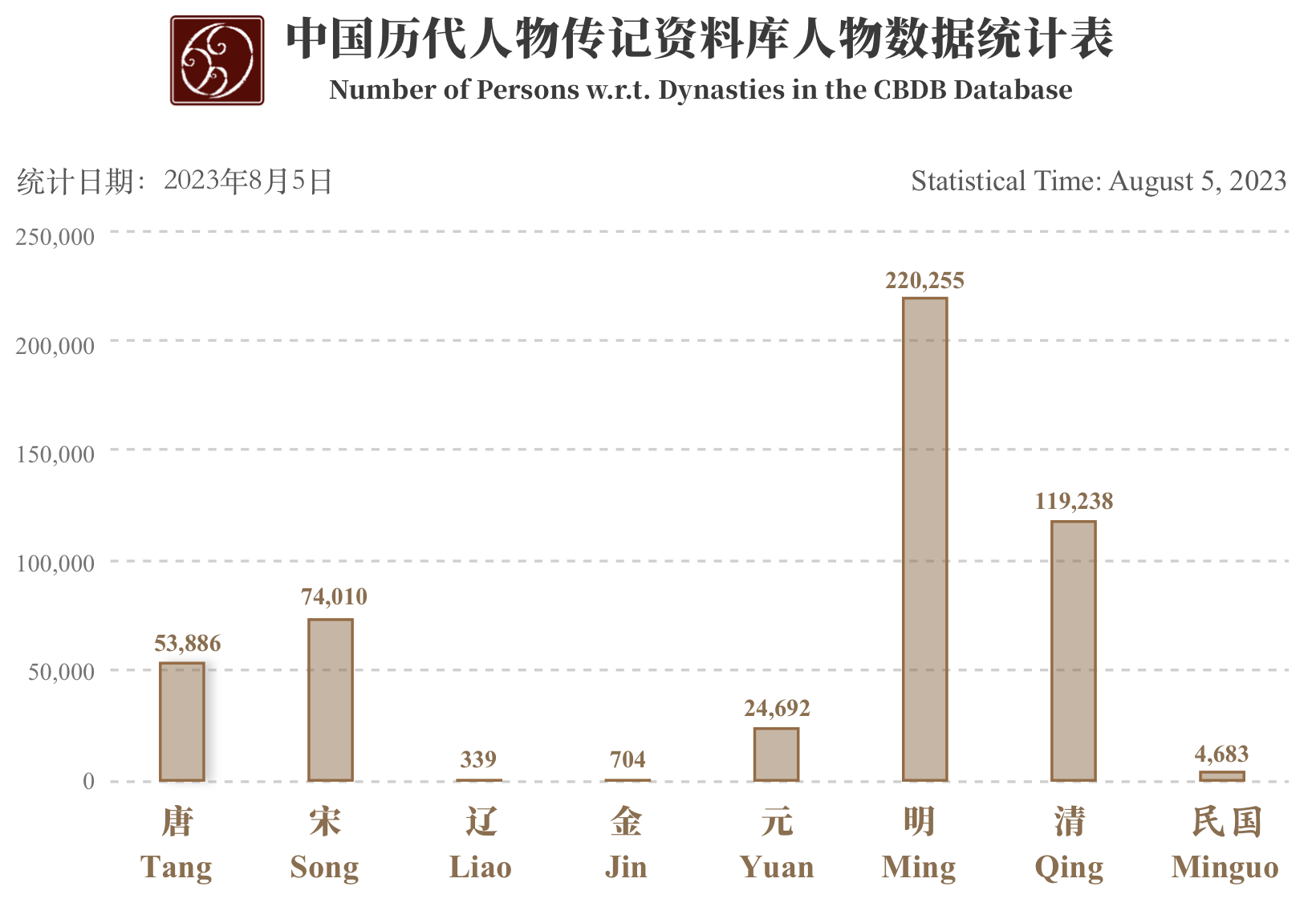
| c\_source | | | error\_types |
| --- | --- | --- | --- |
| 是不是在 TEXT\_CODE 这张表里 | | | NoSource |

| c\_year | c\_nianhao\_id | error\_types |
| --- | --- | --- |
| 1850 | 0 | EmptyNianhao |

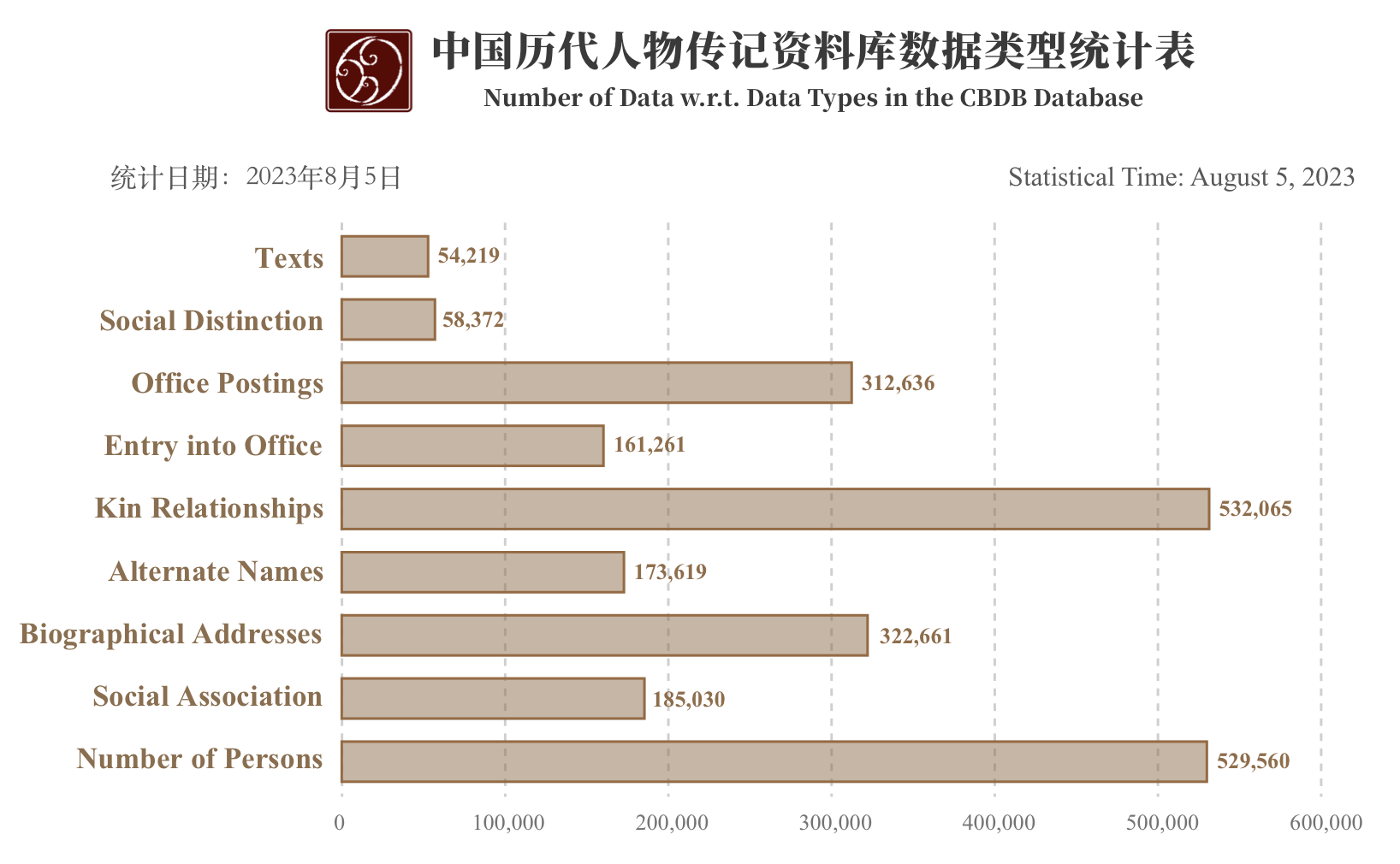
| c\_source | c\_pages | error\_types |
| --- | --- | --- |
| 0 | 769頁 | NoSourceHavePages |

## 6. CBDB数据自动统计程序及可视化

* Person in different dynasties



* Person in different data type



## 7. CBDB 剩余数据表格

| Table\_Name | 繁简 | 中英 | 性别 | 年代 | 角色 | 缺失 | 重复 | 矛盾 | 粒度 | 备注 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ADDRESSES |  |  |  |  |  |  |  |  |  |  |
| ADDR\_BELONGS\_DATA |  |  |  |  |  |  |  |  |  | X |
| ADDR\_CODES |  |  |  |  |  |  |  |  |  |  |
| ADDR\_PLACE\_DATA |  |  |  |  |  |  |  |  |  | X |
| ADDR\_XY |  |  |  |  |  |  |  |  |  | X |
| ALTNAME\_CODES |  |  |  |  |  |  |  |  |  |  |
| ALTNAME\_DATA |  |  |  |  |  |  |  |  |  |  |
| APPOINTMENT\_TYPE\_CODES |  |  |  |  |  |  |  |  |  |  |
| ASSOC\_CODES |  |  |  |  |  |  |  |  |  |  |
| ASSOC\_CODE\_TYPE\_REL |  |  |  |  |  |  |  |  |  | X |
| ASSOC\_DATA |  |  |  |  |  |  |  |  |  |  |
| ASSOC\_TYPES |  |  |  |  |  |  |  |  |  |  |
| ASSUME\_OFFICE\_CODES |  |  |  |  |  |  |  |  |  |  |
| BIOG\_ADDR\_CODES |  |  |  |  |  |  |  |  |  |  |
| BIOG\_ADDR\_DATA |  |  |  |  |  |  |  |  |  |  |
| BIOG\_INST\_CODES |  |  |  |  |  |  |  |  |  |  |
| BIOG\_INST\_DATA |  |  |  |  |  |  |  |  |  |  |
| BIOG\_MAIN |  |  |  |  |  |  |  |  |  |  |
| BIOG\_SOURCE\_DATA |  |  |  |  |  |  |  |  |  | X |
| BIOG\_TEXT\_DATA |  |  |  |  |  |  |  |  |  | X |
| CHORONYM\_CODES |  |  |  |  |  |  |  |  |  |  |
| COPYMISSINGTABLES |  |  |  |  |  |  |  |  |  | X |
| COPYTABLES |  |  |  |  |  |  |  |  |  | X |
| COPYTABLESDEFAULT |  |  |  |  |  |  |  |  |  | X |
| COUNTRY\_CODES |  |  |  |  |  |  |  |  |  | X |
| DATABASE\_LINK\_CODES |  |  |  |  |  |  |  |  |  |  |
| DATABASE\_LINK\_DATA |  |  |  |  |  |  |  |  |  | X |
| DYNASTIES |  |  |  |  |  |  |  |  |  |  |
| ENTRY\_CODE\_TYPE\_REL |  |  |  |  |  |  |  |  |  | X |
| ENTRY\_CODES |  |  |  |  |  |  |  |  |  |  |
| ENTRY\_DATA |  |  |  |  |  |  |  |  |  |  |
| ENTRY\_TYPES |  |  |  |  |  |  |  |  |  |  |
| ETHNICITY\_TRIBE\_CODES |  |  |  |  |  |  |  |  |  |  |
| EVENT\_CODES |  |  |  |  |  |  |  |  |  |  |
| EVENTS\_ADDR |  |  |  |  |  |  |  |  |  | X |
| EVENTS\_DATA |  |  |  |  |  |  |  |  |  |  |
| EXTANT\_CODES |  |  |  |  |  |  |  |  |  |  |
| FIX\_AUTHORS |  |  |  |  |  |  |  |  |  | x |
| FOREIGNKEYS |  |  |  |  |  |  |  |  |  | x |
| FORMLABELS |  |  |  |  |  |  |  |  |  | x |
| GANZHI\_CODES |  |  |  |  |  |  |  |  |  | X |
| HOUSEHOLD\_STATUS\_CODES |  |  |  |  |  |  |  |  |  |  |
| KIN\_DATA |  |  |  |  |  |  |  |  |  |  |
| KIN\_MOURNING |  |  |  |  |  |  |  |  |  |  |
| KIN\_MOURNING\_STEPS |  |  |  |  |  |  |  |  |  | X |
| KINSHIP\_CODES |  |  |  |  |  |  |  |  |  |  |
| LITERARYGENRE\_CODES |  |  |  |  |  |  |  |  |  |  |
| MEASURE\_CODES |  |  |  |  |  |  |  |  |  |  |
| NIAN\_HAO |  |  |  |  |  |  |  |  |  |  |
| OCCASION\_CODES |  |  |  |  |  |  |  |  |  |  |
| OFFICE\_CATEGORIES |  |  |  |  |  |  |  |  |  |  |
| OFFICE\_CODE\_TYPE\_REL |  |  |  |  |  |  |  |  |  | X |
| OFFICE\_CODES |  |  |  |  |  |  |  |  |  |  |
| OFFICE\_CODES\_CONVERSION |  |  |  |  |  |  |  |  |  |  |
| OFFICE\_TYPE\_TREE |  |  |  |  |  |  |  |  |  | X |
| OFFICE\_TYPE\_TREE\_BACKUP |  |  |  |  |  |  |  |  |  |  |
| PARENTAL\_STATUS\_CODES |  |  |  |  |  |  |  |  |  | X |
| PLACE\_CODES |  |  |  |  |  |  |  |  |  |  |
| POSSESSION\_ACT\_CODES |  |  |  |  |  |  |  |  |  |  |
| POSSESSION\_ADDR |  |  |  |  |  |  |  |  |  | X |
| POSSESSION\_DATA |  |  |  |  |  |  |  |  |  | X |
| POSTED\_TO\_ADDR\_DATA |  |  |  |  |  |  |  |  |  | X |
| POSTED\_TO\_OFFICE\_DATA |  |  |  |  |  |  |  |  |  |  |
| POSTING\_DATA |  |  |  |  |  |  |  |  |  | X |
| SCHOLARLYTOPIC\_CODES |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_ADDR |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_ADDR\_TYPES |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_ALTNAME\_CODES |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_ALTNAME\_DATA |  |  |  |  |  |  |  |  |  | X |
| SOCIAL\_INSTITUTION\_CODES |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_CODES\_CONVERSION |  |  |  |  |  |  |  |  |  | X |
| SOCIAL\_INSTITUTION\_NAME\_CODES |  |  |  |  |  |  |  |  |  |  |
| SOCIAL\_INSTITUTION\_TYPES |  |  |  |  |  |  |  |  |  |  |
| STATUS\_CODE\_TYPE\_REL |  |  |  |  |  |  |  |  |  | X |
| STATUS\_CODES |  |  |  |  |  |  |  |  |  |  |
| STATUS\_DATA |  |  |  |  |  |  |  |  |  |  |
| STATUS\_TYPES |  |  |  |  |  |  |  |  |  |  |
| TEXT\_BIBLCAT\_CODE\_TYPE\_REL |  |  |  |  |  |  |  |  |  | X |
| TEXT\_BIBLCAT\_CODES |  |  |  |  |  |  |  |  |  |  |
| TEXT\_BIBLCAT\_TYPES |  |  |  |  |  |  |  |  |  |  |
| TEXT\_CODES |  |  |  |  |  |  |  |  |  |  |
| TEXT\_INSTANCE\_DATA |  |  |  |  |  |  |  |  |  |  |
| TEXT\_ROLE\_CODES |  |  |  |  |  |  |  |  |  |  |
| TEXT\_TYPE |  |  |  |  |  |  |  |  |  |  |
| YEAR\_RANGE\_CODES |  |  |  |  |  |  |  |  |  |  |
| TablesFields |  |  |  |  |  |  |  |  |  | X |
| TablesFieldChanges |  |  |  |  |  |  |  |  |  | X |