### Excel数据原文件数据清洗步骤及数据库导入准备

#### Excel表头模板介绍

##### 1.1录入版各子表介绍

###### 1.1.1 main\_sample子表

样品的基本信息。如表序号、样品原始编号、样品数据类型、样品年龄、年龄测试方法、年龄误差、对应各子表序号、备注。

###### 1.1.2 geo\_location子表

样品的地理位置信息。含有大陆/大洋、国家、具体位置描述、经度、纬度、经纬度误差、备注。

###### 1.1.3 geo\_environ子表

样品周围地质体的地理信息。含有板块名称、地质体、具体地质环境描述、温度、压力、深度、备注。

###### 1.1.4 rock\_property子表

岩石类样品的具体信息。含有岩石名称、TAS分类名称、产状、质地、碱度、镁铁-长英质分类、酸-基性分类、备注。

###### 1.1.5 mineral\_property子表

矿物类样品的具体信息。含有矿物名称、IMA分类名称、颗粒点位号码、点位位置、备注。

###### 1.1.6 experiment\_property子表

实验类样品的具体信息。含有实验名称、温度、压力、时间、样品舱、氧逸度、氧逸度标准、备注。

###### 1.1.7 element子表

样品所含元素的信息。含有元素类别、元素名、元素值、元素单位、备注

###### 1.1.8 collector子表

数据收集者的信息。含有收集者名字、收集时间、收集来源、备注。

###### 1.1.9 reference子表

原数据的来源信息。含有发表年份、doi号、作者、标题、期刊名、备注。

##### 1.2字典表介绍

有字典表的字段为14个。

其中已固定的为：data\_type、continent、attitude、alkalinity、composition、sio2、ima\_name、sopt\_location、elem\_type、elem\_unit共10个。

未固定仍可更新的为：country、plate、tas\_name、elem\_name共4个。

具体文件见HTG数据库网站Data Template页面。

（https://htgdb.deep-time.org/dataTemplate）

#### Excel数据原文件清洗步骤

##### 2.1一般清洗步骤

**2.1.1确定数据类型**

确定样品为岩石样品、矿物样品、实验样品中的哪一类，选择对应的录入版模版。（rock\_property子表、mineral\_property子表、experiment\_property子表三者只能存在一个）

**2.1.2编制samp\_id**

每行数据设置samp\_id唯一编号，每行数据即代表一个样品。值得一提的是，origin\_num并不能代表样品的唯一编号，因为存在一个样本多个测试点位的情况，成分数据有差异，每个点位视为一个样品，拆分成不同的行，即每个点位为唯一的数据，拥有唯一的samp\_id。

若样品为矿物数据，含有母岩信息，则将母岩提取为一个岩石样品，整理出母岩数据表格。此母岩数据表格的samp\_id在矿物数据排列结束后进行编号排列。

**2.1.3数据整理**

根据表头要求（分为岩石样品、矿物样品、实验样品、矿物的母岩样品四大），按照样品类型整理数据并录入对应的表格模板中，其中每一份文件各子表的第一列应为samp\_id（element子表除外）。

对应模板见HTG数据库网站Data Template页面，https://htgdb.deep-time.org/dataTemplate。

**2.1.4数据补全**

运行Location字段补全代码和Reference字段补全代码进行数据信息补全。

**2.1.5 element转置**

运行转置代码，将element按照主量元素、微量元素、同位素转置并填入数据表格中。

##### 2.2特殊清洗步骤

本部分主要为上一部分的填充数据这一步骤的补充。由于每份原数据的特点和整理标准不同，要想将其清洗至适合本数据库录入的格式，需要进行的清洗步骤和处理方式皆有所不同。以下为清洗细则：

###### 2.2.1 Chuntao数据

1. AGE\_ERROR=(max\_age-min\_age)/2；
2. Publish\_year、authortitle皆由原数据的REFERENCE列拆分出来；

###### 2.2.2 QinBen\_CPX/OPX/Spinel/Garnet/Olivine/Xenolith数据

QinBen的六份原数据格式相同，数据清洗方式上差异很小，故在统一列举。

1. QinBen数据主体为矿物数据，每行数据都含有对应的母岩的信息。提取其中母岩数据作为岩石类型的数据录入，赋予其samp\_id。矿物数据的samp\_id编号完成后再编母岩的samp\_id。逻辑关系：子岩的parent\_id即为对应母岩的samp\_id。
2. location、environ子表中：1、原数据的TECTONIC SETTING对应HTG数据模板的terrane；2、原数据的LOCATION COMMENT对应HTG数据模板location表的notes列；3、在LOCATION中搜索continent、country、plate的dic表数据，有则填入，无则不填；4、LOCATION 依照“/”分列后，第一列选取含有carton的数据，第二列删去plate、continent 、country数据，合并后放入location表的loc\_detail列;5、LOCATION 依照“/”分列后，第三列及之后的信息放入environ表的geo\_detail列。
3. Reference子表中的信息皆来自原数据的CITATIONS列与其对应的References dictionary表。References dictionary已清洗，清洗版已放入文件夹中。
4. Latitude、Longitude数据皆取自各自最大最小值的均值，若最大值或最小值仅存在一个，则取用唯一的那个值。
5. 母岩的rock\_name为原数据的ROCK NAME列。其中，先删除“XENOLITH”后缀。出现类似“CLINOPYROXENITE, GARNET, XENOLITH”的时候，rock\_name只保留第一个逗号之前的数据，第一个逗号之后的矿物名称写进notes。
6. 原数据如果有grain size列，则写进mineral\_property表的notes列。

###### 2.2.3 ZhangZhou\_sulfide silicate\_exper数据

1. 数据为实验样品，分别进行了硅酸盐与硫化物两个实验测量部分，在element部分的区分方式不同。故将其拆分为sulfide、silicate两个表分别编制samp\_id并录入，origin\_num分别为Sample #列加上-sulfide/-sulfide后缀。
2. HTG数据模板的reference子表由原数据的REFERENCE列拆分而来。

###### 2.2.4 科马提岩数据

HTG数据模板的字段有如下对应关系：

1. composition对应的是原数据中rock\_facies列只留下mafic、ultramafic、intermediate的数据；
2. attitude对应原数据中rock\_origin列；
3. alkalinity对应原数据中frost\_class2列；
4. loc\_detail对应原数据中plate\_subplate列；
5. plate对应原数据中plate\_major列；
6. terrance对应原数据中prov\_group列；
7. geo\_detail对应原数据中prov\_name列;
8. reference子表中的数据来源于原数据year、author、bibtex列。

###### 2.2.5 Ball数据

HTG数据模板的字段有如下对应关系：

1. test\_method选用原数据的Method列；
2. age\_error选用原数据的Uncertainty列；
3. Continent、country皆选用原数据中Larger\_Region列、Region列中被列入了dic表数据，再根据经纬度进行数据补全；
4. loc\_detail则选用原数据中Larger\_Region列、Region列去掉了continent、country后合并的信息。

#### 注意事项-数据库录入相关

##### 3.1数据库的自动补充说明

后端设置主表自动补充子表的主键id列为逻辑外键，例如collect\_id、loc\_id、rock\_id，数据库皆可自动补充，其中reference子表已实现自动去重功能，下图为主表main\_sample中的逻辑外键关系；

|  |  |
| --- | --- |
| **子表名称** | **id名称** |
| **geo\_location** | loc\_id |
| **geo\_environ** | environ\_id |
| **rock\_property** | rock\_id |
| **mineral\_property** | mineral\_id |
| **experiment\_property** | exper\_id |
| **element** | elem\_id |
| **collector** | collect\_id |
| **reference** | refer\_id |

##### 3.2 Excel数据文件列格式说明

1. collect\_time数据格式采取excel单元格格式中的“yyyy/mm/dd”格式，如：2023/11/1。
2. publish\_year为整数格式。
3. samp\_age、age\_error、longitude、latitude、loc\_error、temperature、press、depth、elem\_value应保留10位小数的数值格式。
4. 其他所有字段皆为文本格式（samp\_id为整数格式），并注意清空字段前后空字符串。

##### 3.3手动数据补充处理说明

数据库设定了origin\_num、rock\_name等这些字段为非空字段。若原数据中应纳入非空字段的内容空缺，则会补充“NOT-GIVEN”；

本数据库的author字段按照英文逗号自动分隔，英文逗号后、作者名字前不应存在空格。