PWS CUP 2016 マスターデータの書式 Ver. 0.2

匿名加工コンテストで提出する匿名加工マスターデータは、以下の書式に従う必要があります。従っていないデータは受け付けられません。トランザクションデータについては別紙を参照願います。

また、再識別コンテストで与えられる匿名加工マスターデータは他チームが提出したデータそのものですので、以下の書式に従っています。

元（加工前）データは本書式に従っているので、参考にしてください。ただし、本書式は元データに比べて、0 サプレスの制限が緩いなど少し緩い書式になっているため、再識別コンテストでは注意してください。

全体的な書式

全体的な書式は次の通りです。

項目	書式
ファイル形式	CSV (ASCII)、二重引用符なし
改行文字	LF (¥n) または CRLF (¥r¥n)
行数	1 行以上、元データ行数以下
各行の文字数	48 文字以下（改行文字は含まない）

各行は 4 列（セル）で構成されます。つまり、各行にちょうど 3 つのカンマが含まれる必要があります。空行があってはいけません。最終行の改行はあってもなくても良いです。たとえば、最終行は

18287,f,1966/10/10,United Kingdom<LF>[EOF]

でも

18287,f,1966/10/10,United Kingdom[EOF]

でも良いです（<LF>は改行文字、[EOF]はファイルの終わりを示します)。

各列の書式

各列の書式は次の通りです。正規表現は Java での表記（http://docs.oracle.com/javase/jp/8/docs/api/java/util/regex/Pattern.html）です。

#	通称	書式パターン（正規表現）	書式の概要
1	顧客 ID	[1-9][0-9]{0,8}	9 桁以下の自然数
2	性別	f|m	「f」か「m」
3	生年月日	<後に掲載>	1930 年 1 月 1 日～1989 年 12 月 31 日の

			YYYY/MM/DD 形式
4	国	<後に掲載>	36 種類の値のいずれか

上記列 3（生年月日）は例外的に、書式パターンにマッチする値に対し、さらに実在する年月日であることを確認する検査がおこなわれます。書式パターン（正規表現）は次の通りです。

```
19[3-8][0-9]/(0?[1-9]|1[0-2])/[0-3]?[0-9]
```

たとえば、値「1930/2/29」はこの書式パターンにマッチしますが、年月日として実在しないため、許されません。一方、値「1932/2/29」であれば、年月日として（1932 年はうるう年なので）実在するため、許されます。

上記列 4（国）の書式パターン（正規表現）は次の通りです。

```
(?x) #コメントモード開始
Australia |
Austria |
Bahrain |
Belgium |
Brazil |
Canada |
Channel¥x20Islands |
Cyprus |
Czech¥x20Republic |
Denmark |
EIRE |
European¥x20Community |
Finland |
France |
Germany |
Greece |
Iceland |
Israel |
Italy |
Japan |
Lebanon |
Lithuania |
Malta |
```

```
Netherlands |
Norway |
Poland |
Portugal |
RSA |
Saudi¥x20Arabia |
Singapore |
Spain |
Sweden |
Switzerland |
USA |
United¥x20Arab¥x20Emirates |
United¥x20Kingdom
```

つまり、次の 36 種類の値のいずれかのみが許されます。

- Australia, Austria, Bahrain, Belgium, Brazil, Canada, Channel Islands, Cyprus, Czech Republic, Denmark, EIRE, European Community, Finland, France, Germany, Greece, Iceland, Israel, Italy, Japan, Lebanon, Lithuania, Malta, Netherlands, Norway, Poland, Portugal, RSA, Saudi Arabia, Singapore, Spain, Sweden, Switzerland, USA, United Arab Emirates, United Kingdom

各列とも、値の前後に不要な空白があるだけでも不正になることに注意してください。

列 1（顧客 ID）の注意点および例

列 1 は自然数で、元データは 5 桁ですが、匿名加工データは 9 桁以下です（符号付き 32bit 整数型に収まります）。

例：

- ○ 987654321
- × 01234　　∵ 0 が先行してはいけません。

列 2（性別）の注意点および例

列 2 は「f」か「m」です。

例：

- ○ f
- × M　　∵ 大文字は使えません。

列 3（生年月日）の注意点および例

列 3 は「年/月/日」の書式で、元データは月と日では先行の「0」がありませんが、匿名加工データでは先行の「0」が許されます。年は 1930~1989 でなくてはならず、月日は実在する値でなくてはなりません。

例：

- ○ 1930/1/1
- ○ 1989/01/01
- × 1989/1/001　　∵ 月も日も 3 桁以上に 0 を先行させてはいけません。
- × 1930/2/29　　∵ 2/29 は（1930 年には）ありません。

列 4（国）の注意点および例

列 4 は上述した 36 種類のいずれかの国名です。大文字小文字は区別されます。

例：

- ○ United Kingdom
- × united kingdom　　∵ 大文字小文字を変えてはいけません。
- × United　Kingdom　　∵ 空白の数を変えてはいけません。
- × UnitedKingdom　　∵ 空白を取り除いてもいけません。

2016 年 8 月 3 日 Ver. 0.2

PWS CUP 実行委員会