**HFCuS**

HaplotypeFrequency

Curation Service

# Input file

ASCII File, XML

## Structure

|  |  |  |  |
| --- | --- | --- | --- |
| Mandatory | XML Tag | Values / Kind of Val. | Description |
| Yes | POP\_ID | URI from POP DB | Population Identifier |
| Yes | HT\_List | List of HT |  |
| Yes | HTL | Pair of HTL\_Name, HTL\_Freq |  |
| Yes | HTL\_Name | GL String |  |
| Yes | HTL\_Freq | 0 < f <= 1 |  |
| Yes | HT\_Lic | License ID | License under which HF data is available |
| Yes | HT\_Res | G, P, gNMDP, gDKMS, n-Field, Serology |  |
| No | GT\_List | List of GTL | List of Genotypes |
| No | GTL | Record of GTL\_Name, GTL\_M\_List |  |
| No | GTL\_Name | GL String | Raw Data |
| No | GTL\_M\_List | List of GTLP\_Meth | Additional Information on GT, Typing |
| No | GTLP\_Meth | Pair of GTLP\_M\_Data,  GTLP\_M\_Value |  |
| No | GTLP\_M\_TYPE | String | Free Text / Predefined Tag |
| No | GTLP\_M\_VALUE | Free |  |
| No | GT\_Lic | License ID | License under which GT data is available |
| No | METHOD\_LIST | List of METHOD |  |
| No | METHOD | Record of METH\_Type, METH\_Value,  METH\_CLASS |  |
| No | METH\_Type | String | Free Text |
| No | METH\_Value | String | Free Text |
| No | METH\_CLASS | String | From predefined list / Text |
| No | QUALITY\_LIST | List of QUALITY |  |
| No | QUALITY | Record of QUAL\_TYPE, QUAL\_VALUE,  QUAL\_CLASS |  |
| No | QUAL\_TYPE | String | Free Text |
| No | QUAL\_VALUE | String | Free Text |
| No | QUAL\_CLASS | String | Predefined list / Free Text |
| No | LABEL\_LIST | List of LABEL |  |
| No | LABEL | Pair of LABEL\_TYPE, LABEL \_VALUE |  |
| No | LABEL \_TYPE | Free Text | Predefined List and Free Text |
| No | LABEL \_VALUE | Free Text |  |
| No | LABEL\_CLASS | Free Text | Predefined List and Free Text |
| No | ACL | ToBeDefined  <Defaults to public/private> ?? | Access Control List |
| No | COHORT\_ID |  | As an alternative to GT List, NOT THE SAME IDs as used in the HFCeS |
| No | METHOD\_ID |  | As an alternative to Method List, NOT THE SAME IDs as used in the HFCeS |
|  |  |  |  |

## Additional Stored Values

* Timestamp
* Submitting UserID

## Direct Output/Feedback

* URI to dataset
* Method\_ID
* Cohort\_ID
* HF\_ID

# Internal Data Structure of HFCuS

## Basic Fields

The basic data structure of the HFCuS mimics the input file

## Additional Fields

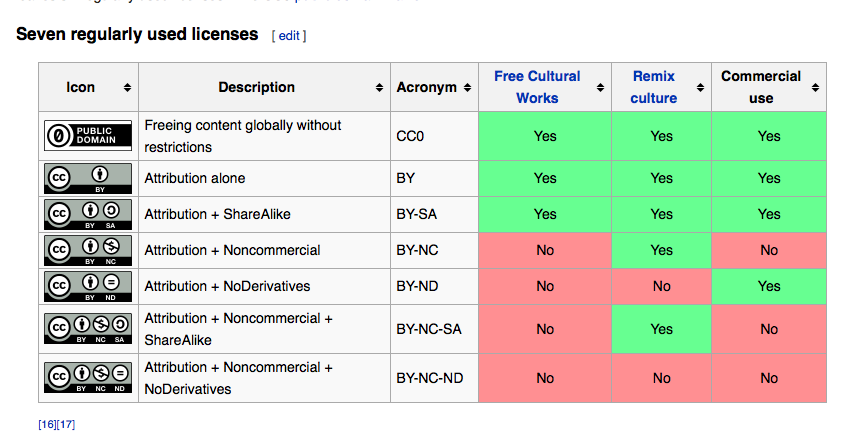
To allow for curation of submitted data sets, comments can be used. They are an independent submission to the HFCuS but refer to an existing HF data set. The field COM\_REF\_SPEC can be used to specifically address a comment to a certain piece of data in the original set.

|  |  |  |  |
| --- | --- | --- | --- |
| Madatory | Data\_Field | Content | Description |
|  | COMMENT | Record of COM\_TIME, COM\_USER, COM\_REF\_HF,  COM\_LIST | Mandatory fields:  COM\_TIME, COM\_USER, COM\_REF\_HF, COM\_LIST |
|  | COM\_TIME | Timestamp | Time of addition of the comment |
|  | COM\_USER | UserID | The user of HFCuS adding the comment |
|  | COM\_REF | Pair of COM\_REF\_TARGET, COM\_REF\_ID | The HF set the comments refer to |
|  | COM\_REF\_TARGET | Free Text/COHORT\_ID, METHOD\_ID, HF\_ID, POP\_ID, COMMENT |  |
|  | COM\_REF\_ID | ID | Appropriate Of the above target |
|  | COM\_LIST | List of COM\_REC |  |
|  | COM\_REC | Record of COM\_TEXT, COM\_REF\_SPEC |  |
|  | COM\_TEXT | Free Text | The comment |
|  | COM\_REF\_SPEC | Free Text | Some hints what the comment is referring to |

Also, if the genotype list is available, GTs can be downloaded, HF resubmitted for the same Cohort\_ID with a different (better!) methodology.

# License Models

People submitting to HFCuS shall choose one of the following options for licensing:



taken from: https://en.wikipedia.org/wiki/Creative\_Commons\_license

# User and Group Models

TBdone, TBdefined

# List of Method Tags

|  |  |  |
| --- | --- | --- |
| METH\_CLASS | HH2016 |  |
| METH\_TYPE | **VALUE** | **DESCRIPTION** |
| EM\_ALGORTIHM | String | The EM Algorithm used |
| EM\_VERSION | String | Version of the EM |
| EM\_ALG\_REF | String | A reference to the algorithm used |
| MAC\_SERVICE | String | The MultiAlleleCodeService used |
| MAC\_SER\_REF | String | A reference to the MAC Service |
| MAC\_VERSION | String | Version of the MAC Service |
| ARS\_SERVICE | String | The service used to translate typing resolutions |
| ARS\_SERV\_REF | String | A reference to the ARS service used |
| ARS\_VERSION | String | Version of the ARS Service |
| HWE\_METHOD | String | The Method used for HWE deviation estimation |
| HWE\_REF | String | A reference to the HWE deviation estimation method |
| LD\_METHOD | String | The LD estimation Method |
| LD\_METHOD\_REF | String | A reference to the LD estimation method |
| EM\_PARAM\_... |  |  |
| ARS\_PARAM\_... |  |  |
| HWE\_PARAM\_.. |  |  |

# List of Geotype-Method Tags

|  |  |  |
| --- | --- | --- |
| METH\_CLASS | HH2016 |  |
| GTLP\_M\_TYPE | **GTLP\_M\_VALUE** | **DESCRIPTION** |
| TYPING\_METHOD | SSO, SSP, Serology, SangerSequencing, NGS,  Free Text | The Typing Method used |
| TYPING\_REF | String | A reference to the typing method |
| TYPING\_DATE | Date | Date of typing |
| TYPING\_IMGT\_VER | String | Version on IMGT(/HLA) used to type the sample |
| MIRING\_REF | Reference | A reference to a MIRING compliant set of details to the typing of the sample |

# List of Quality Tags

|  |  |  |
| --- | --- | --- |
| QUAL\_CLASS | HH2016 |  |
| QUAL\_TYPE | **VALUE** | **DESCRIPTION** |
| DIV\_LAMBDA | real, < 0 | Exponent of Power Law fit  to HTF distribution (This is  called alpha in Slater et al.  Power Laws for Heavy-  Tailed Distributions?) |
| DIV\_50 | integer | Number of haplotypes  needed (in descending  order of frequency) to have  the cumulative sum be > 0.5  (Sample size sensitive!) |
| DIV\_50\_REL | Real, 0 <= x <= 1 | Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5 divided by the number of HT |
| SAM\_SIZE | integer | Number of GT |
| SAM\_POP | integer | Size of Population (approx.) |
| DIV\_PGD | Real, 0 <= x <= 1 | Population genetics diversity (1-sum f\_i ^2 N/(N-1)) |
| DIV\_HEAVY\_TAIL | Real, 0 <= x <= 1 | a is an independence  parameter of the Bayesian  SHF model that describes  how allele frequency  products correlate with  haplotype frequencies (also  correlates with the fraction  of nonzero categories) –  From Yoram SHF MS |
| RES\_TRS\_COUNT | Real, 0 <= x <= 1 | Jan knows that – (Average  number of possible  genotypes per individual?) |
| RES\_TRS | Real, 0 <= x <= 1 | Typing Resolution Score –  Average sum of square of  genotype probabilities  (imputation using  population-specific HF  estimate, could also do  uniform HF global) |
| RES\_SHARE\_AMBIG | Real, 0 <= x <= 1 | Fraction of GT with a lower resolution than definied in the resolution tag |
| RES\_MISS\_LOCI | Real, 0 <= x <= 1 | Fraction of GT with missing  loci (separate qual\_type per  locus?) |
| DEV\_HWE | Real | Deviation from HWE (using  HWE with ambiguity  method) |
| ERR\_STD | Real, 0 <= x <= 1 | Weighted average of  standard errors across all  haplotypes |
| ERR\_SAMP\_80\_100 | Real | Laurent, Excoffier “If”  between frequencies  derived from 100% set and  80% training set |
| SUM\_FREQ\_GAP | Real | Sum of haplotype  frequencies for unobserved  haplotypes that are  expected in population by  SHF model |
| ERR\_OFFSET | Real, 0 <= x <= 1 | 1-sum f\_i (Difference  between predicted full HF  distribution using SHF  versus actual including test  set?) |
| LD\_MEASURE | Real | Define in Method section –  (Where is LD measured for  quality?) |
| KFOLD\_IMPUTE | Real, 0 <= x <= 1 | % of imputable GT in 20%  test set from HT generated  in 80% training set |
| KFOLD\_PRED\_ACTUAL | Real, 0 <= x <= 1 | Divergence between  predicted and actual with  Log Loss function (for test  set predictions on simulated  lower-resolution typings) |
| KFOLD\_N | integer | Number of independent  training-test folds (k) |

# List of Labels

|  |  |  |
| --- | --- | --- |
| LABEL\_CLASS | HH2016 |  |
| LABEL\_TYPE | **VALUE** | **DESCRIPTION** |
| GT\_REGISTRY | String | ION or other description of the entity hosting the GT |
| HT\_ESTIMATION\_ENT | String | ION or other description of the entity performed the HTF analysis |
|  |  |  |

# Open Issuses for HFCuS

* User / Group Management, Schema, Specs
* ACL: Management and Specs
* Governance structure
* Implementation

# Images

