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, g_{NMDP}, g_{DKMS}, 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	

J. Sauter, DKMS, Aug. 11th, 2016

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.

J. Sauter, DKMS, Aug. 11th, 2016

License Models

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

Seven regularly used licenses [edit]

Icon	Description	Acronym	Free Cultural Works	Remix culture	Commercial use
	Freeing content globally without restrictions	CC0	Yes	Yes	Yes
	Attribution alone	BY	Yes	Yes	Yes
	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes
	Attribution + Noncommercial	BY-NC	No	Yes	No
	Attribution + NoDerivatives	BY-ND	No	No	Yes
	Attribution + Noncommercial + ShareAlike	BY-NC-SA	No	Yes	No
	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	No	No	No

[16][17]

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

J. Sauter, DKMS, Aug. 11th, 2016

List of Quality Tags

QUAL_CLASS	HH2016	
QUAL_TYPE	VALUE	DESCRIPTION
DIV_LAMBDA	real, < 0	Exponent of Power law fit to HTF dsitribution
DIV_50	integer	Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5
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	a	Martin knows that
RES_TRS_COUNT	Real, 0 <= x <= 1	Jan knows that
RES_TRS	Real, 0 <= x <= 1	Resolution score
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
DEV_HWE	real	Devition from HWE, method described in the method section
ERR_STD	Real, 0 <= x <= 1	Weighted average of standard error
ERR_SAMP_80_100	real	Laurent, Excoffier
SUM_FREQ_GAP	Real	Expected but unobserved, LD!
ERR_OFFSET	Real, 0 <= x <= 1	1-sum f_i
LD_MEASURE	real	Define in Method section
KFOLD_IMPUTE	real	% of imputable GT from HT
KFOLD_PRED_ACTUAL	Real	Divergence between prediction and actual
KFOLD_N	integer	Number of independent iterations

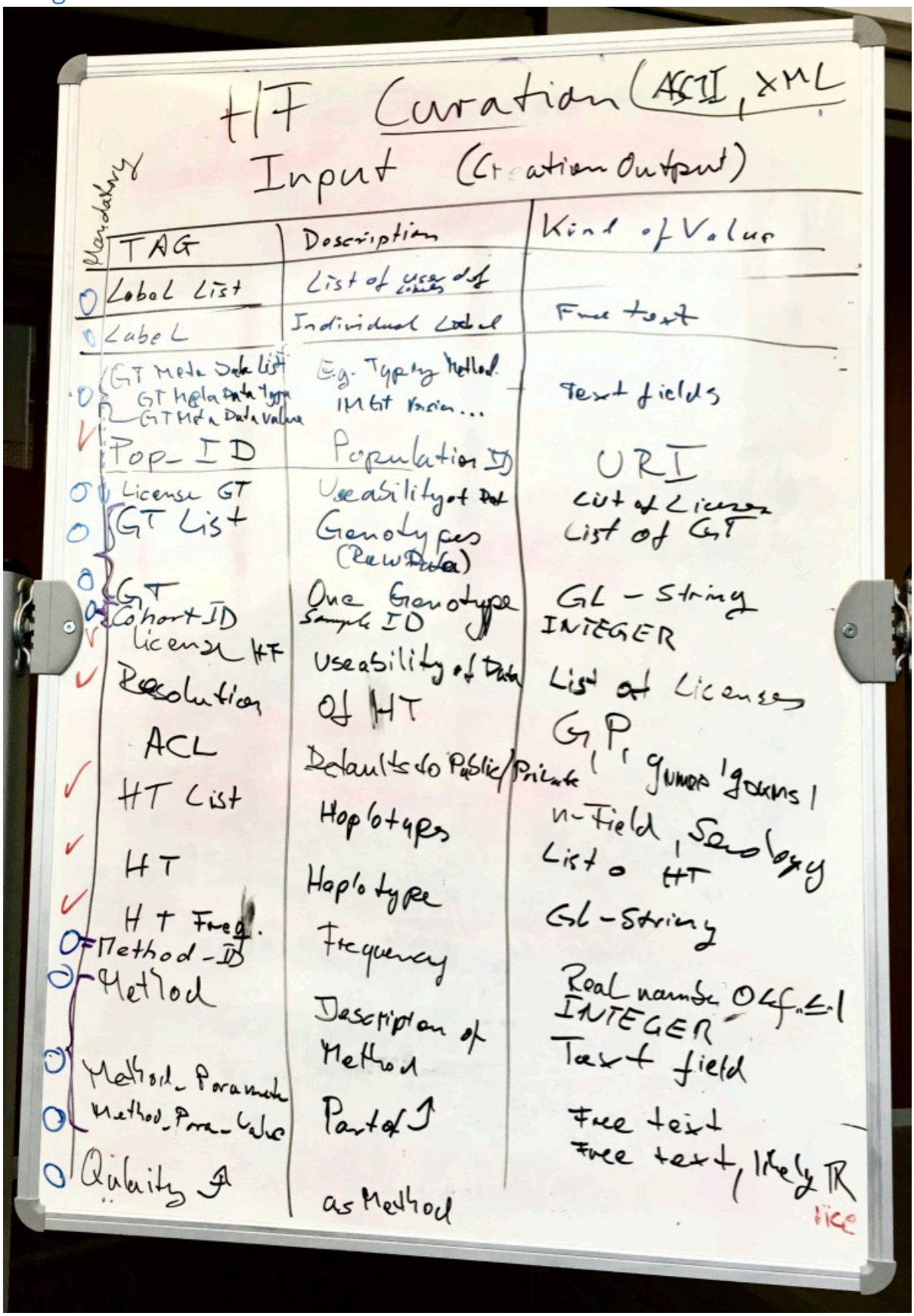
List of Labels

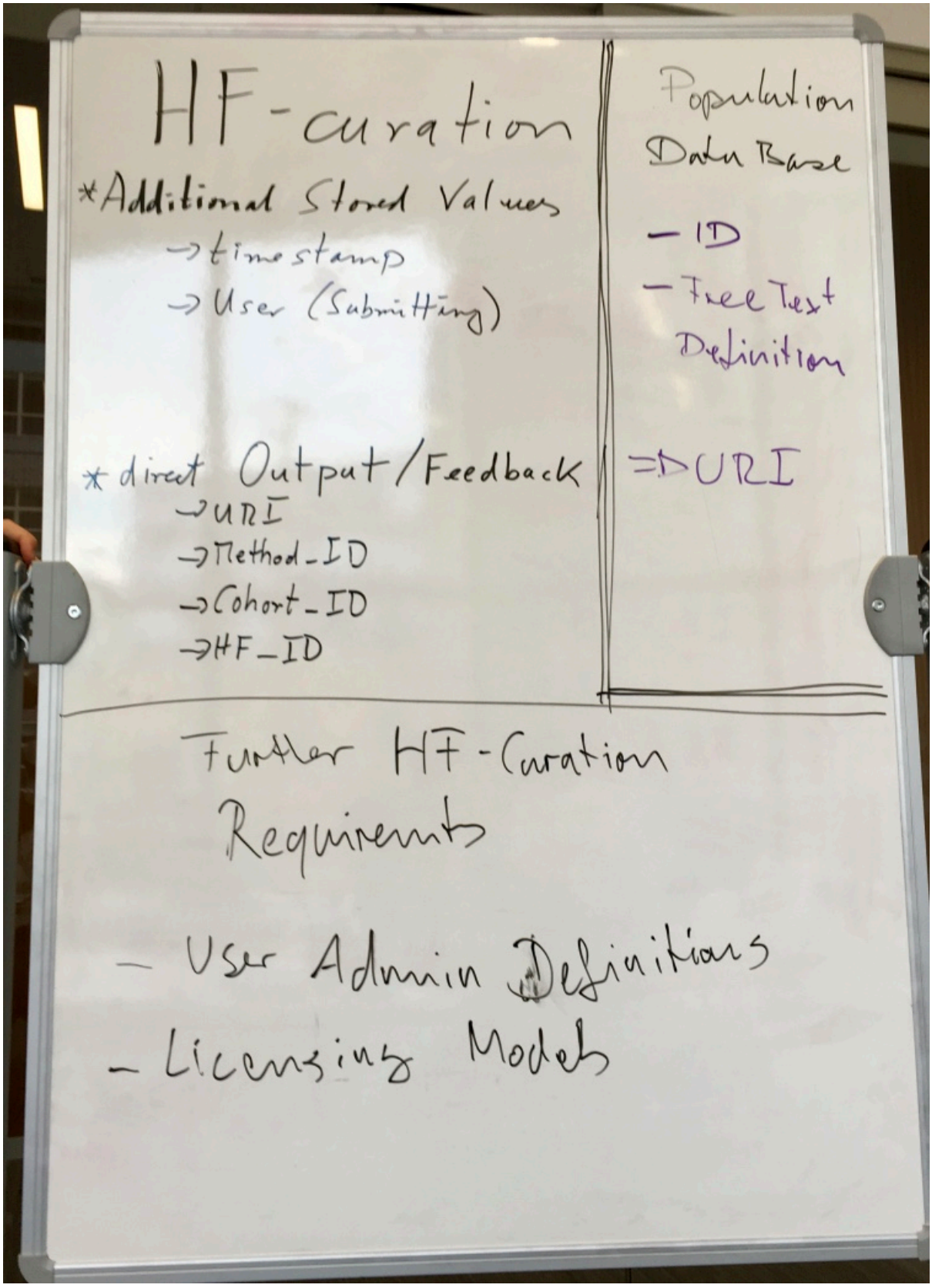
LABEL_CLASS	HH2016	
LABEL_TYPE	**VALUE**	**DESCRIPTION**
GT_REGISTRY	String	ION or other description of the entity hosting the GT
HT_ESTIMATION_ENT	String	ION or other description of the entity performed the HTF analysis

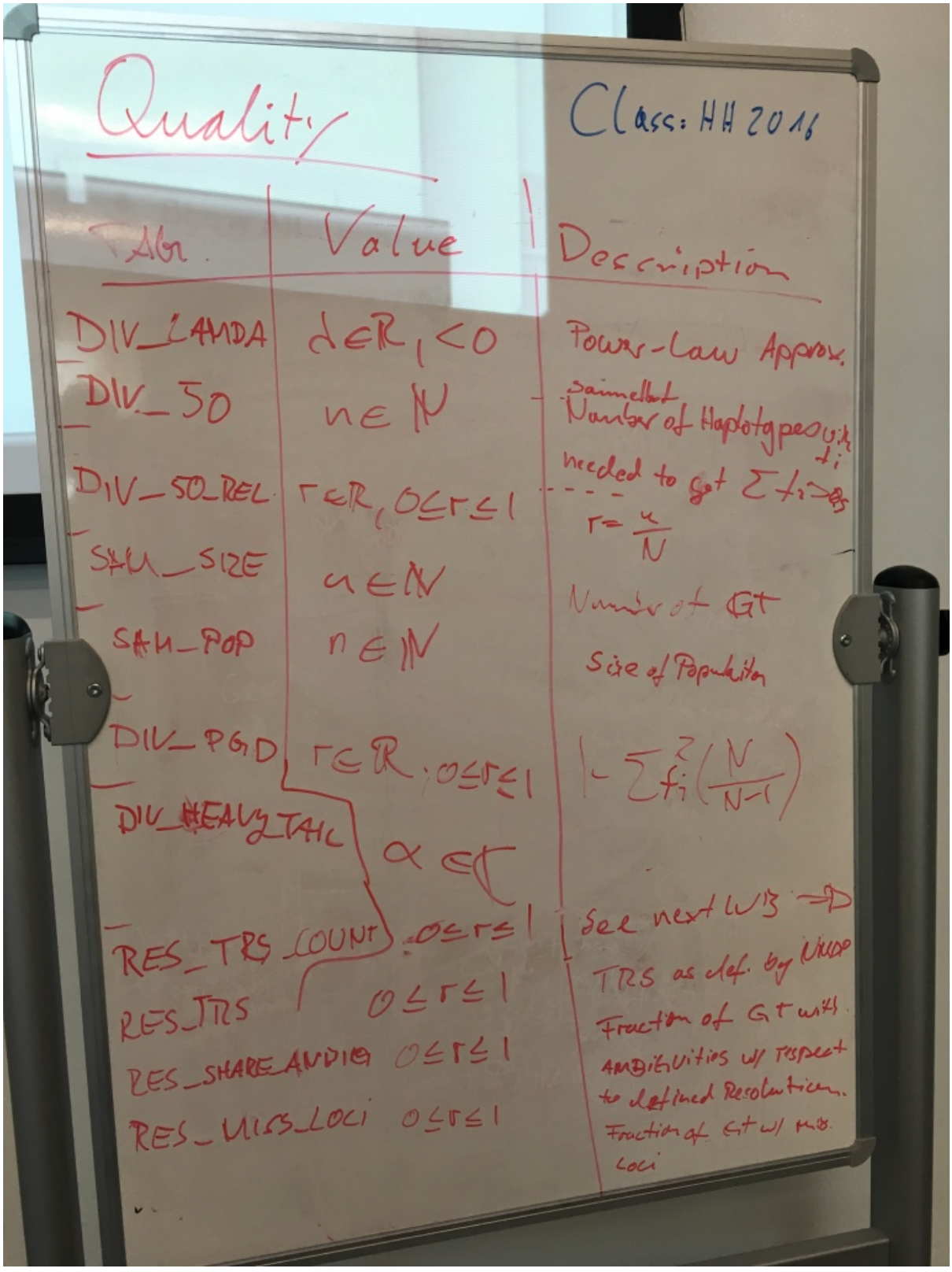
Open Issuses for HFCuS

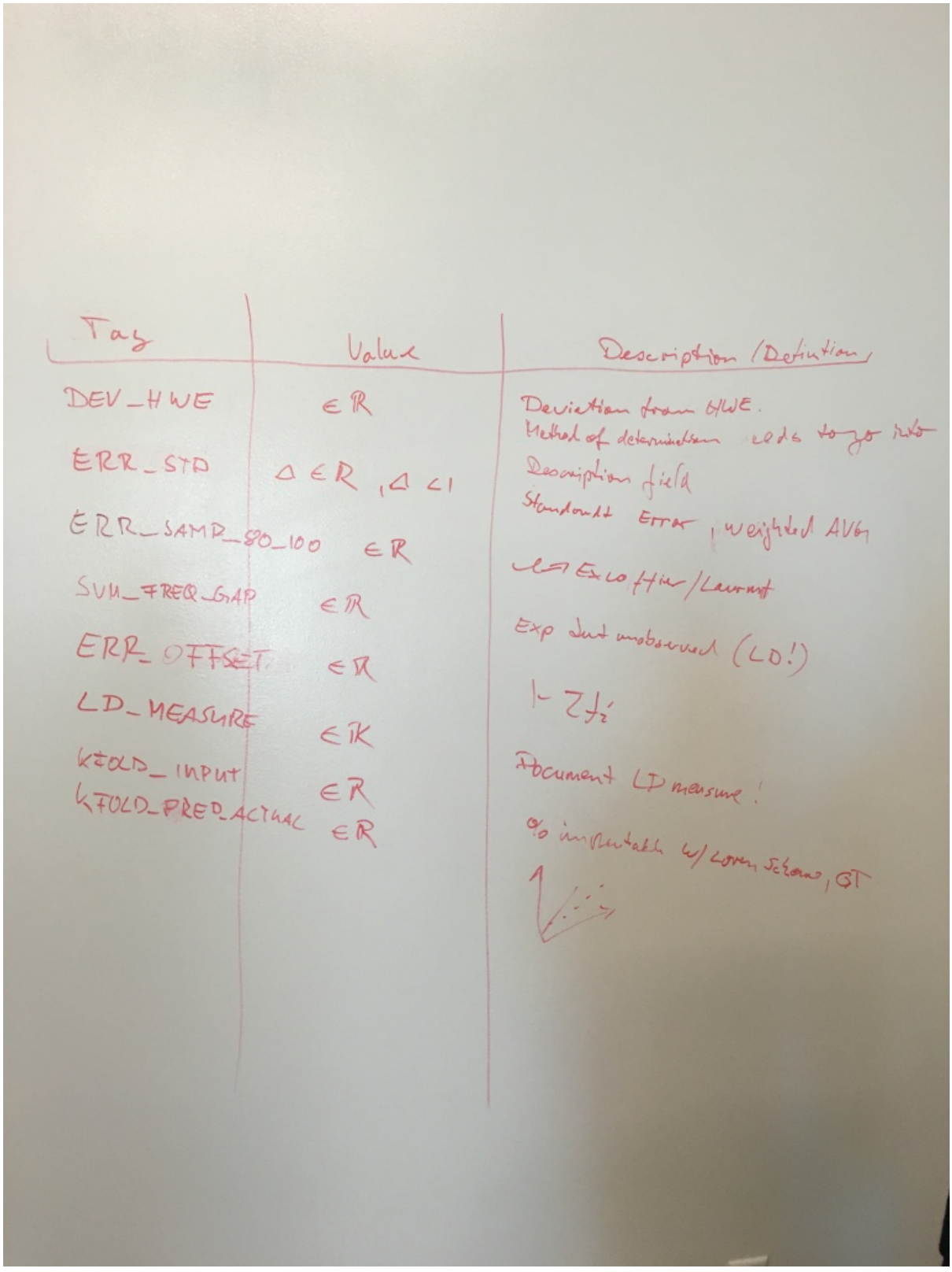
- User / Group Management, Schema, Specs
- ACL: Management and Specs
- Governance structure
- Implementation

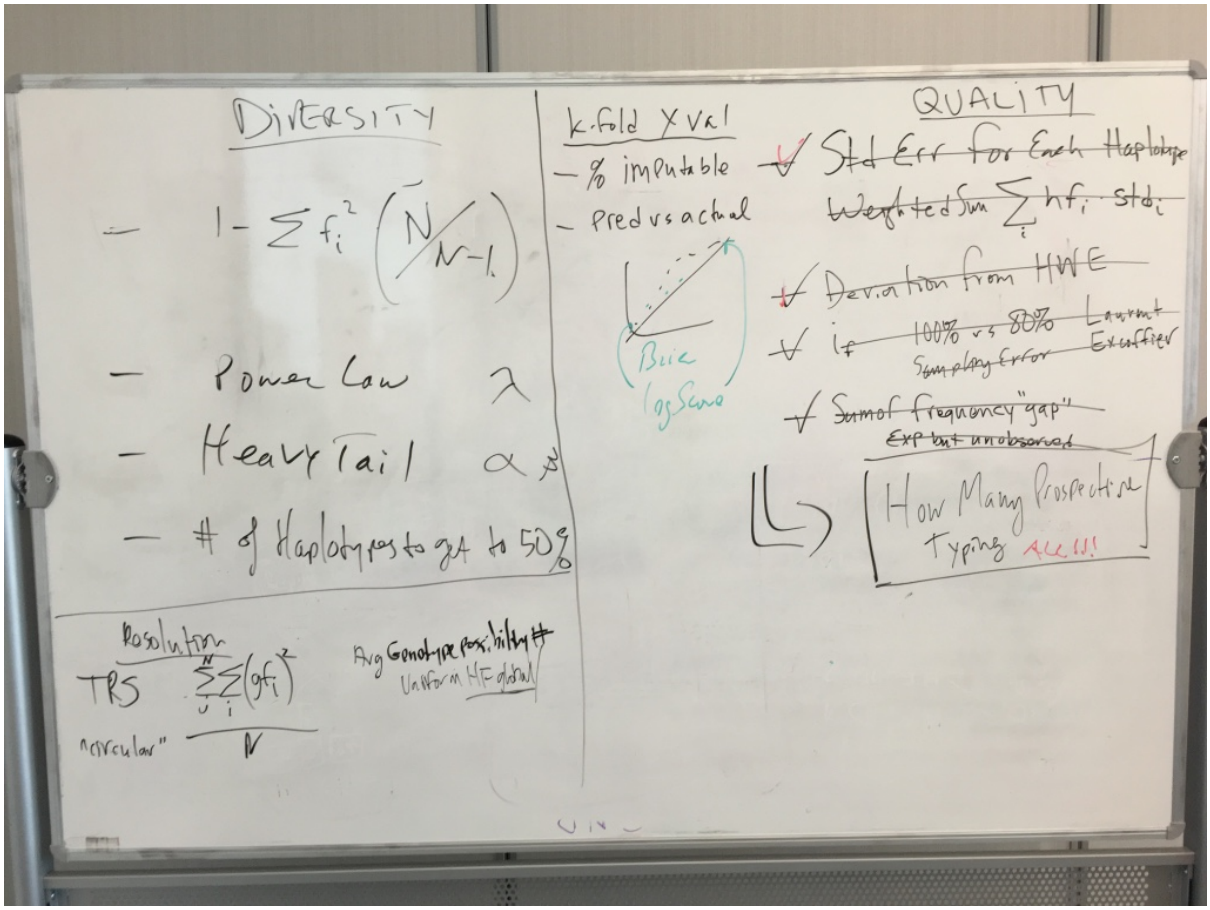
Images


HT Curation (ASCII, XML
Input (Creation Output)
Mandatory
TAG
Description
Kind of Value
Label List
List of
Label
Individual Label
Free text
GT Meta Data List
GT Meta Data Type
GT Meta Data Value
E.g. Typing Method
IMGT Version...
Text fields
Pop_ID
Population ID
URI
License GT
Useability of
List of License
GT List
List of GT
Genotypes
(Raw
GT
Cohort ID
One Genotype
Sample ID
GL - String
INTEGER
license HT
Useability of
Resolution
Of HT
List of Licenses
ACL
Defaults to Public/Private
HT List
Haplotypes
HT
Haplotype
HT Freq.
Frequency
GL-String
Method-ID
Method
Real name of
INTEGER
Description of
Method
Text field
Method_Parameter
Method_Para_Value
Part of
Free text
Quality
as Method
Free text, likely IR


HF-curation
*Additional Stored Values
→timestamp
→User (Submitting)
*direct Output/Feedback
→URI
→Method_ID
→Cohort_ID
→HF_ID
Population Data Base
- ID
- Free Text Definition
=>URI
Further HF-Curation Requirements
- User Admin Definitions
- Licensing Models


Quality
Class: HH 2016
TAG
Value
Description
DIV_LAMDA
λ∈R, λ<0
Power-Law Approx.
DIV_S0
n∈N
Number of Haplotypes with ti needed to get Σti=S
DIV_S0_REL
r∈R, 0≤r≤1
r = n/N
SAM_SIZE
n∈N
Number of GT
SAM_POP
n∈N
Size of Population
DIV_PGD
r∈R, 0≤r≤1
DIV_HEAVY_TAIL
α∈
RES_TRS_COUNT
0≤r≤1
See next WB
RES_TRS
0≤r≤1
TRS as def. by NMDP
RES_SHARE_AMBIG
0≤r≤1
Fraction of GT with AMBIGUITIES w/ respect to defined Resolution
RES_MISS_LOCI
0≤r≤1
Fraction of GT w/ miss. Loci


Tag
Value
Description / Definition
DEV_HWE
∈ ℝ
Deviation from HWE.
ERR_STD
Δ ∈ ℝ, Δ < 1
Description field
Standard Error, weighted AVG
ERR_SAMP_80_100
∈ ℝ
SUM_FREQ_GAP
∈ ℝ
Exp but unobserved (LD!)
ERR_OFFSET
∈ ℝ
LD_MEASURE
∈ ℝ
Document LD measure!
KFOLD_INPUT
∈ ℝ
KFOLD_PRED_ACTUAL
∈ ℝ


DIVERSITY
$1 - \sum f_i^2 \left(\frac{N}{N-1}\right)$
Power Law λ
Heavy Tail α
# of Haplotypes to get to 50%
Resolution
TRS
"circular"
$\frac{\sum_u \sum_i (gf_i)^2}{N}$
Avg Genotype possibility #
Uniform HF plateau
k-fold XVal
% imputable
Pred vs actual
Bias
QUALITY
Std Err for Each Haplotype
Weighted Sum $\sum_i hf_i \cdot std_i$
Deviation from HWE
If 100% vs 80% Simplify Error
Learner Exciffier
Sum of frequency "gap"
Exp but unobserved
How Many Prospective Typing ALL!!!