Evaluation of ESPERANTO
Case study based on the curation of multiple single datasets and their integration

1. Aims

ESPERANTO is a tool designed to ensure a GLP-compliant (Good Laboratory Practice) streamlined and standardized harmonisation of toxicogenomics (TGx) metadata. The process, carried in an user-friendly interface, results in curated datasets characterized by higher FAIR-ness. The tool offers comprehensive documentation that tracks all modifications and operations executed on the datasets. Additionally, in GLP mode, the user is required to add a mandatory comment to every operation.
ESPERANTO also manages the integration of different single curated datasets in a consistent fashion, creating the best premises to generate high-quality insights from the data.

2. Brief random notes about ESPERANTO

This document does not provide a detailed description of how ESPERANTO functions. For this purpose, the link to the updated and detailed User Guide can be found in the readme at https://github.com/fhaive/esperanto.
ESPERANTO´s performance is based on the cross-comparison between the dataset(s) and a reference *ad hoc* vocabulary. Each curation round allows the user to enrich and update the vocabulary previously used.
The user is encouraged to think of both dataset and vocabulary files in terms of data tables, with "***labels***" naming the columns, and "***contents***" in the different cells.
The vocabulary is build as a nested synonym vocabulary, where a key reference label is linked to label synonyms and to a series of key reference contents. In turn, each of the latter is associated with its synonyms.

3. Evaluation strategy

To test ESPERANTO's efficiency, we evaluated its performances in two different scenarios. At the start, ESPERANTO generates a unique ***SHA-256 ID*** for each running session: each file saved, and report generated result associated to that unique identifier.
Input GSE datasets, vocabulary used, saved session files and reports are provided in the *case_study_files* folder uploaded at https://github.com/fhaive/esperanto.

3.1 Single dataset

For this case study, we utilise two distinct datasets publicly available on Gene Expression Omnibus (GEO) repository (GSE199152 and GSE53845) and are relevant for Idiopathic Pulmonary Fibrosis (IPF).
The reference vocabulary to upload can be the starting version provided with this tool, or a pre-existing customized dictionary from the user if they have the same structure.
The starting reference vocabulary can be considered empty since the implemented filler row is removed as soon as the vocabulary starts to be used and populated.

At the same time, the curation of each dataset provided a list of label and content potential candidates: the user was asked to evaluate whether incorporate them into an updated version of the uploaded vocabulary.
The cross comparison of the following dataset will be performed with the updated dictionary.
GLP mode was not activated to privilege the operational linearity for the descriptive purpose of this document, but any performed operation on the data was recorded in the procedural track report. ESPERANTO generates a unique sha-256 ID that identifies all the files and reports of the same session.
Despite the GLP/procedural track report, ESPERANTO documents different aspects of the curation through several reports that guide any future user to replicate the same curated outcome obtained by the original curator.
Regarding the vocabulary update, three different reports identify the candidates accepted, those discarded and those that currently present issues to revise in a second session.
After each curation round and the correspondent potential incorporation of new-entry candidates, we evaluated:

1. the enrichment of the reference vocabulary in terms of new entries incorporated
2. the GLP-fication of the curation by providing a detailed report regarding the operative pipeline.

3.2 Integration of multiple datasets

In the second scenario, we considered the integration of the two datasets previously curated, using the last updated version of the vocabulary to categorise the entries of the integrated table and speed up the harmonisation quality check. As for single datasets, GLP/procedural track is one of the reports that composes the documentation of the analysis. In particular, documentation will be generated to record “consistent” as well as problematic entries to revise in a second curation session.
As for single datasets, we evaluated:

1. the merging of the curated dataset into a main harmonised table
2. the GLP-fication of the integration process by supplying a detailed report documenting the operations performed.

4. Curation of Single Datasets

The following tables belong to ESPERANTO analysis report, where the main results of the curation are summarised. The tables presented in the following paragraphs cover the pre-curation vs post-curation condition of the dataset, the vocabulary before and after the update, and an overview of the updating process.

4.1 Dataset I: GSE199152

As previously mentioned, this first dataset was curated using an empty reference vocabulary. (Sha256 ID: *e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e*)

4.1.1 Phenodata (I)

The current dataset required less than 30 operations to be curated (Table 2). Due to the empty nature of the input vocabulary (Table 4), the majority of operations falls under the

"Duplicate Removal" and "Content Homogenization" section of ESPERANTO, where the automation is limited to implementing the operations column by column.

Loaded Dataset Features

file name	#samples (rows)	#variables (columns)	Total Fields
GSE199152.xlsx	27	40	1080

Table 1 - Characteristics of pristine GSE199152

Synthesis of the Operations Performed to Curate the Dataset

Type of Operation	#operation(s)
Accepted Suggested Relabelling	0
Rejected Suggested Relabelling	0
Duplicate Removal	3
Deletion	0
Recoding Full Column	14
Skipped Full Column Recoding	10
Specials	0
Total	**27**

Table 2 - Number and type of the operations required to complete the curation of the current dataset

Curated Dataset Features

file name	#samples (rows)	#variables (columns)	Total Fields
curated_GSE199152.xlsx	27	37	999

Table 3 - Characteristics of curated GSE199152

4.1.2 Vocabulary (I)

The curation round of GSE199152, although a relatively small dataset (Table 1), generated 56 entries that were positively evaluated for vocabulary enrichment.

Overview of Original Vocabulary Features

Entry Type	#Entries
Labels	0
Label Synonyms	0
Contents	0
Content Synonyms	0
Total Entries	**0**

Table 4 - Characteristics of the starting vocabulary

Overview of the Updated Vocabulary Features

Entry Type	#Updated Entries
Labels	22
Label Synonyms	13
Contents	13
Content Synonyms	8
Updated Total Entries	**56**

Table 5 - Characteristics of the updated vocabulary

4.1.3 New entries vocabulary evaluation (I)

As outlined in the manual, the user can classify the candidate for vocabulary enrichment as either “Issue”, “Discard”, or “Accepted”. The outcome of this process is shown in Table 6. To ensure transparency and accountability, the Curator and Arbiter responsible for the classification are also recorded (Table 7). These measures aim to enhance the overall quality and reliability of the dataset by providing clear documentation of the curation process.

Type of Operation performed during Candidate Evaluation

Classification Type	#Operations
Issue	0
Discarded	0
Accepted	24
Unprocessed	0
Total Operations for Candidate Evaluation	**24**

Table 6 - Type of operations to evaluate entry candidates

It shows only the first 6 accepted entries, but the complete reports are available in *case_study_files* folder (par. 3).

New entries evaluated as Correct

ID	Label Type of Storing	Label	Label Synonym	Content Type of Storing	Content	Content Synonym	Decision	Curator	Arbiter
1	2	type	NA	4	SRA	NA	Newly Accepted	Simol	Simol
2	2	channel_count	NA	4	1	NA	Newly Accepted	Simol	Simol
3	1	tissue_source	source_name_ch1	3	biopsy	Lung biopsy	Newly Accepted	Simol	Simol
4	1	organism	organism_ch1	3	homo_sapiens	Homo sapiens	Newly Accepted	Simol	Simol
5	1	molecule	molecule_ch1	3	total_RNA	total RNA	Newly Accepted	Simol	Simol
6	1	extract_protocol	extract_protocol_ch1	4	NA	NA	Newly Accepted	Simol	Simol

Table 7 - First 6 entries accepted for vocabulary enrichment

4.1.4 GLP-fication of the curated dataset (I)

Table 8 shows part of the complete procedural track report available in *case_study_files f*older (par. 3). It lists the operation performed and the classification of the recoded entries as potential candidate for vocabulary enrichment.

The curation session has improved the FAIR-ness of the dataset. In particular, the procedural track is the way ESPERANTO ensures the reuse of the dataset and the reproducibility of the harmonisation process.

The other reports highlight specific aspects of the curation process, and they also support the user in integrating the information delivered by the procedural track.

Step ID	Type of Operation
1	**SESSION ID:** e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e
2	USER - Current session was carried over by: Simol (2023-03-15)
3	MODE - Current session takes SINGLE dataset as input.
4	INPUT FILE - GSE199152.xlsx
5	VOCABULARY VERSION FILE - empty_dict.xlsx
6	The column 'geo_accession' is kept, while its duplicate '…1' is deleted.
7	The column 'diagnosis.ch1' is kept, while its duplicate 'characteristics_ch1.1' is deleted.
8	The column 'tissue.ch1' is kept, while its duplicate 'characteristics_ch1' is deleted.
9	SAVED the current session: as 2023-03-15_13-42-16_GSE199152_current_session.RData
10	Confirmed Label: type Confirmed Content: SRA As confirmed value, 'type' is temporarily stored only as 'Reference Label' and discarded as 'Synonym'. As confirmed value, 'SRA' is temporarily stored only as 'Reference Content' and discarded as 'Synonym'.
11	Confirmed Label: channel_count Confirmed Content: 1 As confirmed value, 'channel_count' is temporarily stored only as 'Reference Label' and discarded as 'Synonym'. As confirmed value, '1' is temporarily stored only as 'Reference Content' and discarded as 'Synonym'.
12	Edited Label: tissue_source Edited Content: biopsy 'source_name_ch1' is temporarily stored as a 'Label Synonym' of the edited label. 'Lung biopsy' is temporarily stored as a 'Content Synonym' of the edited content.
13	Edited Label: organism Edited Content: homo_sapiens 'organism_ch1' is temporarily stored as a 'Label Synonym' of the edited label. 'Homo sapiens' is temporarily stored as a 'Content Synonym' of the edited content.
14	SAVED the current session: as 2023-03-15_13-45-11_GSE199152_current_session.RData
15	Edited Label: molecule Edited Content: total_RNA 'molecule_ch1' is temporarily stored as a 'Label Synonym' of the edited label. 'total RNA' is temporarily stored as a 'Content Synonym' of the edited content.
16	Edited Label: extract_protocol Confirmed Content: 1 out of 1 contents evaluated. 'extract_protocol_ch1' is temporarily stored as a 'Label Synonym' of the edited label. Skipped full recoding, and all original entries were discarded from being potential 'Reference Content'.

Table 8 - Subsection of the procedural track for GSE199152 curation

4.2 Dataset II: GSE53845

Compared to the previous curation round, GSE53845 will use the updated vocabulary version generated during GSE199152 harmonisation session.

(Sha256 ID: *e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12*)

4.2.1 Phenodata (II)

Curation of GSE53845 required 54 steps to appear as a harmonised version lighter of the 20% of the original entries (Table 9 and Table 11).

Loaded Dataset Features

file name	#samples (rows)	#variables (columns)	Total Fields
GSE53845.xlsx	48	52	2496

Table 9 - Characteristics of pristine GSE53845

The type of operations needed to curate GSE53845 is influenced by the increased complexity of the vocabulary. Compared to par. 4.1, the user now authorized the relabelling suggestions automatically proposed by ESPERANTO (Table 10).

Synthesis of the Operations Performed to Curate the Dataset

Type of Operation	#operation(s)
Accepted Suggested Relabelling	6
Rejected Suggested Relabelling	1
Duplicate Removal	12
Deletion	2
Recoding Full Column	19
Skipped Full Column Recoding	14
Specials	0
Total	**54**

Table 10 - Number and type of the operations required to complete the curation of the current dataset

In fact, in the long run, the vocabulary enrichment will constantly reduce the active modification of the entries by the user, privileging the automatized retrieval of the reference label/content from the reference vocabulary and limiting the user intervention to supervision.

Curated Dataset Features

file name	#samples (rows)	#variables (columns)	Total Fields
curated_GSE53845.xlsx	48	42	2016

Table 11 - Characteristics of curated GSE53845

4.2.2 Vocabulary (II)

By using as reference vocabulary the one originated by the curation of GSE199152, harmonisation of GSE53845 was able to generate 50 accepted new entries. 34% of them are new labels establishing new categories, 24% new contents and 42% label/content synonyms.

Overview of Original Vocabulary Features

Entry Type	#Entries
Labels	22
Label Synonyms	13
Contents	13
Content Synonyms	8
Total Entries	**56**

Table 12 - Characteristics of the input vocabulary

The motivation behind curation is reflected in the increase in the number of entries. Although data in GEO are presented in a certain format, there is always some degree of "subjectivity", which leads to the identification of new candidates for vocabulary after each round of curation.

Overview of the Updated Vocabulary Features

Entry Type	#Updated Entries
Labels	39
Label Synonyms	28
Contents	25
Content Synonyms	14
Updated Total Entries	**106**

Table 13 - Characteristics of the updated vocabulary

4.2.3 New entries vocabulary evaluation (II)

As previously mentioned in par. 4.1.3 and outlined in more detail in the manual, the user can classify candidates for vocabulary enrichment as "Issue", "Discard", or "Accepted". The result of that process is shown in Table 14. Curator and Arbiter are also recorded allowing users to trace the origin and classification of each entry (Table 15).

Type of Operation performed during Candidate Evaluation

Classification Type	#Operations
Issue	0
Discarded	0
Accepted	33
Unprocessed	0
Total Operations for Candidate Evaluation	**33**

Table 14 - Type of operations to evaluate entry candidates

The complete documentation is available in the *case_study_files* folder (par. 3).

New entries evaluated as Correct

ID	Label Type of Storing	Label	Label Synonym	Content Type of Storing	Content	Content Synonym	Decision	Curator	Arbiter
1	2	data_processing	NA	4	NA	NA	Accepted	Simol	Simol
2	2	contact_institute	NA	4	NA	NA	Accepted	Simol	Simol
3	2	contact_city	NA	4	NA	NA	Accepted	Simol	Simol
4	2	extract_protocol	NA	4	NA	NA	Accepted	Simol	Simol
5	2	organism	NA	4	homo_sapiens	NA	Accepted	Simol	Simol
6	2	molecule	NA	4	total_RNA	NA	Accepted	Simol	Simol

Table 15 - First 6 entries accepted for vocabulary enrichment

4.2.4 GLP-fication of the curated dataset (II)

Table 16 shows a selection of the complete procedural track report available in the *case_study_files* folder (par. 3). This report provides a list of operations performed during the curation session and classifies the recoded entries as potential candidate for vocabulary enrichment.

Through this process, the FAIR-ness of the dataset has been significantly improved. The procedural track ensures that the dataset is reusable and the harmonisation process is reproducible through the standardized pipeline of ESPERANTO.

The additional reports highlight specific aspects of the curation process and support the user in integrating the information delivered by the procedural track reports.

Step ID	Type of Operation
1	**SESSION ID:** e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12
2	GLP mode disabled
3	USER - Current session was carried over by: Simol (2023-03-27)
4	MODE - Current session takes SINGLE dataset as input.
5	INPUT FILE - GSE53845.xlsx
6	VOCABULARY VERSION FILE - updated_vocabulary_v2023-03-15_14-08-14_e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e.xlsx
7	SAVED the current session: as 2023-03-27_10-54-12_GSE53845_current_session.RData
8	'tissue_source' is rejected as relabelling candidate for 'source_name_ch1'.
9	'organism' relabels 'organism_ch1'.
10	'molecule' relabels 'molecule_ch1'.
11	'extract_protocol' relabels 'extract_protocol_ch1'.
12	'taxid' relabels 'taxid_ch1'.
13	'disease' relabels 'diagnosis.ch1'.
14	'tissue' relabels 'tissue.ch1'.
15	SAVED the current session: as 2023-03-27_10-56-51_GSE53845_current_session.RData
16	The column 'geo_accession' is kept, while its duplicate '…1' is deleted.
17	All duplicates ("organism", "organism_ch2") are kept.
18	The column 'source_name_ch2' is kept, while its duplicates 'characteristics_ch2', 'sample.type.ch2' are deleted.
19	All duplicates ("treatment_protocol_ch1", "treatment_protocol_ch2") are kept.
20	All duplicates ("extract_protocol", "extract_protocol_ch2") are kept.
21	All duplicates ("label_protocol_ch1", "label_protocol_ch2") are kept.
22	All duplicates ("taxid", "taxid_ch2") are kept.
23	The column 'disease' is kept, while its duplicate 'characteristics_ch1.2' is deleted.
24	The column 'disease.state.ch1' is kept, while its duplicate 'characteristics_ch1' is deleted.
25	The column 'gender.ch1' is kept, while its duplicate 'characteristics_ch1.4' is deleted.
26	The column 'source.ch1' is kept, while its duplicate 'characteristics_ch1.3' is deleted.
27	The column 'tissue' is kept, while its duplicate 'characteristics_ch1.1' is deleted.
28	SAVED the current session: as 2023-03-27_11-00-01_GSE53845_current_session.RData
29	RESTORED the previous session at 2023-03-27_11-19-25
30	USER - From now, the restored session is carried over by: Simol (2023-03-27)
31	Confirmed Label: type Confirmed Content: RNA As confirmed value, 'type' is temporarily stored only as 'Reference Label' and discarded as 'Synonym'. As confirmed value, 'RNA' is temporarily stored only as 'Reference Content' and discarded as 'Synonym'.
32	Confirmed Label: channel_count Confirmed Content: 2 As confirmed value, 'channel_count' is temporarily stored only as 'Reference Label' and discarded as 'Synonym'. As confirmed value, '2' is temporarily stored only as 'Reference Content' and discarded as 'Synonym'.

Table 16 - Subsection of the procedural track for curation session

5. Integration of multiple curated datasets

5.1 Integrated Dataset (M)

Curated GSE199152 and GSE53845 datasets were successfully uploaded and merged together into a table with 4200 fields (Table 17).

Loaded Dataset Features

file name	#samples (rows)	#variables (columns)	#Total Fields
updated_GSE53845v2023-03-30_18-17-47_e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12.xlsx	48	43	2064
updated_GSE199152v2023-03-29_16-51-39_e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e.xlsx	27	38	1026
Merged version	**75**	**56**	**4200**

Table 17 - Characteristics of single curated datasets and of the integrated table

5.2 Vocabulary (M)

The integrated dataset was cross-compared with the latest updated version of the vocabulary (in this case the one resulting from the curation of GSE53845, shown in Table 18). The cross-comparison resulted in color-coded entries, with green indicating that the data labels/contents are already present in the reference vocabulary, and red to claim that they are not.

Overview of Original Vocabulary Features

Entry Type	#Entries
Labels	39
Label Synonyms	28
Contents	25
Content Synonyms	14
Total Entries	**106**

Table 18 - Characteristics of the uploaded reference vocabulary to evaluate the integration of multiple curated datasets

5.3 Integration Evaluation (M)

55 operations were needed to evaluate the quality of the integration of the datasets previously curated.

Full File Analysis

Classification Type	#processed columns
Issue	0
Consistent	55
Processed Total	55

Table 19 - Type of operations to evaluate the integration of multiple curated datasets

The color-coding is the base on which the user can evaluate the entries to identify coherent terms as well as those that may require additional curation at the level of individual datasets. Table 20 presents the first 6 columns evaluated as successfully integrated.

	Consistent Entries	Arbiter
1	**type** is accepted. (CONSISTENT)	Simol
2	**channel_count** is accepted. (CONSISTENT)	Simol
3	**source_name_channel_1** is accepted. (CONSISTENT)	Simol
4	**organism** is accepted. (CONSISTENT)	Simol
5	**molecule** is accepted. (CONSISTENT)	Simol
6	**label_channel_1** is accepted. (CONSISTENT)	Simol

Table 20 - First 6 entries classified as consistent during the evaluation of the integration of multiple curated datasets

5.4 GLP-fication of the integration of curated dataset (M)

Table 21 shows an extract of the complete procedural track report available in *case_study_files* folder (par. 3)

Step ID	Type of Operation
1	**SESSION ID:** 368fd61ea8c54896be566c5c01235a5892cf134e9134d70b1a17a3d7accf9259
2	**SESSION ID:** 368fd61ea8c54896be566c5c01235a5892cf134e9134d70b1a17a3d7accf9259
3	GLP mode disabled
4	USER - Current session was carried over by: Simol (2023-03-30)
5	MODE - Current session takes MULTIPLE datasets as input.
6	INPUT FILES - updated_GSE53845v2023-03-30_18-17-47_e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12.xlsx updated_GSE199152v2023-03-29_16-51-39_e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e.xlsx
7	INPUT FILES - updated_GSE53845v2023-03-30_18-17-47_e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12.xlsx updated_GSE199152v2023-03-29_16-51-39_e535c4690a151551d5c740ea55c0100d5ab0736a4432bf8560397443e14ca31e.xlsx
8	VOCABULARY VERSION FILE - updated_vocabulary_v2023-03-30_18-17-59_e0d04f2dd4522fb33ad3d9f7f0411f5598d278da32f543054bd35b5900060f12.xlsx
9	SAVED the current session: as 2023-03-30_18-24-48_multiples_current_session.RData
10	**type** is accepted. (CONSISTENT)
11	**channel_count** is accepted. (CONSISTENT)
12	**source_name_channel_1** is accepted. (CONSISTENT)
13	**organism** is accepted. (CONSISTENT)
14	**molecule** is accepted. (CONSISTENT)
15	**label_channel_1** is accepted. (CONSISTENT)
16	**taxid** is accepted. (CONSISTENT)
17	**organism_channel_2** is accepted. (CONSISTENT)
18	**molecule_channel_2** is accepted. (CONSISTENT)
19	**label_channel_2** is accepted. (CONSISTENT)
20	**taxid_channel_2** is accepted. (CONSISTENT)
21	**disease** is accepted. (CONSISTENT)
22	**gender** is accepted. (CONSISTENT)
23	**tissue_source** is accepted. (CONSISTENT)
24	**tissue** is accepted. (CONSISTENT)
25	**library_selection** is accepted. (CONSISTENT)
26	**library_source** is accepted. (CONSISTENT)
27	**library_strategy** is accepted. (CONSISTENT)
28	SAVED the current session: as 2023-03-30_20-04-55_multiples_current_session.RData
29	**title** is accepted. (CONSISTENT)
30	**geo_accession** is accepted. (CONSISTENT)
31	**status** is accepted. (CONSISTENT)
32	**submission_date** is accepted. (CONSISTENT)

Table 21 - Extract of the procedural track for the integration of multiple curated datasets

6. Conclusions

The two datasets have been meticulously curated to ensure a smooth integration, an essential step towards performing further analysis.
The curation process and the entire decision-making pipeline have been well-documented to ensure reproducibility for new users.

The generated documentation provides a step-by-step guide to replicate the exact outcome achieved by the first researcher.
The GLP mode will supplement the main by providing additional information about the list of operations performed and the reasoning behind each decision made by the researcher, enabling future users to gain a deeper understanding of the applied methodology. Additionally, the vocabulary has been enriched as consequence of the curation round and subjected to an additional quality check by the user to ensure consistency and accuracy.
Overall, the curated datasets and accompanying reports provide a comprehensive resource for researchers seeking to leverage this valuable resource for data harmonisation and integration. By providing this level of documentation, ESPERANTO contributes to the dataset's reliability and reproducibility, ensuring that future analyses can be conducted with confidence.