Overview of TagWorks file formats

TagWorks collects annotations using two types of task presenters, the Highlighter presenter and the Data Hunt presenter, which each create data in a format specific to their purpose.

When exporting contributor data then, we have two main types:

- Highlighter
- Data Hunt

The Data Hunt format has additional columns for schema data, that is, the schema, the topic, question, and answer data and related data like question and answer numbers for each row.

Once data from multiple contributors has been merged algorithmically or adjudicated by an expert into a result set, those tags are saved in a “tag container”. The tag container export formats also vary by whether the source data came from a Highlighter or a Data Hunt. The main difference is that the Data Hunt exports an answer_uuid instead of the Highlighter topic_name column.

Filename changes for exports from Public Editor after January 19, 2020:

The **task_uuid** column was changed to **source_task_uuid** column, which has different values than the old column. (It is not just a rename of the column heading.) However, it is used for the exact same purpose, identifying the unique task that was presented to obtain the data in that row.

- Choosing *Export Task Runs* for Highlighter projects now exports two files instead of one:

Filename suffix	Description
-Highlighter.csv.gz	Same as before. One row per text highlight, with dis-contiguous selections of a case on separate rows.
-HighlighterByCase.csv.gz	Same data but the last column is a JSON object containing all text spans for a single case.

- Choosing *Export Task Runs* for Data Hunt projects now exports two files instead of four:

Filename suffix	Description
-DataHuntHighlights.csv.gz	Suffix now *-DataHunt.csv.gz*

-DataHuntSubmitted.csv.gz	Suffix now *-DataHuntByCase.csv.gz*
-Schema.csv.gz	No suffix change. Now exported separately using Generator detail page menu *Export Data Hunt Schema.*
-DataHuntAnswers.csv.gz	No suffix change. Deprecated format offered as a separate export choice under *Export legacy cross tab format for Pybossa Data Hunts* Not available for Mechanical Turk projects. Will be removed in the future.

Hierarchy and repeating rows

TagWorks exports data using CSV formats. Because TagWorks data has several levels of hierarchy in all formats, the data is de-normalized. That is, higher level data structures are repeated in each row as the lowest detail level changes.

For example, a Data Hunt is configured by a Schema that has a four level hierarchy:

1. Schema
2. Topic
3. Question
4. Answer

When a schema is exported, the schema, topic, and question levels are repeated as often as needed to output one or more rows per answer in the schema. An answer will be represented on more than one row if the contributor identified more than one case for the answer, or if discontiguous highlights are being output one per line.

In general, each column of data is either a key or a value that is a function of a key in that row, that is, the value came from the database table and row corresponding to the key.

Schema export file format

This file is exported for Data Hunt Schemas only. The export menu is on the Generator Menu page in TagWorks. Schemas are immutable once they are uploaded, so exporting a schema once is sufficient.

Highlighter schemas are not currently exported because the only information in a Highlighter schema that makes it to the output formats is the “topic_name” string column.

Many of the Data Hunt Schema columns are exported as-is to the Data Hunt exported task runs file formats.

Column name	Type	Example
schema_namespace	Text Usually the source filename.	SemanticsTriager
schema_sha256	SHA-256 of the file uploaded with the schema. Used as unique id externally instead of a uuid.	e534146ed609abf86aa24c7b9776d336315ffd1bd3d4e74ad41fa4ae6e930e35
topic_uuid	uuid	02fa4d97-67eb-4b36-bb24-a5978391f070
topic_name	text	language
topic_options	JSON	{"highlight": true, "version": "4", "hint_type": ""}
question_uuid	uuid	fd82ec78-5df5-4039-bc98-24af5a60f6e0
question_label	text	T1.Q1
question_text	text	To which extent do you think the bolded text is slang?
question_type	one of: RADIO CHECKBOX SELECT_SUBTOPIC TEXT DATE TIME	RADIO
question_hint_type	text	
question_next_questions	text, a list of question labels	
alpha_distance	One of: nominal, ordinal, interval, ratio	ordinal
question_options		{"version": "4", "alpha_distance": "ordinal", "require_one_or_more": true, "hint_type": ""}
answer_count	Number of answers for this question.	5
answer_uuid	uuid for this answer	b843d25f-6b8a-401f-ac3f-2e7d40302fa9

answer_label	text in the format Tx.Qy.Az, where x is the topic number, y is the question number, and z is the answer number.	T1.Q1.A1
answer_content	text	Very likely this is slang
answer_next_questions	text, a list of question labels	T1.Q31, T1.Q32, T1.Q33, T1.Q34, T1.Q91
highlight	boolean: 0 or 1 Derived from answer_options on export	0
require_highlight	boolean: 0 or 1 Derived from answer_options on export	0
answer_options	JSON	{"case_numbers": false, "highlight": false, "version": "4", "require_highlight": false}

HighlighterByCase.csv export format

These columns are used to export contributor data for projects that use the Highlighter presenter.

highlight_task_uuid	This uuid identifies a unique combination of a schema and an article that schema as applied to.	8d8a057b-6be4-4d6b-9c50-f290c34ac1a8
task_url	Text - Only applicable to projects using Pybossa	https://pe.goodlylabs.org/project/NYU_Semantics/task/2287
tua_uuid	Currently unused. In the future, Highlighter tasks will be able to show bolded subsets of the text like Data Hunts. This column will be the uuid of the bolded text. (a text-unit-of-analysis).	

article_batch_name	This is the path that the article was loaded from, either from a zip file directory structure or from S3. The path and filename are not required to be unique. Generators apply researcher supplied regular expressions to this value to select articles.	NYU_Articles/Pilot/FormerUkraineProsecutorSaysHe.txt
article_number	If the filename starts with a number, the system uses that number if it is not already taken, otherwise, the next highest number over 100000 is assigned. This number must be unique.	100028
article_filename	text	FormerUkraineProsecutorSaysHe.txt
article_sha256	SHA-256 of article text	f7eb2314bb13aae4542caf7ee10c336890c3e911c50093d575b1849e107abfb7
article_text_length	integer	8477
destination	Choice field: PYBOSSA or MTURK	PYBOSSA
task_redundancy	Integer - Requested task redundancy	3
taskrun_count	Task runs retrieved so far	1
ah_taskrun_uuid	task run uuid	3d26e90b-5ef0-47a3-8481-eb896cdb629b
contributor_uuid	Contributor uuid	e1ae8875-a398-4dde-8f4e-4b21109784e3
created	ISO date time	2019-10-23 20:55:26.259220
finish_time	ISO date time	2019-10-23 21:56:02.529520
elapsed_seconds	float	3636.2703
hg_tua_uuid	Uuid for this result (varies for each case number)	31cd4587-89e5-4df4-9ebf-5304068ecf3c
namespace	Original filename of schema used for task.	NYU_Semantics

topic_name	String - a topic_name from schema	Language
case_number	integer	1
highlight_count	Integer Number of text spans in this highlight. Span overlap is possible.	5
submitted_tua	JSON	[{"start": 918, "case_number": 1, "end": 926, "text": "vendetta"}, {"start": 975, "case_number": 1, "end": 1033, "text": "Let\u2019s put this through prosecutors, not through presidents"}, {"start": 1112, "case_number": 1, "end": 1134, "text": "just for the interests"}, {"start": 1573, "case_number": 1, "end": 1581, "text": "obsessed"}, {"start": 3114, "case_number": 1, "end": 3121, "text": "dropped"}]

Highlighter.csv export file format

The Highlighter.csv has the same initial columns as the HighlighterByCase format, but unrolls the submitted_tua column into one row per highlight, using the following output columns:

start_pos	integer	918
end_pos	integer	926
target_text	text	vendetta

Notice that in the submitted_tua JSON, the data accessors are ‘start’ and ‘end’, and in the CSV format, it is ‘start_pos’ and ‘end_pos’.

DataHuntByCase.csv export file format

These columns are used to export contributor data for projects that use the Data Hunt presenter.

namespace (**schema_namespace** prior to May 17, 2020)	schema filename	NYU_Reasoning2
schema_sha256	SHA-256	eace698d45562a4832f0753fc291ea8199df836c5c98517a897c6d112e9f1545

quiz_task_uuid	This uuid identifies a unique combination of a schema and an article that schema as applied to.	4f3bca38-79d9-4a37-aebd-db9a33657fe8
task_url	Text - Only applicable to projects using Pybossa	https://pe.goodlylabs.org/project/NYU_Reasoning2/task/2467
tua_uuid	The unique id used by the source tag container to identify the text in the article to show in bold (the text-unit-of-analysis).	896b08c1-6d02-4bd5-886d-5e6045370fc3
article_batch_name	This is the path that the article was loaded from, either from a zip file directory structure or from S3. The path and filename are not required to be unique. Generators apply researcher supplied regular expressions to this value to select articles.	NYU_Articles/Day1/RogerStoneTrialEndsRick.txt
article_number	If the filename starts with a number, the system uses that number if it is not already taken, otherwise, the next highest number over 100000 is assigned. This number must be unique.	100033
article_filename	text	RogerStoneTrialEndsRick.txt
article_sha256	SHA-256 of article text	86d0e1839d825b8b56f2c7430efa716f4777e45dd001dd29ff68a3d232069e4c
article_text_length	integer	2442
destination	Choice field: PYBOSSA or MTURK	PYBOSSA
task_redundancy	Integer - Requested task redundancy	3

taskrun_count	Task runs retrieved so far	5
quiz_taskrun_uuid	task run uuid	1ef11ef2-4e33-42be-91b2-ec9250fc8c00
contributor_uuid	Contributor uuid	2b2f1081-fac7-4884-b56c-28501b89abb8
created	ISO date time	2019-11-13 22:16:46.525040
finish_time	ISO date time	2019-11-13 22:18:17.894846
elapsed_seconds	float	91.369806
topic_name	uuid for this result (varies for each case number)	Reasoning Specialist V4
question_label		T1.Q1
question_text		Does the passage contain... (check all that apply):
answer_label		T1.Q1.A5
answer_content		Arguments or quotes from both sides
answer_uuid		70b43066-c1d4-4b85-935e-62ba5be15578
submitted_tua_uuid		d28fca9a-9359-49ce-82d5-4d80e3270dfc
answer_text		Arguments or quotes from both sides
case_number		1
highlight_count	Integer Number of text spans in this highlight. Span overlap is possible.	2
submitted_tua		[{"case_number": 1, "text": "I do not recall discussing WikiLeaks with him", "answer_id": 4746, "end": 1369, "start": 1324}, {"case_number": 1, "text": "more information would be coming", "answer_id": 4746, "end": 1160, "start": 1128}]

DataHunt.csv export file format

The DataHunt.csv has the same initial columns as the DataHuntByCase format, but unrolls the submitted_tua column into one row per highlight, using the following output columns:

start_pos	integer	1324
end_pos	integer	1369
target_text	text	I do not recall discussing WikiLeaks with him

Notice that in the submitted_tua JSON, the data accessors are ‘start’ and ‘end’, and in the CSV format, it is ‘start_pos’ and ‘end_pos’.

Tag Containers

Each Tag Container exports two files - one ending in *Tags.csv.gz*, and one ending in *NegativeTasks.csv*.

A Tag Container usually represents either the output of a consensus algorithm or the result of expert adjudication. It is possible to import a Tag Container that was generated by some external method.

The *Tags.csv* columns are:

article_batch_name		NYU_Articles/Pilot/FormerUkraineProsecutorSaysHe.txt
article_number		100028
article_filename		FormerUkraineProsecutorSaysHe.txt
article_sha256		f7eb2314bb13aae4542caf7ee10c336890c3e911c50093d575b1849e107abfb7
article_text_length		8477
tua_group_uuid	uuid for this tag container.	1eb2228d-4038-4671-8655-fb1aa9301f91
tua_group_name	Text- the name of the container	NYU_Semantics.adjudicated
tua_batch_uuid	The uuid for a tag batch. Tags are in a batch, and batches are in a tag container. Adjudicator output uses one batch for each task. Consensus algorithms can save tags for many tasks in one batch.	30533b18-8c5c-440b-84d5-9592afd29328

tua_batch_name	A name that indicate the process that generated this TUA, whether algorithmic or adjudication.	Adjudicator nick task 1783 article number 100028
tua_batch_final	Boolean - True if this batch of data is final and can be used to make downstream tasks.	TRUE
source_task_uuid	The highlight_task_uuid or the quiz_task_uuid of the task that collected the source data for this result.	8d8a057b-6be4-4d6b-9c50-f290c34ac1a8
tua_uuid	The uuid for this text-unit-of-analysis, that represents one topic-case number combination, or one answer-case number combination.	e8cd3a4b-7d82-41f0-8eaa-8d701f524c58
namespace	Usually the filename of the Schema used to generate the task that gathered this data.	NYU_Semantics
topic_name	Highlighters will export the text that was shown as the topic to the user. Data Hunts will assign a topic name of the form Tx.Qy.Az, or T1.Q2.A4 to indicate that the highlight is for topic 1, question 2, answer 4.	Language
case_number	integer	1
answer_uuid	Provided if the source task was a Data Hunt. Not applicable to Highlighter projects. This is the unambiguous version of Tx.Qy.Az that implicitly	

	specifies which schema the answer is for.	
extra	JSON. For future use with Gold Standard Trainers.	{}
highlight_count	Integer Number of text spans for this contributor/answer/case combination.	5
start_pos	integer	918
end_pos	integer	926
target_text	text	vendetta

The *NegativeTasks.csv* format has the same initial columns as *Tags.csv*, up to and including **source_task_uuid**. After that column, NegativeTasks has just one additional column, **tua_negative_uuid**.

A row in NegativeTasks indicates that the article has no tags in this container. Absence of highlights in the Tags.csv file can only be used to infer that the article hasn't been processed yet. A row in NegativeTasks means that processing resulted in no tags for the specified article.

The *NegativeTasks.csv* columns are:

article_batch_name		
article_number		
article_filename		
article_sha256		
article_text_length		
tua_group_uuid		
tua_group_name		
tua_batch_uuid		
tua_batch_name		
tua_batch_final		

source_task_uuid		
tua_negative_uuid		