accenture

Data Labelling Manual

Turing

myWizard® 360°

Table of Contents

Revision History

Date	Version	Description	Author
7-1-2018	Version 0.1	Initial document	Reuben
8-1-2018	Version 0.2		

Table 1 - Revision History Table

Glossary

Word	Meaning
MTurk	Amazon **M**echanical **Turk.**
Micro-Service	Multiple sub components that constitute an application as a single program.
Data Labelling	A microservice for Turing used to acquire data labels using MTurk, for Machine Learning.
Data Management	A microservice in Turing to store and manage all the data for further processes, such as providing a dataset for Machine Learning training.
Machine Learning Core	A microservice responsible for all machine learning related tasks, such as model training, model prediction and so on.
HIT	Human Intelligence Task, basic tasks to be completed by Mturk workers. Often a reading a number or handwriting and documenting an answer.
Data Label	A machine learning concept for supervised learning where every training example needs to have an accompanying label i.e a desired training goal.
OCR	Optical Character Recognition: a form of text recognition

Table 2 - Glossary Table

1 Introduction

Data Labelling is a Turing micro-service that provides users with a streamline method of acquiring data labels for machine learning. Data Labelling utilizes the Amazon web services Mturk application to efficiently and rapidly acquire data labels.

2 What is MTurk and HIT.

MTurk or Amazon Web services **m**echanical **turk** services is a method of crowdsourcing and distributing processes or jobs. MTurk distributes work to people all over the world to perform a set of required tasks. It is a platform to hire a large-scale temporary workforce for simple tasks and process. Turing utilises this service to provide low-cost, high accuracy data labels to construct machine learning models. These models are what allows the Turing Orchestrator to automatically interpret forms in an adaptable and configurable method. Whilst MTurk is a fast and efficient method to labelling datasets it is not required. Should the data be sensitive or confidential companies can opt to manually label the dataset internally instead. This can be done using the Data Management micro-service.

Human Intelligence Tasks (HIT) are the currency of the MTurk marketplace. They are essentially the tasks that workers complete for a request. For example, a task might be reading a few snippets of a handwritten form that a worker reads and provides their interpretation for. HITs exist in 5 states:

- **Assignable:** a HIT can be accepted for work by workers.
- **Unassignable:** a HIT cannot be worked on any more
- **Reviewable:** all workers have submitted completed answers for a HIT. These can now be reviewed
- **Reviewing:** the requester is Reviewing the responses to the HIT
- **Disposed:** the HIT has been deleted and cannot be retrieved

For successfully completing a HIT, workers receive a small payment often in the range of a few US cents.

2.1 Justification

For the models to be constructed the Turing machine learning core needs labelled data to train on. This is where the power of MTurk is utilised. For a relatively small cost, MTurk distributes the process of labelling this data for virtual classification. That is, across the world humans perform the task of manually reading the snippets of extracted numbers and names from forms and entering a human interpretation of that snippet. These human interpretations are the labels that the Machine Learning Core uses as true labels and learns a method of interpreting the snippets so that they closely match the human derived value. This process would typically take a large temporary workforce, however MTurk makes this accessible and easy for many businesses.

The value of this processing overhead comes from the fact that once a dataset has been labelled, any further Machine Learning Models can be retrained using this dataset, as long as the labelled dataset is the same type of data.

3 The MTurk Main Page

To begin this process, the user must first log into the main Turing page. After logging into the main Turing page, users can access the Mturk management page using the navigation icon.

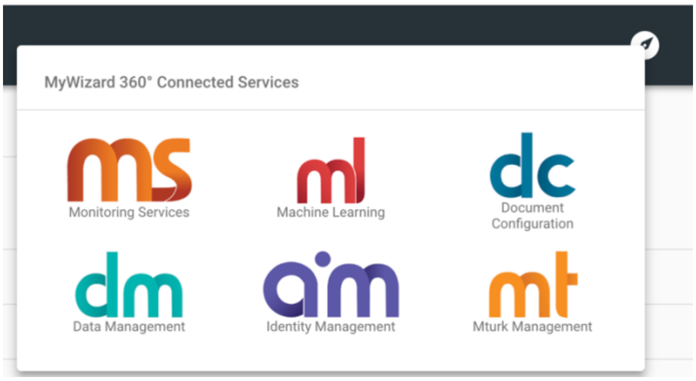
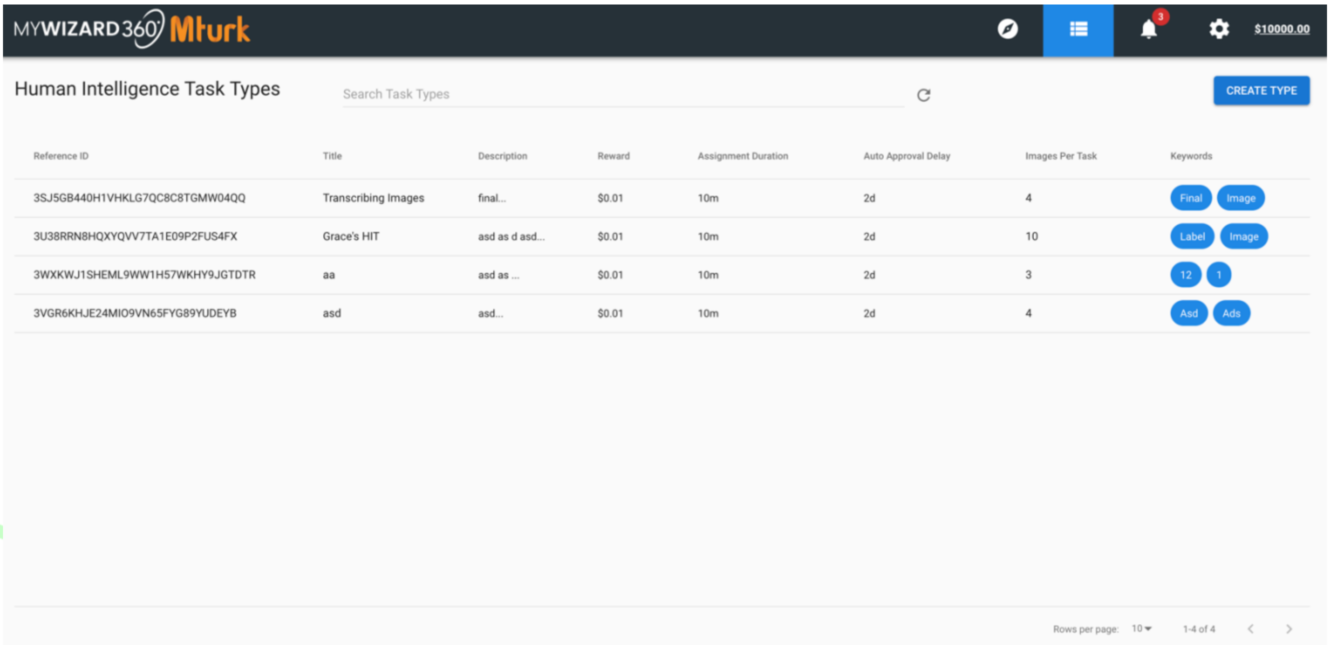


Figure 1 Navigating to MTurk

This lands users upon the main MTurk management page. From here users can create a new HIT type and dispatch batches of work to be completed by workers from the selected countries.



Human Intelligence Task Types

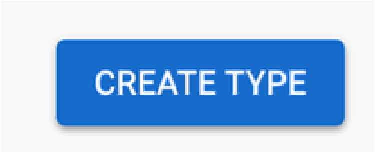
Reference ID	Title	Description	Reward	Assignment Duration	Auto Approval Delay	Images Per Task	Keywords
3SJ5GB440H1VHKLG7QC8C8TGMW04QQ	Transcribing Images	final...	$0.01	10m	2d	4	Final Image
3U38RRN8HQXYQVV7TA1E09P2FUS4FX	Grace's HIT	asd as d asd...	$0.01	10m	2d	10	Label Image
3WXKWJ1SHEML9WW1H57WKHY9JGTDTR	aa	asd as ...	$0.01	10m	2d	3	12 1
3VGR6KHJE24MIO9VN65FYG89YUDEYB	asd	asd...	$0.01	10m	2d	4	Asd Ads

Rows per page: 10 ▾   1-4 of 4

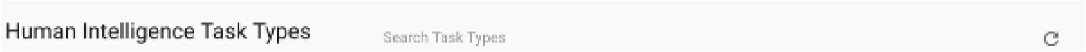
Figure 2 Main MTurk Page

4 Creating a new HIT

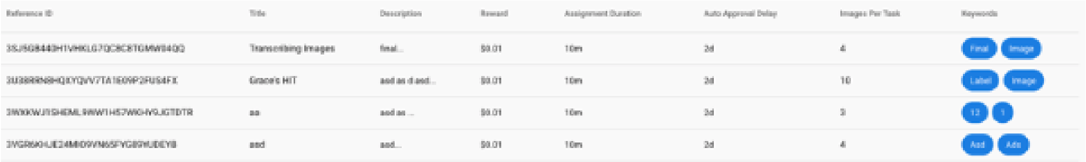
First users must click the CREATE TYPE button. This will open an expandable side panel.



New HIT creation Button



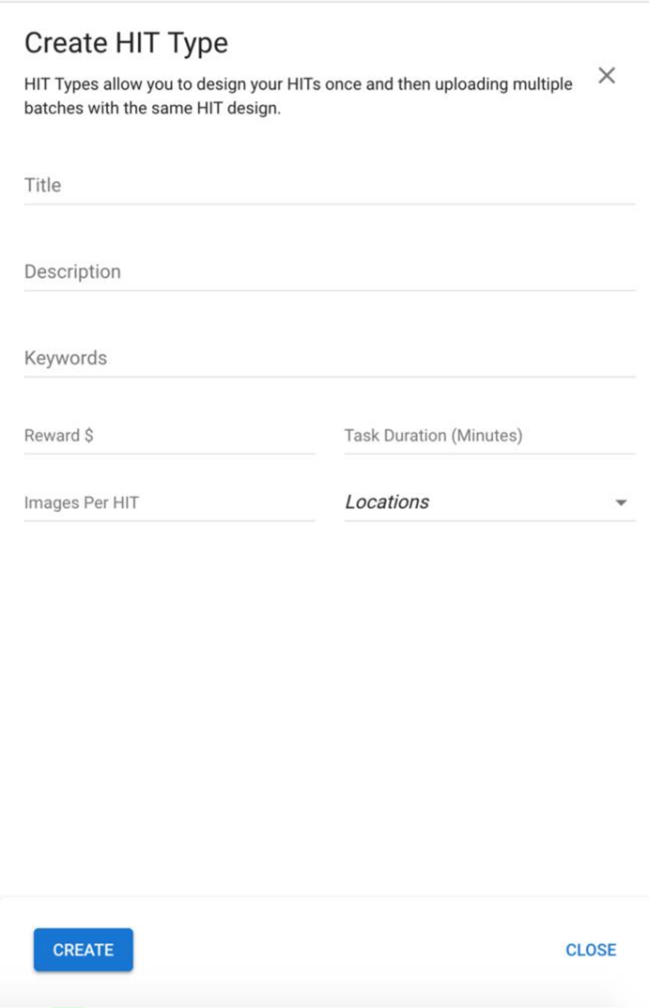
Search bar for created HITS



Reference ID	Title	Description	Reward	Assignment Duration	Auto Approval Delay	Images Per Task	Keywords
3SJ5GB443H1VHKLG7QC8C8TGMW84QQ	Transcribing Images	final...	$0.01	10m	2d	4	Final Image
3U38RRN8HQXYQVV7TA1E09P2FU54FX	Grace's HIT	asd as d asd...	$0.01	10m	2d	10	Label Image
3WKKWJ1SHEML9WW1H57WKHV9JGTDTR	aa	asd as ...	$0.01	10m	2d	3	12 1
3YGR6KHUE24MIO9VN65FYG89UDEY8	asd	asd...	$0.01	10m	2d	4	Asd Ads

Created HIT list

3 Main Features from MTURK page



Create HIT Type

HIT Types allow you to design your HITs once and then uploading multiple batches with the same HIT design.

Figure 4 New HIT creation panel

To create a new HIT, users must provide a:

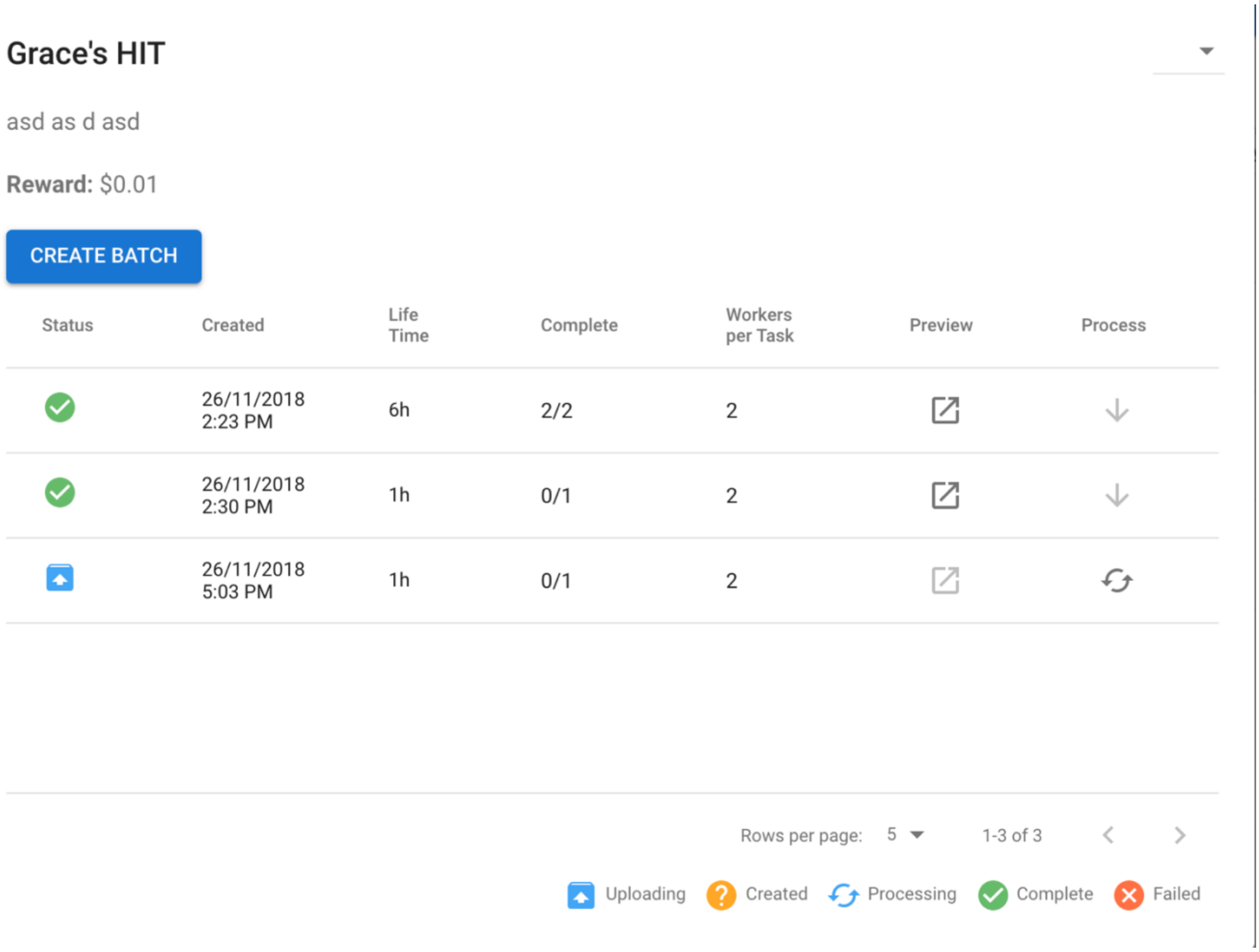
- Title for the HIT
- Short description
- Key words for searching and identifying HIT
- Reward (how much workers earn for completing the HIT)
- Duration of the HIT, how long it will take workers to complete typically
- Number if images in each HIT
- Specify from which countries MTurk workers will be offered the job.

The minimum reward offered for a HIT is 1 cent or $0.01, a new HIT type cannot be created if the offered reward is lower than this.

After entering these details, the new HIT will appear in the created HIT list. If users have created a large number of HITs, they may need to navigate to further pages using the navigation buttons at the bottom of the page. Alternatively, users can search for the HIT keywords in the HIT search bar to narrow down the created HIT list to show just those with the searched terms.

5 Dispatching HITs

Creating the HITS is just the first step in obtaining data labels using MTURK. After finding the desired HIT from the list shown clicking on the desired HIT will open the HIT batch panel. Here batches can be created, or previous batches viewed.



Grace's HIT

asd as d asd

Reward: $0.01

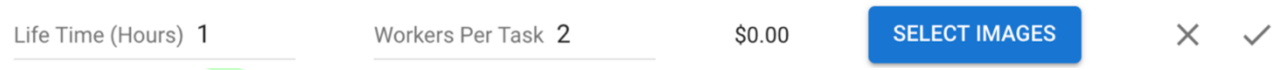
Status	Created	Life Time	Complete	Workers per Task	Preview	Process
✓	26/11/2018 2:23 PM	6h	2/2	2	↗	↓
✓	26/11/2018 2:30 PM	1h	0/1	2	↗	↓
⇧	26/11/2018 5:03 PM	1h	0/1	2	↗	⟳

Rows per page: 5 ▾ 1-3 of 3 ‹ ›

⇧ Uploading ? Created ⟳ Processing ✓ Complete ✕ Failed

Figure 5 HIT Batch Panel

In this panel there will be the name of the HIT, its description and the HIT reward. Clicking the CREATE BATCH button will open a number of new options such as the number of hours the task will be available for workers to complete and the number of workers required to perform each the task. Assigning multiple workers to one task will mean that the results can be cross validated. The more workers assigned to each task means there will be more internal validation and therefore higher accuracy in the resulting dataset. This means the Machine Learning Model will have a more accurate set of labels to train off, the final result being the main Turing process will perform more accurately. Note that assigning 3 workers is the minimum required to converge on majority vote consensus, which provides high levels of accuracy.



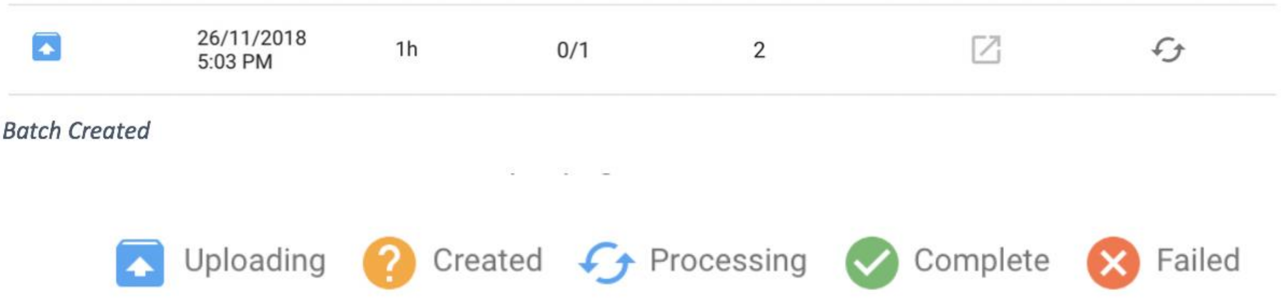
Life Time (Hours) 1

Workers Per Task 2

$0.00

Figure 6 Additional Create Batch options

After assigning a duration and number of workers, the user can select the images to be processed and viewed by the workers using the SELECT IMAGES button. This will open a directory to select the files.



26/11/2018 5:03 PM 1h 0/1 2

Batch Created

Batch Status Legend

Figure 7 Batch Information

The status of each batch created is shown by the status icon. To dispatch an uploaded source of images, click the far-right circular dispatch icon. This will send out the HIT to be processed.

A batch of images does not have to be manually chosen from a directory. By choosing the MTurk labelling option for document configurations in the Data Management Microservice (see Data Management User for more details), users will be redirected to the Data Labelling Microservice, this will be very similar to the process described above. Firstly, users will be asked to select or create a HIT type. Creating a batch is the same process, however the user will not be prompted to provide a directory of images for labelling. The Data Management microservice will provide the segmented images to the Data labelling Microservice.

Final part of MTurk microservice incomplete. Further revision required.

6 Monitoring Progress

Since the time taken to receive completed tasks in MTurk is undetermined, it can be difficult to track multiple task flows when there are multiple datasets in queue.

Users can use the notification icon (‘the bell’ in the header menu) to check on the creation of HITs or the progress of labelling.

Click on the HIT type to show the list of batches under this HIT. Clicking on the symbol in the preview column will open a new tab which shows a preview of what MTurk participants see when they accept the task.

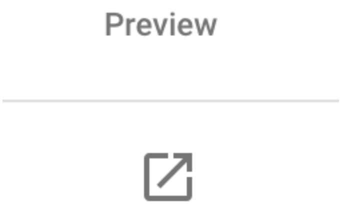
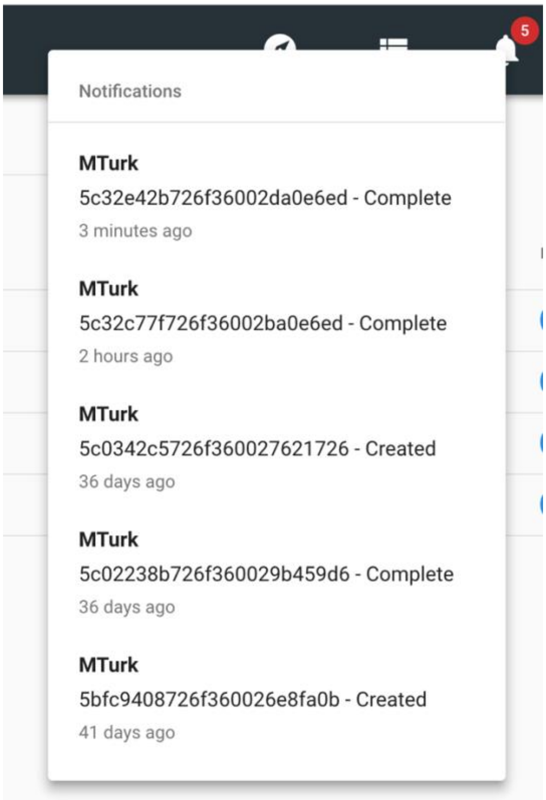


Figure 8 Batch Preview



Notifications

MTurk
5c32e42b726f36002da0e6ed - Complete
3 minutes ago

MTurk
5c32c77f726f36002ba0e6ed - Complete
2 hours ago

MTurk
5c0342c5726f360027621726 - Created
36 days ago

MTurk
5c02238b726f360029b459d6 - Complete
36 days ago

MTurk
5bfc9408726f360026e8fa0b - Created
41 days ago

Figure 9 MTurk Notifications Menu

This functionality can be used to track and manage different dispatches across multiple datasets.

7 Using the results

The final data labels are automatically routed to Data Management where they are applied to the documents. Refer to the Data Management User Manual for more details.

8 Contact Information

If you have any questions or queries, do not hesitate to contact the following people:

Riley Green
Email: riley.green@accenture.com
Mobile: +61 487 317 143
Address: International House, 3 Sussex St, Barangaroo NSW 2000

Xiwen Sun
Email: xiwen.sun@accenture.com
Mobile: +61 398 388 954 Address:
Melbourne Level 5, 161 Collins Street