Use Cases: Test Plan
First Evaluation 2015

Deliverable Nr Title:	Deliverables 7.3.1 & 8.3.1 Use Cases: Test Plan (First Evaluation)
Delivery Date:	May 2015
Author(s):	Chris Lintott, Grant Miller and Alex Bowyer (University of Oxford - Zooniverse) Marcello Colacino, Piero Savastano, David Riccitelli, and Andrea Volpini (InsideOut 10)
Publication Level:	Public

Table of Contents

Documentation Information

Project (Title/Number)	MICO - “Media in Context” (610480)
Work Package / Task	Work Package 7 - Use Case: Crowd Sourcing Platform Work Package 8 - Use Case: Video Sharing Platform
Responsible person and project partner	Grant Miller (University of Oxford - Zooniverse) Andrea Volpini (InsideOut10)

Executive Summary

This document outlines the Test Plan (First Evaluation) describing the setup of the first evaluation round in the two use cases (WP7 and WP8) and the functionalities that are evaluated. It is currently restricted to the testing conducted during the development of the MICO platform. The aim is to compare each MICO Technology Enabler (TE) prior to beginning end-to-end testing of the system.

For each MICO TE included in these tests we will assess the following:

1. **output accuracy** - how accurate, detailed and meaningful each single response is when compared to our best estimates using our existing databases and analysis;
2. **technical performance** - how much time each task requires and how scalable the solution is when we increase in volume the amount of contents being analysed;
3. **usability** evaluated both in terms of **integration, modularity** and **usefulness**.

Note that a low score on these metrics does not indicate a failure. It is important to collect these metrics to correctly understand what the system does and does not do - but this does not constitute a success/fail judgement of the platform as a whole.

A: Zooniverse Showcase - Test Plan

References

Reference ID	Link
REF-PR	http://en.wikipedia.org/wiki/Precision_and_recall
REF-SS	http://www.snapshotserengeti.org/
REF-AGG	https://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/view/9431
REF-F1	http://en.wikipedia.org/wiki/F1_score

Definitions

Term	Definition
Precision	In pattern recognition and information retrieval with binary classification, precision is the fraction of retrieved instances that are relevant. In a classification task, a precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labeled correctly). **[REF-PR]**
Recall	In pattern recognition and information retrieval with binary classification, recall is the fraction of relevant instances that are retrieved. In a classification task, a recall score of 1.0 means that every item from class C was labeled as belonging to class C (but says nothing about how many other items were incorrectly also labeled as belonging to class C). **[REF-PR]**
F-measure	In binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r **[REF-PR]** of the test to compute the score. **[REF-F1]**

Background

This document provides prerequisites and guidelines for the test planning, test design, test implementation, test execution and test evaluation processes. It is currently restricted to the testing conducted during the development of the MICO platform.

The test planning comprises two scenarios:

1. **Technical validation** - this will be performed on staging using real contents coming from production environments. At this step, our goal is to check that everything works as expected and that all functional requirements are properly implemented. Once we've checked that MICO outputs are consistent we can move to the production environments and begin testing with the real end-users;

2. **Real-world evaluation** - this will be performed in production. Assuming the MICO platform becomes stable and can be integrated in our application workflows, we will start the evaluation phase using MICO outputs. A/B tests will be designed and performed and all involved KPI will be monitored and compared. Note that since this is a research project, we do not expect that the platform has to succeed a real world evaluation to be deemed a success. This additional testing will help identify areas for future improvement and will not be part of any success assessment for the project.

This document focuses on the technical validation of the MICO platform. All suggested tests consider the constraints / limitations of the ongoing development.

It is applicable for the development of systems undertaken by MICO, and to the organisation that assumes the responsibility for its implementation and testing.

Goals

We aims to compare each MICO TE prior to beginning end-to-end testing of the system. For each MICO TE included in these tests we will assess the following:

4. **output accuracy** - how accurate, detailed and meaningful each single response is when compared to our best estimates using our existing databases and analysis;
5. **technical performance** - how much time each task requires and how scalable the solution is when we increase in volume the amount of contents being analysed;
6. **usability** evaluated both in terms of **integration, modularity** and **usefulness**.

Note that a low score on these metrics does not indicate a failure. It is important to collect these metrics to correctly understand what the system does and does not do - but this does not constitute a success/fail judgement of the platform as a whole.

Methodology and planning

Output accuracy

In order to validate MICO TE output accuracy we follows these steps:

1. **accuracy definition** - define what *accuracy* means within each single TE scope (end-user expectation has to be considered / included in the accuracy definition);
2. **accuracy KPI** - define a set of KPI that can be used to measure TE accuracy and compare the results with other data;
3. **borderline cases** - identify borderline critical cases that could require dedicated validation tasks with ad hoc datasets.

TE	Accuracy Definition	Accuracy KPI	Borderline
TE-202	**Emptiness** - Properly determine whether or not a subject contains animals	• **precision [REF-PR]** here defined as the number of properly categorized	**day/night** - Validate emptiness, animal count and animal type on

	Animal Count - Properly determine how many animals are present **Animal Type** - Correctly identify animal type. The identified animal types are: cat-like, dog-like, pig-like, hoofed, birds, monkeys, giraffes, elephants, humans	subjects (for emptiness, counts, types) compared to the number of subjects analyzed	both daytime and nighttime images **simple/complex** - Validate emptiness, animal count and animal type on images with >5 animals as well as those with 0 or 1. **animal size** - Validate emptiness, animal count and animal type on images with different sizes of animal - giraffe/elephant, large ground animals (e.g. hippo), deer/cattle, big cats, rodents) **mixed animal types** - Validate results for images containing species from more than one of the identified "animal types"
TE-501	**N/A** - Zooniverse internal, not part of MICO platform	**N/A**	**N/A**
TE-506 (WP5)	**Recommendation** - Provide cross-media contextual content suggestions	● **precision [REF-PR]** here defined as the number of recommendations that are good recommendations. ● **recall [REF-PR]** the proportion of good recommendations that appear in top recommendations. ● **F1-measure**	**Changed behaviour -** Validate that if a user's preference changes subsequently to initial training, that the recommendation will adapt accordingly when provided with new user behaviour data.

Accuracy test planning will require these preparatory steps:

1. define datasets for each test;
2. identify alternative ways to evaluate accuracy, either using Zooniverse internal data or scripts, or by finding third-party solutions to be used as a benchmark in accuracy evaluation;
3. perform dataset normalization tasks if required;
4. define useful algorithms;

Accuracy test datasets

ID	Dataset Description
TD-12	a dataset of around 300,000 images from Univ. of Minnesota/Snapshot Serengeti **[REF-SS]** containing a mixture of different species, numbers of animals, and blank/non-blank images (also known as "Season 8"). All of these images have been fully classified by the crowd therefore we know with confidence which species, how many animals and which ones are empty.

TD-13	a dataset of ~11 million user events collected by Geordi, the Zooniverse analytics collector (TE-501) over the period February-May 2015.
TD-14	a set of user profiles, derived by script from TD-13, which indicate which species each user prefers, based on their past shares, favourites and other indicators of interest
TD-15	a modified version of TD-14, where the favorited animals are switched for at least 50% of the users.
TD-16	Derived from TD-12, a mapping of subject > species present, for verifying results.
TD-19	a dataset of 20,000 images from Univ. of Minnesota/Snapshot Serengeti **[REF-SS]** containing ~20,000 images which have never been classified (also known as “the lost season”)

Benchmark solutions

TE	Benchmark solutions
TE-202	For TD-12 dataset, we already know from consensus which species are present, and how many animals there are, and which images are empty. Therefore we can use our consensus data as control for measuring accuracy.
TE-506	We will benchmark the performance of the recommendation engine against our own programmatic user profile generator (as detailed in Architecture section). The results should be equal or better.

Dataset Normalization

TD	Normalization required
TD-12	.We will create indexes & subsets of this dataset to easily identify: ● emptiness of image ● number of animals (per consensus) ● species present (per consensus) ● animal types present (per the list of animal types identified in TE-202 and summarized in the table above) ● borderline groupings as per table in 4.1. above (simple/complex, etc)
TD-13>TD-14	TD-14 will be generated from TD-13 using pre-existing Zooniverse scripts. This organizes the data by mapping user (ID or IP address) to a list of liked species.
TD-14>TD-15	TD-15 will be generated by programmatically and randomly rotating each species to a different species for 50% of the users, so that those users all are now recorded as liking different animals.
TD-16	From the normalized TD-12 data, we will also generate an index showing which species are present in each subject. Again this will be done using pre-existing Zooniverse scripts.

Algorithms

We will use established techniques of aggregation when analysing TD-12 to determine the classified contents of subjects, as described in the paper “Aggregating User Input in Ecology Citizen Science Projects” **[REF-AGG].**

For the other normalizations, we use existing Zooniverse scripts. In the case of day/night detection we will write a new algorithm using timestamp data.

Technical performance

Technical performance will be measured in terms of:

1. **latency** - time required to perform a single task on a given dataset. Measures will be repeated 10 times;
2. **scalability** - an assessment of whether the given TE is suitably efficient and practical when applied to large situations (e.g. a large input dataset and / or, a large number concurrent requests)

Usability

The TE usability requires a qualitative evaluation which will consider:

- **integration** - how simple it is to integrate each single TE into Zooniverse technologies;
- **modularity** - how simple it is to configure each TE and/or a combination of multiple TEs in a chain from within pre-existent application workflows.
- **usefulness** - looking at the degree to which the TE delivers valuable information and tools that Zooniverse applications will be able to harness in future, and some consideration towards a cost-benefit analysis of doing so.

Test Implementation

ID	Test	Description
TP-202-01	**Test emptiness detection across a season of subjects**	• use dataset **TD-12** • process **TD-12** with TE-202(emptiness detector) • calculate precision, recall and F1-measure on **TE-202** outputs. • calculate manual blanks, for **TD-12** per normalization procedure. • calculate precision, recall and F1-measure on outputs. • compare **KPIs**
TP-202-02	**Test animal type detection across a season of subjects**	• use dataset **TD-12** • process **TD-12** with TE-202(group detector) • calculate precision, recall and F1-measure on **TE-202** outputs. • calculate manual animal types present, for **TD-12** per normalization procedure. • calculate precision, recall and F1-measure on outputs. • compare **KPIs**
TP-202-03	**Test animal counting across a season of subjects**	• use dataset **TD-12** • process **TD-12** with TE-202(animal detector) • calculate precision, recall and F1-measure on **TE-202** outputs. • calculate manual animal counts for **TD-12** per normalization procedure. • calculate precision, recall and F1-measure on outputs. • compare **KPIs**

TP-506-01	**Test species recommended in recommended images**	• use dataset **TD-14** • Train **TE-506** (recommendation engine) with TD-14. • calculate precision, recall and F1-measure on **TE-506** outputs. • Process results from TE-506 per user against TD-16, and identify distinct species that were recommended to each user. • Compare species recommendation to TD-14 preferences. • Assess KPIs
TP-506-02	**Test subjects recommended**	• use dataset **TD-14** • Train **TE-506** (recommendation engine) with **TD-14**. • calculate precision, recall and F1-measure on **TE-506** outputs.
TP-506-03	**Test subjects recommended**	• use dataset **TD-15** • Train **TE-506** (recommendation engine) with **TD-15**. • compare recommended subjects to those in TP-506-02 - ensure the correct species are displaced according to the changes that were made between TD-14 & TD-15 • assess **KPIs**
TP-506-04	**Real-world tests using recommender results**	• use dataset **TD-11** • Run a full experiment over a period of days/weeks, per the Happy User Experiment plans, inserting recommended images per user. • Evaluate session times and number of classifications for experimental vs control user. • Determine whether the recommended images increased or decreased user participation. • assess **KPIs**

B: Use Cases: First Prototype - Video News Showcase - Test Plan

References

Reference ID	Link
REF-PR	http://en.wikipedia.org/wiki/Precision_and_recall
REF-WER	http://en.wikipedia.org/wiki/Word_error_rate
REF-F1	http://en.wikipedia.org/wiki/F1_score
REF-LEV	http://en.wikipedia.org/wiki/Levenshtein_distance
REF-ERS	http://aimotion.blogspot.it/2011/05/evaluating-recommender-systems.html
REF-DAM	http://photo.greenpeace.org
REF-BS-01	http://betafaceapi.com/demo.html
REF-MAG	http://mag.greenpeace.it
REF-BS-02	https://rekognition.com/demo/face
REF-BS-03	http://www.ispeech.org/
REF-LI	https://loadimpact.com/api-testing
REF-AL-01	http://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance#PHP

Definitions

Term	Definition
Precision	In pattern recognition and information retrieval with binary classification, precision is the fraction of retrieved instances that are relevant. In a classification task, a precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labeled correctly). **[REF-PR]**
Recall	In pattern recognition and information retrieval with binary classification, recall is is the fraction of relevant instances that are retrieved. In a classification task, a recall score of 1.0 means that every item from class C was labeled as belonging to class C (but says nothing about how many other items were incorrectly also labeled as belonging to class C). **[REF-PR]**

Word Error Rate	Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system. It is derived from the Levenshtein distance - where the distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other - working at the word level instead of the phoneme level. **[REF-WER]**
F-measure	In binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r **[REF-PR]** of the test to compute the score. **[REF-F1]**
ASR	In computer science, **speech recognition** (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT).

Background

This document provides prerequisites and guidelines for the test planning, test design, test implementation, test execution and test evaluation processes. It is currently restricted to the testing conducted during the development of the MICO platform.

The test planning comprises two scenarios:

1. **Technical validation** - it will be performed on staging using real contents coming from production environments. At this step, our goal is to check that everything works as expected and that all functional requirements are properly implemented. Once we've checked that MICO outputs are consistent we can move to the production environments and begin testing with the final end-users;
2. **On field evaluation** - it will be performed in production. Assuming the MICO platform becomes stable and can be integrated in our application workflows, we will start the evaluation phase using mocked / static MICO outputs. A/B tests will be designed and performed and all involved KPI will be monitored and compared.

This document focus on the technical validation of the MICO platform. All suggested tests consider the constraints / limitations of the ongoing development.

It is applicable for the development of systems undertaken by MICO, and to the organisation that assumes the responsibility for its implementation and test.

Goals

We aim to compare each MICO TE prior to begin testing end-to-end the system. For each MICO TE included in these tests we're willing to asses:

1. **outputs accuracy** - how much each single response is accurate, detailed and meaningful in terms of precision/recall and eventually where MICO TE stand when compared to other similar technologies available in the market;

2. **technical performances** - how much time each single task requires and how scalable the solution is when we increase in volume the amount of contents being analysed;
3. **usability** evaluated both in terms of **integration** and **modularity**.

Methodology and planning

Outputs accuracy

In order to validate MICO TE outputs accuracy we follows these steps:

1. **accuracy definition** - define what *accuracy* means within each single TE scope (end user expectations has to be considered / included in the accuracy definition);
2. **accuracy KPI** - define a set of KPI that can be used to measure TE accuracy and compare the results with existing alternatives ones;
3. **borderline cases** - identify borderline critical cases that could require dedicated validation tasks with ad hoc datasets.

TE	Accuracy Definition	Accuracy KPI	Borderline
TE-204	**Face Detection** - Properly detect faces on images and video	• **precision [REF-PR]** here defined as the number of properly detected faces compared with the number of faces matches. • **recall [REF-PR]** here defined as the number of properly matched faces compared with the total number of faces in the dataset • **F1-measure [REF-F1]**	**Face orientation** - Validate face detection on a dataset including non-frontal faces. **Face size** - Validate face detection on a dataset where faces are smaller than X pixels.
TE-214	**Speech to text** - Extract meaningful text transcriptions from video contents.	• **WER [REF-WER]**	**Noise** - Validate speech to text on noisy videos **Language** - Validate speech to text on different languages domain from standard english (Italian and Arabic)
WP5	**Recommendation** - Provide cross-media contextual content suggestions	• **precision** here defined as the number of recommendations that are good recommendations[1].	**Small datasets** - Validate WP5 output accuracy for small and heterogeneous dataset (ea: “Greenpeace News” contents)

[1] The most highly preferred items in the test set are the good recommendations, and the rest aren't. It is important to give an threshold that divides good recommendations from bad ones - see **[REF-ERS]** for more details about precision & recall usage on recommendation engine evaluation. Given a rating range from 1 to 5 where 5 means “absolutely preferred” we define a rating of 4 (four) as threshold for a good recommendation. See § 4.1.3 about dataset normalization required for this test.

		● **recall** the proportion of good recommendations that appear in top recommendations. ● **F1-measure**	

Accuracy test planning will require these preparatory steps:

1. define datasets for each test;
2. identify alternative third solutions in the market to be used as benchmark in accuracy evaluation;
3. perform dataset normalization tasks if required;
4. define useful algorithms;

Accuracy test datasets

ID	Dataset Description
TD-01	a dataset of 50 images from **Greenpeace photo archive [REF-DAM]** containing faces from a frontal perspective.
TD-02	a dataset of 50 images from **Greenpeace photo archive [REF-DAM]** containing faces **not** from a frontal perspective.
TD-03	a dataset of 50 textual / multimedia assets coming from “Greenpeace News” digital magazine **[REF-MAG]** properly annotated.
TD-04	a dataset of 500 randomly selected “Greenpeace News” **[REF-MAG]** magazine profled users with their interactions on D-03 datasets items.
TD-05	a dataset of 10 video from Greenpeace video archive **[REF-MAG]** containing voiceover in native English without noise and / or music. Related video voiceover transcriptions to use as reference.

Benchmark solutions

TE	Benchmark solution
TE-204	BetafaceApi **[REF-BS-01]**, Rekognition.com **[REF-BS-02]**
TE-214	iSpeech **[REF-BS-03]**

Dataset Normalization

TD	Normalization required
TD-04	TD-04 is a set of real profiled users along with their interactions on TD-03 items. These info are tracked via / stored on Google Analytics. The only interaction we store at the moment is the simple page / content view. We need to transform the pageview in a rating with a range value from 1 to 5 considering the time spent on the page by the user: more time a single user

	spent on a given page, more he likes the page itself. We will consider 20, 40, 60, 80 percentile of time spent distribution to define thresholds useful to convert time in ratings.

Algorithms

For Word Error Rate calculation we will start from standard Levenshtein distance implementation. **[REF-AL-01]**

Technical performances

Technical performances will be measured in terms of:

1. **latency** - time required to perform a single task on a given dataset. Measures will be repeated 10 times;
2. **scalability** - it the given TE suitably efficient and practical when applied to large situations (e.g. a large input dataset and / or, a large number concurrent requests). External platform as LoadImpact **[REF-LI]** could be used.

Usability

The TE usability requires a qualitative evaluation which will consider:

- **integration** - how simple it is to integrate each single TE within pre-existent application workflows;
- **modularity** - how simple it is to configure each TE and/or a combination of multiple TEs in a chain from within pre-existent application workflows.

Test Implementation

ID	Test	Description
TP-204-01	**Test in-front face detection on images**	• use dataset **TD-01** • process **TD-01** with TE-204 • calculate precision, recall and F1-measure on TE-204 outputs. • process **TD-01** trough **BetaFaceApi** [REF-BS-01] • calculate precision, recall and F1-measure on TE-204 outputs. • compare **KPIs**
TP-204-02	**Test lateral face detection on images**	• use dataset **TD-02** • process **TD-02** with TE-204 • calculate precision, recall and F1-measure on TE-204 outputs. • process **TD-02** trough **BetaFaceApi** [REF-BS-01] • calculate precision, recall and F1-measure on TE-204 outputs. • compare **KPIs**
TE-214-01	**Test ASR on videos containing voiceover in english without noise and/or music**	• use dataset **TD-05** • process ASR on **TD-05** via TE-214 • calculate **WER** on ASR results transcriptions in D-05 as reference. • process ASR on **TD-05** via **iSpeech** [REF-BS-03] • calculate **WER** on ASR results transcriptions in D-05 as reference. • compare **KPIs**

TP-WP05-01	**Item similarity based use case test**	• use dataset **TD-03** • train WP5 on **TD-03** • query WP5 to generate recommendation for each item in D-03 • calculate precision, recall [REF-PR] and F1-measure [REF-F1] for each WP5 outputs.
TP-WP05-02	**User based use case test**	• use dataset **TD-04** • train WP5 on **TD-04** - use 70% as training set, 30% as test set • calculate precision, recall [REF-PR] and F1-measure [REF-F1].