Change Log

Version	Date	Changes
1.21	2019-06-12	- Updated section 9 Schedule to extend system description due date to July 12, 2019
1.20	2019-04-30	- Updated section 8 Prize to remove winner MATERIAL PI meeting invitation
1.19	2019-04-02	- Updated section 6.1 Submission Limits and Feedback - Added table comparing original vs. revised schedule and details
1.18	2019-03-18	- Updated section 6.1 Submission Limits and Feedback - Updated section 8 Prize with updated baseline score thresholds - Updated section 9 Schedule with revised dates
1.17	2019-02-19	- Removed mention of distractor documents/languages in section 4.3.3 Evaluation
1.16	2019-02-13	- Added link to Query Relevance guidelines to section 3 The Main Scoring Idea: A Detection System - Added system description length requirement in section 7 System Description - Added baseline score thresholds to section 8 Prize
1.15	2019-02-12	- Added section 4.3.1.1 Dry Run on DevTest - Added link to prize challenge rules and regulations document in section 8 Prize - Updated section 9 Schedule with revised dates
1.14	2019-01-28	- Fixed typo and revised notations for Equation 2 and Equation 3 in section 3.1 The Main Detection Metric: AQWV (Actual Query Weighted Value) for accuracy and clarity
1.13	2018-11-29	- Updated section 5.2 System Output Format to indicate the filename does not need “q-” - Updated section 5.3 Reference Format to indicate the filename does not need “q-”
1.12	2018-11-13	- Updated section 4.5 Data Usage Restrictions - Updated section 6.1 Submission Limits and Feedback
1.11	2018-10-18	- Updated section 3.1 The Main Detection Metric: AQWV (Actual Query Weighted Value) to include modified AQWV - Added section 4.2 Additional Training Data - Updated rules in section 4.5 Data Usage Restrictions
1.10	2018-08-28	- Updated section 5.2 System Output Format

		● Updated section 5.3 Reference Format ● Updated section 6.2 Evaluation Submission Format
1.9	2018-08-15	● Updated section 4.5 Dataset Structure ● Updated section 6.2 Evaluation Submission Format ● Updated section 8 Prize ● Updated 6.1 Submission Limits and Feedback to remove submission limit for development cycle ● Extended registration period to November 30, 2018
1.8	2018-08-09	● Added section 8 Prize
1.7	2018-07-12	● Initial public version

Original vs. Major Revision Schedule and Details

Item	**Original versions (Pre February 2019)**	**February 2019 Revision**	**March 2019 Revision**	**Section**
Evaluation period	1 week Jan 28 - Feb 1, 2019	1 week March 11-15, 2019	12 weeks March 11 - May 31, 2019	9 Schedule
Evaluation set submission limits	1	1	1/week	6.1 Submission Limits and Feedback
Evaluation set score feedback	yes	yes	yes, only on overall score	6.1 Submission Limits and Feedback
AQWV threshold to be eligible to win	Text: TBD Speech: TBD	Text: ≥ 0.3 Speech: ≥ 0.2	Text: ≥ 0.2 Speech: ≥ 0.1	8 Prize
Winner MATERIAL PI meeting invitation	yes	yes	no	8 Prize

OpenCLIR 2019 Evaluation Plan

(Open Cross Lingual Information Retrieval 2019 Evaluation)

Contents

1 Introduction

The goal of the first OpenCLIR (Open Cross Language Information Retrieval) evaluation is to develop methods to locate content in speech or text documents in low-resource languages, using English queries. This capability is one of several expected to ultimately support effective triage and analysis of large volumes of data, in a variety of less studied languages. Successful systems will be able to adapt to new languages and new genres.

The OpenCLIR evaluation was created out of the IARPA (Intelligence Advanced Research Projects Activity) MATERIAL (Machine Translation for English Retrieval of Information in Any Language) program that encompasses more tasks, including domain classification and summarization, and more languages. The purpose of OpenCLIR is to provide a simplified, smaller scale evaluation open to all. Please see the MATERIAL website for more information on the MATERIAL program.[1]

The first OpenCLIR evaluation will declare winners and award prizes; see section 8 Prize.

Please see the OpenCLIR evaluation website for up-to-date information and resources pertaining to the OpenCLIR evaluation.[2]

2 Evaluation Task

The OpenCLIR evaluation task is Cross Language Information Retrieval: Given a set of documents in a given foreign language and a set of English queries, retrieve the documents relevant to each query.

3 The Main Scoring Idea: A Detection System

Given a query (English word string), the system must detect which documents out of a set of documents are responsive to the query. The standard that system output will be scored against is a set of human annotations of relevance created according to the OpenCLIR19 Query Relevance Guidelines.

3.1 The Main Detection Metric: AQWV (Actual Query Weighted Value)

Each system will calculate a numerical score in the range [0,1] for every query-document pair. Participants will choose a value for a detection threshold θ that will optimize the system's performance in terms of the metric described below. Given an OpenCLIR query, all documents scored at or above the threshold value will be marked by the system as relevant to the query and all documents scored below will be marked as not relevant. The detection threshold is envisioned as being used as a dial by the end-user of a system, to be adjusted depending on user preference for higher precision versus higher recall.

For a given OpenCLIR query Q, let the number of documents that are relevant to Q be $N_{Relevant}$, and let the number of non-relevant documents to be $N_{NonRelevant}$. Let the total number of documents in the corpus be $N_{Total} = N_{Relevant} + N_{NonRelevant}$. For a given value of the detection threshold θ, let the number of relevant documents that a participant system did not mark as relevant be N_{Miss}, and let the number of non-relevant documents that the system marked as relevant be N_{FA}. Then, we define the Query Value QV for query Q at detection threshold θ as

$$QV(Q,\theta) = 1 - [\, P_{Miss}(Q,\theta) + \beta P_{FA}(Q,\theta) \,] \quad \text{(Equation 1)}$$

[1] https://www.iarpa.gov/index.php/research-programs/material
[2] https://www.nist.gov/itl/iad/mig/openclir-evaluation

where

- $P_{Miss}(Q, \theta) = \frac{N_{Miss}}{N_{Relevant}}$ is the probability of a missed detection error (i.e., the system failed to find a relevant document),
- $P_{FA}(Q, \theta) = \frac{N_{FA}}{N_{NonRelevant}} = \frac{N_{FA}}{N_{Total} - N_{Relevant}}$ is the probability of a false alarm error (i.e., the system retrieved a non-relevant document as relevant),
- β is defined as
    - $\beta \equiv \frac{C}{V}\left(\frac{1}{P_{Relevant}} - 1\right)$,
- C is the cost of an incorrect detection, here defined a-priori as 0.0333 (0.1/3) for the CLIR evaluation,
- Values of C may change as we converge on plausible applications,
- V is the value of a correct detection, here defined a-priori as 1.0, and
- $P_{Relevant}$ is an a-priori estimate, across datasets, of the prior probability that a document is relevant. Note that the value of $P_{Relevant}$ incorporated in β does not enter into the calculation of P_{Miss} or of P_{FA}.

$P_{Relevant}$ will be determined after relevance annotation of the evaluation datasets is complete.

β is defined as a constant a-priori so that all systems will optimize their performance in the same P_{Miss} vs. P_{FA} tradeoff space. Using the constants above (for C, V, and $P_{Relevant}$) gives $\beta = 20.0$ for the CLIR evaluation.

All queries will be weighted equally regardless of their respective $N_{Relevant}$. This value Query Weighted Value QWV at threshold θ is defined as

$$QWV(\theta) = \frac{\sum_{i=1}^{NQ} QV(Q_i, \theta)}{NQ} \quad \text{(Equation 2)}$$

where

- Q_i is a specific query
- NQ is the total number of queries
- QV is defined in Equation 1

$AQWV(\theta)$ is $QWV(\theta)$ when the system is running at its actual decision threshold.

The reader will note the following:

- $AQWV(\theta)$ = 1.0 for a perfect system
- $AQWV(\theta)$ = 0.0 for a system that puts out nothing (all misses, no false alarms)
- $AQWV(\theta)$ is negative for a system that produces excessive false alarms
    - $AQWV(\theta) = -\beta$ if none of the documents that are relevant (according to the answer key) are returned (so that $P_{Miss} = 1.0$), while all the documents that are actually non-relevant (according to the answer key) are returned (so that $P_{FA} = 1.0$)

Some queries may not have any relevant documents. Since AQWV is biased when a query has no relevant document, the following two AQWV alternatives will also be calculated. The second variant, shown in Equation 3 below and referred to as modified AQWV, is the **primary metric**.

- AQWV for queries with only relevant documents: Prior to scoring, queries without any relevant documents will be removed, and AQWV will be calculated the same way as with Equation 1 and Equation 2 above.

- Modified AQWV: Using P_{Miss} on queries with relevant documents and P_{FA} on all queries with the formula

$$QWV\ (\theta) = 1 - (\frac{\sum_{i=1}^{NQ_{Relevant}} P_{Miss}(Q_i,\theta)}{NQ_{Relevant}} + \beta \frac{\sum_{j=1}^{NQ} P_{FA}(Q_j,\theta)}{NQ}) \qquad \text{(Equation 3)}$$

where $NQ_{Relevant}$ is the number of queries with relevant documents.

4 Data Resources

At various times during the evaluation period, data packs will be released for system development and testing. The data packs are described below, while their distribution timeline is specified in the schedule.

4.1 Build Packs

Participants will receive build packs for Automatic Speech Recognition (ASR) and Machine Translation (MT) training. There will be approximately 50 hours of audio for ASR (with 40/10 training/development recommended division) and 800,000 words of bitext for MT training. Participants may wish to use some of the build-pack transcribed audio and bitext for DevTest purposes (e.g., doing deleted interpolation or n-fold cross-validation).

These build packs will consist of the following:

- Language-specific peculiarities and/or language specific design document(s) which contains information on the language:
  - What family of languages it belongs to
  - Dialectal variation
  - Orthographic information (including notes on any encodings that occur in our datasets)
    - Information on the character set
    - For a language written in a non-Latin character set, a transliteration into Latin characters
- Files of transcribed conversational audio in that practice language
  - The directory structure of the build pack will identify some of this as a DevTest [3]set, but participants are free to re-partition this data in any way desired
- Conversational audio: some in 8-bit a-law .sph (Sphere)[4] files and some in .wav files with 24-bit samples
- 800,000 words of bitext (sentences in the language and corresponding English translations)
  - Likely to include source URLs but probably little or no other metadata

4.2 Additional Training Data

In addition to the build packs, participants are allowed to use publicly available data for system training purposes. Such resources must be listed in the system description.

[3] While these files have a similar purpose to those described in Section 4.2.1, these are two distinct sets of files.

[4] Some tools to manipulate NIST Sphere format are available at https://www.nist.gov/itl/iad/mig/tools. Basic information about the Sphere format can be found at https://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section_02/text/nist_sphere.text

Note, the DevTest, Analysis, and Evaluation document packs described in section 4.3 are not permissible for any system training purposes.

4.3 Document Packs

There are three types of document packs: *DevTest, Analysis,* and *Evaluation*.

The genres for the text and audio modes are listed in Table 1:

Mode	Genre	Abbreviation
Text	News Text	NT
	Topical Text	TT
	Blog Text	BT
Audio	News Broadcast	NB
	Topical Broadcast	TB
	Conversational Speech	CS

Table 1: Genres of OpenCLIR document genres and their abbreviations

Some metadata including the genre information will be provided in the document packs. Audio files will be in .wav file format, and text files will be in UTF-8 ASCII .txt file format. The volume of text (number of documents as well as number of words) is substantially larger than the volume of speech.

Conversational Speech data will originate as two-channel audio and will be provided to participants as two-channel audio with the two channels temporally aligned. When any of that data is transcribed, the two channels will be transcribed separately, and then those two transcripts will be combined/interleaved into a single transcript that reflects the temporal alignment. Conversational Speech transcripts provided to participants (for example, in the Analysis Pack) will all be of that combined/interleaved form.

Audio data may have background speakers or music. We do not intend to transcribe what is clearly background speech, and we do not expect to score such background speech for retrieval.

The subsections below give more detail about each document pack type.

4.3.1 DevTest

To assist participants with system development, we will provide some data similar to the Evaluation dataset, which participants can use as a development test. The DevTest dataset is intended for the participants to use only for internal testing purposes.

4.3.1.1 Dry Run on DevTest

Every participating team is required to make at least one submission on the DevTest by the date specified in section 9 Schedule, and work with NIST in the event of problems with the submission until a valid submission is achieved. This is to ensure that potential issues can be corrected before the evaluation week. Failure to do so may result in being removed from further participation in the evaluation.

4.3.2 Analysis

To assist participants with error analysis, we will provide an Analysis dataset. English translations and transcriptions of the audio documents and query relevance will be included in each pack. The Analysis dataset will be larger than the DevTest dataset, and its composition will be similar to the DevTest.

4.3.3 Evaluation

The Evaluation dataset will be released at the start of the evaluation week.

4.4 Query Packs

The queries will be distributed to participants in two packs. The first query pack (QUERY-DEV) will contain *open* queries, where participants can conduct any automatic or manual exploration or data harvesting activities on the open queries as long as they are documented and disclosed. The second query pack (QUERY-EVAL) will contain *closed* queries, where participants are only allowed to submit to NIST (National Institute of Standards and Technology) for scoring their results produced against the Analysis, DevTest, or Evaluation document packs. These results must be generated by their fully automatic systems with no human in the loop.

Results on the open queries will not be counted toward the final AQWV.

4.5 Data Usage Restrictions

This section describes the rules governing the use of documents, queries, and query relevance annotations. An overview for each type of dataset is outlined in Table 2:

	Dataset			
	Build	**DevTest**	**Analysis**	**Eval**
Manually examine documents **before** the end of the Challenge	Yes	No	Yes	No
Manually examine documents **after** the end of the Challenge	Yes	Yes	Yes	Yes
Manually examine QUERY-DEV and relevance annotations	-	Yes	Yes	No
Manually examine QUERY-EVAL and relevance annotations **before** the end of the Challenge	-	No	No	No
Manually examine QUERY-EVAL and relevance annotations **after** the end of the Challenge	-	Yes	Yes	Yes
Automatic processing of all queries	-	Yes	Yes	Yes
Mine vocabulary from released documents and queries for ASR and MT development	Yes	No	No	No
Train ASR and MT models using released documents and queries	Yes	No	No	No
Automatically extract and process vocabulary from documents and queries for IR	-	Yes	Yes	Yes
Parameter tuning	Yes	Yes	Yes	No
Index data for automated modeling	Yes	Yes	Yes	Yes
Use IR models built from DevTest or Analysis	-	Yes	Yes	No
Build and apply cross-lingual training models from languages not currently evaluated	Yes	Yes	Yes	Yes
Score locally	-	Yes	Yes	No

Table 2: Rules outlining allowable actions for query and document sets

Participants should use the DevTest dataset to test their systems and can also use the DevTest dataset as a held-out dataset to set the values of general system parameters. Participants should not use the DevTest dataset for system training.

Unlike the DevTest dataset, participants are free to examine the Analysis dataset in detail, although it too should not be used as training data. We envision that the Analysis dataset will help participants to do glass-box testing to understand why and how their systems generated particular outputs, including how their system made miss errors and false-alarm errors. Participants may use the Analysis dataset and the QUERY-DEV annotations for glass-box analysis and parameter tuning of their systems or system components that are trained using other data. Participants should be mindful, however, of possible overfitting that may result from maximizing their components' performance on such a small set.

Because transcriptions and translations for the Analysis dataset will be provided, participants may calculate ASR word error rate scores and MT BLEU[5] scores on the Analysis dataset.

The Evaluation dataset is to be treated as a blind test.

Participants may mine the web for additional publicly available training and/or development test data. Any such data harvested for training or development must be specified in the system description. Participants must not hire native speakers for data acquisition, system development, or analysis. For example, it is forbidden to use native speaker consultants to find or post-process any data.

Participants may not use third-party commercial software in any part of their pipeline (e.g., transcription, translation, retrieval). Participants may use web-based MT software for translating a few words or phrases from the Analysis dataset as a potential way to understand errors in their systems.

Participants may use the QUERY-DEV queries in any way they wish, but must document their usage in the system description.

Participants must treat the QUERY-EVAL queries as part of the blind Evaluation dataset (i.e. no examination, no probing, no human in the loop). All QUERY-EVAL queries remain closed unless specified otherwise.

While data crawling may continue during an evaluation, models applied to Evaluation data cannot be modified using any data collected by the crawling during the evaluation period. All machine learning or statistical analysis algorithms should complete training, model selection, and tuning prior to running on the Evaluation data. This rule does not preclude online learning/adaptation during Evaluation data processing at evaluation time so long as the adaptation information is not reused for subsequent runs of the evaluation collection. The single exception to this is that participants are not allowed to use text Evaluation data for adaptation of their ASR models to the speech Evaluation data. Participants must document the ways their online learning and adaptation approaches incorporate information extracted from the Evaluation corpus in the system description.

No data or annotations may be distributed outside of the OpenCLIR Evaluation.

4.6 Dataset Structure

The following is a directory tree for a given dataset. Transcriptions, translations, query relevance annotations will only be provided for the Analysis datasets.

```
OPENCLIR_<EvalPeriod>-<LangID>/
      <DatasetName>/
            README.TXT
            file.tbl
            index.txt
            audio/
                  src/
                        <DocID>.wav
                  transcription/
                        <DocID>.transcription.txt
                  translation/
                        <DocID>.translation.eng.txt
```

[5] BiLingual Evaluation Understudy. See the original paper, “BLEU: a method for automatic evaluation of machine translation” at http://aclweb.org/anthology/P/P02/P02-1040.pdf.

```
            text/
                  src/
                        <DocID>.txt
                  translation/
                        <DocID>.translation.eng.txt

<EvalPeriod> ::= { 2019 | ... }
<LangID> ::= ID of the language
<DatasetName> ::= { ANALYSIS | DEV | EVAL }
<DocID> ::= base file name
```

5 File Formats and Their Interpretation

NIST has implemented a scoring tool[6] to calculate AQWV of a CLIR submission. The scoring tool requires the system output and reference to follow certain formats. This section describes these formats.

File formats will be UTF-8 ASCII text, with fields on the same line separated by a tab character. Lines are to be terminated by only a line feed character (no carriage-return), as is typical for Unix-based systems. Syntactically, a field may be empty.

5.1 Query Format

A query will consist of a query string (a word string) with no extra periods, spaces, or tabs.

```
QueryString = [“, a-zA-Z0-9()+:<>[]_] (i.e., includes parentheses and square brackets)
Query ::= QueryString
```

Here are two examples:

```
wheat
ebola
```

Queries will request different types of information, for example:

- lexical – requests the system to find documents that contain translation equivalent of the query string. Translation equivalent is not restricted to a word-to-word equivalent but should sound natural to a native speaker.
- conceptual – requests the system to find documents that contain topic or concept of interest suggested by the query string.
- hybrid – consists of part lexical and part conceptual and requests the system to find documents that satisfy the lexical part and/or conceptual part.
- morphological – requests the system to find a particular English morphological form, e.g. past tense verbs, plural nouns.
- EXAMPLE_OF: request the system to find entities that are members of a particular category, e.g. EXAMPLE_OF (mammal).

Refer to the MATERIAL Query Overview[7] and the OpenCLIR Evaluation Query Language Specification[8] for a complete description of the query syntax including what is allowed and not allowed.

[6] NIST will make public the scoring tool for participants to use.

[7] https://www.nist.gov/sites/default/files/documents/2018/07/12/openclirqueriesandrelevance.pdf

[8] https://www.nist.gov/sites/default/files/documents/2018/07/12/openclirqueryspecification.pdf

5.2 System Output Format

There will be one file per query in the CLIR task, containing the system decision whether each document is relevant to the query or not. Those files will be named:

```
<QueryID>.tsv
```

A legal name for such a file for a query would be query00043.tsv. The file will have one line for each document. These lines will be formatted as follows:

```
<DocID><tb><[Y|N]><tb><ConfidenceFactor>
```

The DocID of a document is the name of the corresponding file (as delivered to the teams by IARPA/NIST) without its extension (e.g. a file DOCUMENT_12345678.wav will have DOCUMENT_12345678 as its DocID).

Y|N will indicate for each document whether the system decided it is relevant to the query (Y) or not (N).

Confidence factors are specified in more detail in section 5.4 Confidence Factors.

A legal example of the first few lines of the query000043.tsv would be:

```
DOCUMENT_12345678  Y      0.85
DOCUMENT_52763409  Y      0.840
DOCUMENT_32198765  Y      0.840
DOCUMENT_98765432  N      0.5
```

5.3 Reference Format

The reference files for the CLIR task will be named as:

```
<QueryID>.tsv
```

For example:

```
query00043.tsv
```

The format of the CLIR reference is similar to that of the CLIR system output format, except without a confidence factor field, with each file containing one DocID and a Y|N field per line:

```
<DocID><tb><[Y|N]>
```

The first few lines of an example CLIR reference file for query query000043:

```
DOCUMENT_12345678  Y
DOCUMENT_28324932  N
DOCUMENT_52763409  Y
DOCUMENT_98765432  N
```

5.4 Confidence Factors

OpenCLIR systems will return a list of documents that are responsive to a query (a separate file for each query), and for each returned document the system will return a confidence factor in the range 0.0 through 1.0, where 0.0 means “definitely non-relevant” and 1.0 means “definitely relevant.” A system that has not [yet] implemented confidence scores should return a constant 0.5 as its confidence factor for each returned document.

The confidence factor is to always have exactly one digit to the left of the decimal point, with at least one digit to the right of the decimal point, and no more than five digits to the right of the decimal point. The number of digits to the right of the decimal point need not be constant.

The confidence factor is not to be in any other floating-point formats such as 5.0e-2. Examples of allowed confidence factors are:

```
0.0
0.5
0.54
0.54321
1.0
```

Examples of illegal confidence factors are:

`1`	(must have a decimal point and at least one digit to the right of the decimal point)
`0.543211`	(must have no more than five digits to the right of the decimal point)

Confidence factors of exactly 0.0 or exactly 1.0 have the same meaning across all systems. But this comparability *across systems* does not hold in between those values. More formally, for all confidence factors *cf* such that 0.0 < *cf* < 1.0 there is *no* assumption that the confidence factors returned by one system are comparable to the confidence factors returned by another system. On the other hand, confidence factors returned by the *same system* on different queries or on different datasets are assumed to be comparable.

6 Evaluation Scoring Server

NIST will provide an automated scoring server for the OpenCLIR evaluation. To make submissions, each team PI must sign up for an evaluation account via https://openclir.nist.gov by filling in all applicable fields. The PI must also complete a data license agreement. There will be one account per team. The team PI who signed up is to share login credentials with anyone on their team who needs to be able to access the web server. Both manual and programmatic submissions will be supported.

6.1 Submission Limits and Feedback

Participants can submit their system output on the different datasets for scoring to help their system development. Submission limits are listed in Table 3, along with the type of feedback provided.

Each submission will be validated prior to scoring. Only submissions that pass validation will count toward the submission limit. Submissions must follow the format given in section 6.2 Evaluation Submission Format.

Timeline	**Data & Query Sets**	**Limit per week**[9]	**Feedback (score)**
Development Cycle	ANALYSIS & QUERY-DEV	unlimited	yes
	DEV & QUERY-DEV	unlimited	yes
Evaluation Period	EVAL & QUERY-EVAL	1	yes, only overall score

[9] This limit may be increased if experience shows that more would present no implementation problems.

Table 3: Submission quota by dataset and cycle

6.2 Evaluation Submission Format

Each submission will be an archive file named as follows:

```
<SysLabel>.tgz
```

<SysLabel> is an alphanumeric [a-zA-Z0-9] label. This label will in part be created from hard-coded information and participant account information, and in part from the following information that participants will specify prior to uploading the submission file to the scoring server:

```
<DatasetName> ::= { ANALYSIS | DEV | EVAL }
<QuerysetName> ::= {QUERY-DEV, QUERY-EVAL}
```

There should be no parent directory when the submission file is untarred. The tar command should be:

```
> tar MySystemSubmissionFile.tgz query*.csv
```

The server will validate the submission file content to make sure the system output files conform to the format described in section 5.2.

6.3 Reporting Scores

This section describes the analyses and scores that will be reported for the various kinds of evaluations.

In addition to overall results, results on various factors (e.g., genre, mode, query length, possibly others) will also be reported. We expect such factors will include various characteristics of queries such as the number of words in the query string, linguistic characteristics such as polysemy of the word(s) in the query string, homophony, named entities, etc., in order to provide maximal insight. During the development cycle, participants will also get these breakdowns for the Evaluation datasets. However, once the development cycle ends and evaluation cycle starts, participants will only receive top level AQWV results on the Evaluation datasets.

Because the full evaluation data is released incrementally, some queries may have no relevant documents for the released subsets. In such cases, if a system also retrieves nothing for a query with no relevant document, the $P_{Miss} = 0$ and $P_{FA} = 0$.

7 System Description

To facilitate maximal information exchange and understanding of the systems developed for the OpenCLIR evaluation, teams will be required to submit a system description of at least seven pages, describing the designs and methods as well as any data harvested and how it was used. The system description will count 20% toward the determination of winners. A system description template detailing the format and what is expected to be in the system description is available on the OpenCLIR website.

8 Prize

The first OpenCLIR evaluation will declare a winner in two separate categories, text and audio data. Current MATERIAL performers are excluded from consideration for prizes.

The winning submissions will be determined using a combination of AQWV score (required to be at or above a baseline of 0.2 for the text category and 0.1 for the speech category) and rating of system description. The AQWV score will be weighted at 80% towards the prize determination; the system description rating will be weighted at 20%. Submissions by participants who fail to submit the required system description will not be eligible to win.

The winners will receive a monetary award. The award will be USD 10,000 for the text category and USD 20,000 for the audio category. Prizes will be awarded in accordance with the laws of the USA and of the winning participants' countries.

Detailed rules regarding the awards are specified in the Prize Challenge Rules & Regulations document.

9 Schedule

Milestone	**Date**
Release of evaluation plan	July 2018
Registration period	July 2018 – November 30, 2018
Release of Build Packs (Training Data)	August 21, 2018
Development Cycle ● Release of ANALYSIS, DEV, & QUERY-DEV (encrypted data & decryption keys) Scoring server accepts submissions for QUERY-DEV on ANALYSIS and on Dev ● Development cycle dry run submission on DEV due	August 21, 2018 – May 31, 2019 August 21, 2018 September 2018 February 14, 2019
Release of EVAL & QUERY-EVAL (encrypted data)	March 4, 2019
Evaluation Period ● Release of EVAL and QUERY-EVAL (decryption keys) Scoring Server accepts submissions for QUERY-EVAL on EVAL ● System output due to NIST	March 11 – May 31, 2019 March 11, 2019 May 31, 2019
System description due to NIST	July 12, 2019