Evaluation Plan for the IARPA MATERIAL Program

(**MA**chine **T**ranslation for **E**nglish **R**etrieval of **I**nformation in **A**ny **L**anguage)

Revision History Highlighted version number indicates released to the public.			
Version Number	**Date**	**By**	**Description**
6.0.4	20181203	Ilya Zavorin	• 1S Domain Z release information updated: in-scope, out-of-scope and addtl feedback parts for Religion updated in the Appendix
6.0.3	20181130	Audrey Tong Richard Tong Ilya Zavorin	• 1S Domain Z release information: updated Table 4 in Section 6.2, added domain definition to the end of the document • Added 3rd paragraph to Section 8.1 to clarify submission for E2E and CLIR
6.0.2	20181121	Audrey Tong Richard Tong	• Modified Section 7.2.2 to indicate that for DomainID, CS documents should not be included in the system output because they are not scored. • Modified Section 7.2.4 to update new json format and clarify system output for E2E. • Modified Section 8.1 regarding no DEV/ANALYSIS submission during eval week. • Modified Section 8.2 on submission file format and where to make submissions.
6.0.1	20181030	Audrey Tong	• Revised E2E submission directory location. • Revised Table 7 to clarify that looking at relevance annotations is not same as looking at underlying documents and that looking at underlying documents is allowed after the language is released by T&E.
6.0.0	20181018	Carl Rubino Audrey Tong	• Revised Section 2.1 to include Beta for 1S • Revised Section 6.5 to make the data restrictions clearer. • Revised Section 7 to update the system output and reference file format. • Merged E2E system output file format (from SummaryFormatSpec_3.1.pdf) to Section 7.2.4 with 2 main differences: an extra column for hard decision and a single type of submission, only TAR of directories (see Section 8.2.2)
5.9.5	20181003	Audrey Tong	• Updated 1S schedule. • Updated Table 4 to include latest domains released for each language.
5.9.4	20180906	Ilya Zavorin	• Added to Section 6.5 a restriction on the use of text Eval data for speech adaptation.
5.9.3	20180810	Audrey Tong	• Fixed typo in Table 8, submission to eval set is always once per week, not 200 :-)
5.9.2	20180720	Audrey Tong	• Updated Table 8 to explicitly state when DomainID and LangID are available on the scoring server. • Revised Table 8 regarding when Analysis and Eval will be available for 1S. During eval week for 1S and beyond, there will only be one submission for the eval query and document sets.

5.9.1	20180711	Audrey Tong	● Updated Table 8 in Section 8.1 to include meeting dates, domain release dates. This table is now the definitive Base Period schedule. ● Removed section 9 and updated section 6 to point to Table 8.
5.9.0	20180702	Audrey Tong	● Added Language Identification task in the Introduction and Section 5, and corresponding description for system output (Section 7.2.3) and reference format (Section 7.3.3) ● Added pointer to E2E spec in Section 7.2.4 ● Updated evaluation rule Section 6.5 to clarify data crawling during program period and usage of eval data ● Updated Table 8 submission rules table in Section 8.1 to match submission rules given in spreadsheet
5.8.4	20180518	Ilya Zavorin Audrey Tong	● Added section 6.2 that lists all 1A/1B domains X, Y and Z. ● Added Appendix Description of Domains (X, Y and Z) ● Replaced references to Domains 1 & 2 etc to Domains X etc. ● Corrected broken cross-references ● Removed Epoch 1 restriction from Section 5 ● Updated Table 8 to indicate submissions for DEV during eval week is allowed.
5.8.3	20180419	Audrey Tong	● Made clear that BBN's proposed AQWV is the primary metric. ● Fixed error in Table 8 (during eval cycle only QUERY2 can be submitted).
5.8.2	20180416	Audrey Tong	● Modified section 2.1 to include BBN's proposed AQWV. ● Modified Table 8 to clarify the submission limit, what can be submitted and when.
5.8.1	20180301	Audrey Tong	● Modified section 5 to indicate that the server now provides X_4. ● Fixed error in section 7.3.2 where the first line of the DomainID AllDocIDs file was wrong.
5.8	20180223	Audrey Tong Richard Tong Ilya Zavorin	● Revised section 2.1 to use consistent nomenclature. ● Revised sections 3 & 4 to better explain E2E metric. ● Revised sections 7.3.1 & 7.3.2 to include information about the reference filenames. ● Revised section 8 to include specific registration instructions. ● Modified section 8.1 to better explain the submission limit. ● Modified section 8.2 to include instructions on how to tar up the submissions. ● Modified sections 8.3.1 & 8.3.2 on the types of results reported by the scoring server. ● Removed the schedule in section 9 and referred readers to a separate document so we don't have to maintain both. ● Minor format and cosmetic fixes.
5.7	20171215	Audrey Tong	● Reorganized section 7 which included removing detailed MATERIAL query syntax and CFG from the eval plan and referring readers to a separate document, dividing remaining content into 4 subsections - query format, system output format, reference format (new), and confidence factor. ● Added a subsection in 8 to discuss submission limit. ● Updated base period schedule regarding site visits, mid-period meetings. ● Corrected various inconsistencies in the eval plan (percentages of X_1, X_2, X_3, X_4 will be reported rather counts; DomainID is the official name).
5.6	20171130	Audrey Tong	Minor edits requested by T&E
5.5	20171128	Audrey Tong	Reorganized Data Resources section

5.4	20171120	Audrey Tong	Added task description Added information about categories of queries (open/closed) and their restrictions Increased precision of confidence factor to 5 decimal points
5.3	20171114	Ilya Zavorin	Added figure and table captions and cross-references; modified query counts in Table 3
5.2	20171106	Audrey Tong	Updated data resource & data release structure Added evaluation scoring server section Minor reorganization and cosmetic fixes
5.1	10 / 26 / 2017	Greg Sanders	Update reflecting changes during first few days after kickoff mtg.
5.0	9 / 15 / 2017	Greg Sanders	Updates to create version released at MATERIAL kickoff meeting
4.3 – 4.6	9 / 6–15 / 2017	Greg Sanders	Minor edits requested by WERB reviewers
4.2	8 / 31 / 2017	Greg Sanders	First version with complete content

CONTENTS

1 INTRODUCTION

The goal of MATERIAL is to develop methods to locate text and speech content in "documents" (speech or text) in low-resource languages using domain-contextualized English queries and to display a summary in English of the information of interest in the relevant documents. This capability is expected to enable effective triage[1] and analysis of large volumes of data, and to do so in a way that takes into account an analyst's domains of interest in a variety of less studied languages. The program will require that the capability be constructed using limited amounts of ground truth bitext data (~800K words) and no domain adaptation data. Successful systems will be able to adapt to new domains and new genres.

The queries will be in English, the material to be searched will be in different languages, and the summaries must be displayed in English.[2] It should be noted that in real-world use, the output from the system would represent documents from multiple languages, mingled in one output "queue."

A summary could be a word-cloud, an extractive summary, or an abstractive summary. The summary will be required to be formatted as static text: possibly with multiple colors, sizes, and spatial alignments and orientations, but with no animations, and no lines/arrows or other graphic elements. The goal is that the summary must suffice for the user to judge relevance of the summarized item to the domain-contextualized query. Research in MATERIAL will include work on effective summarization.

A central goal of the MATERIAL system is to identify both information needs and topics of interest to potential users. For this reason, multiple domains, or high level topics, will be investigated per language, determined by the data collected. For the base period of the program (which is what this evaluation plan covers) there will be five domains per language.

It is possible that MATERIAL systems will, eventually, make use of ontologies (ultimately to be provided or customized by the user or by the user's organization) to achieve several goals:

First, an ontology can provide domain context for the user's statement of information need (which we will refer to as a query in domain) to add constraints on the relevancy of a document to the query.

Second, by learning how to associate the words in an English ontology with the words in the low resource language, the system can expand or leverage the resources available in a low-resource language and in English to address (partially substitute for) the lack of domain-specific bitext needed to adapt the MT system for a new domain.

Third, the ontology can be leveraged to identify related words/phrases for use by the Information Retrieval portions of a system.

No involvement of IARPA or NIST in ontologies is anticipated during the base year of the program, and ontologies are not mentioned further in this Evaluation Plan.

The evaluation tasks for each program period and language stage are given below:

1) Cross Language Information Retrieval (CLIR) – given a set of foreign language documents and English queries, retrieve documents that are relevant to each query.

---

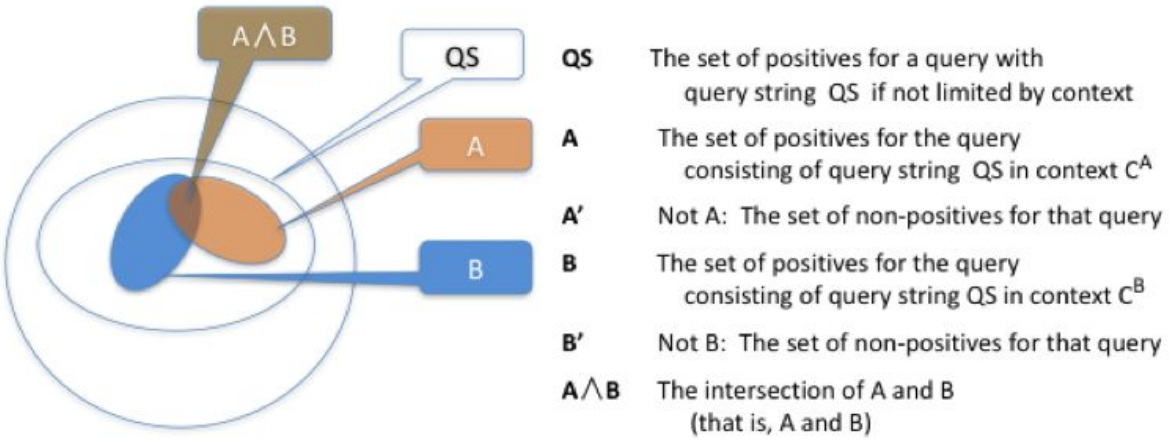
[1] "Triage" may refer to the entire end-to-end system, rather than just finding the documents. Similarly, triage may encompass a system filtering a continuous incoming document stream or a system clustering-by-context a set of documents that are responsive to a query string in a domain.

[2] Developers are free to use any techniques they wish, but in developing this evaluation plan we have considered that methods from cross-language Information Retrieval (CLIR), machine translation (MT), and summarization could provide a possible base for the development of successful novel approaches.

2) Cross Language Information Retrieval and Summary (CLIR+S or E2E) – given a set of foreign language documents and English queries, retrieve documents that are relevant to each query and generate a summary[3] in English for each document the system deems relevant to a query.
3) Domain Identification (DomainID) – given a set of foreign language documents, identify which of those documents are relevant to a given domain.
4) Language Identification (LangID) – given a set of foreign language documents for a given MATERIAL language, identify which of those documents are in that language.

2 The Main Scoring Idea: MATERIAL as a Detection System

Given a query (a domain as context and a query-string of words), the MATERIAL system must detect which documents are responsive.



	B	B'
A	A∧B (A and B)	A∧B' (A only)
A'	A'∧B (B only)	A'∧B' (neither A nor B)

Figure 1: Material evaluation diagram

If the world in which a MATERIAL system operates is viewed as shown in Figure 1, we will *not* be evaluating the ability of a system to retrieve the set labeled QS (that is, the positives if not limited to a context). In MATERIAL, a query is always a query string *plus* a context/domain. If we were to draw a

[3] The summary is to tell English speakers how the document is relevant to the query, not a summary of the document.

similar diagram with an area A for documents relevant to the query string and an area B for documents relevant to the domain, that diagram would also be like the diagram above, in which case $A \wedge B$ would be documents that are relevant to the query string in the domain (i.e., to both the string and the domain).

The syntax for a query in MATERIAL does not support formulating a query whose context is "A only" or a query whose context is "B only" (that is, "A but not B," excluding the intersection of A with B). The user of the system will not be able to formulate such a query, and thus there will not be any such queries in MATERIAL.

However, if the query is in context/domain A, then documents that are in "B only" will be non-targets, and evaluation needs to identify the system's ability to not return such non-targets.

2.1 The Main Detection Metric: AQWV

Each performer system will calculate a numerical score in the range [0,1] for every query-document pair. As described in Section 1.B.2.1 of the MATERIAL BAA, performers will choose a value for a detection threshold θ that will optimize system's performance in terms of the program metric described below. Given a MATERIAL query, all documents scored at or above the threshold value will be marked by the performer system as relevant to the query and all documents scored below will be marked as not relevant[4].

For a given MATERIAL query Q, let the number of MATERIAL 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 osf relevant documents that a performer 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 and detection threshold theta θ as

$$QV(Q,\theta) = 1 - [\, P_{Miss}(Q,\theta) + \beta\, P_{FA}(Q,\theta)\,] \qquad \text{Equation1}$$

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),
- $\beta \equiv \frac{C}{V}\,(\frac{1}{P_{Relevant}} - 1)$
    - C is the cost of an incorrect detection, defined in Table 1. Values of C may change each program period, as we converge on plausible applications of MATERIAL systems.
    - V is the value of a correct detection, defined in Table 1.
    - $P_{Relevant}$ is an a-priori estimate, across datasets, of the prior probability that a document is relevant, defined in Table 1. Note that the value of $P_{Relevant}$ incorporated in β does not enter into the calculation of P_{Miss} or of P_{FA}.

      Our initial analyses suggest 1/600 (approximately 0.001667) is a reasonable estimate of $P_{Relevant}$ for the actual data (value subject to change as we experiment further with our datasets and our baseline system).

[4] The detection threshold is envisioned as being used as a dial by the end-user of a MATERIAL system, to be adjusted depending on user's preference for higher precision versus higher recall.

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:

Language	CLIR				CLIR+S			
	V	C	$P_{relevant}$	β	C	C	$P_{relevant}$	β
1A	1	0.0333	1/600	20	1	0.1	1/600	59.9
1B	1	0.0333	1/600	20	1	0.1	1/600	59.9
1S	1	0.0668	1/600	40	1	0.0668	1/600	40

Table 1: V, C, $P_{relevant}$ and β for each language and task.

All queries will be weighted equally regardless of their respective $N_{Relevant}$[5]. This value Query Weighted Value is defined as

$$QWV(\theta) = average_{\top Q}\ [\![QV(Q,\theta)]\!] \qquad \text{Equation2}$$

$AQWV(\theta)$ 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)
- $AQWV(\theta)$ can go negative if greatly excessive false alarms
    - $AQWV(\theta) = -\beta$ if none of the documents that are actually 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 $P_{FA} = 1.0$)

Since MATERIAL evaluation data will be released incrementally, some queries may not have any relevant documents. Because AQWV is biased when a query has no relevant document, two AQWV alternatives will also be calculated. The second variant 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 using Equation1 and Equation2 given above.
- AQWV using P_{Miss} on queries with relevant documents and P_{FA} on all queries with the formula:

$$QWV\ (\theta) = 1 - [\ average_{Rel-q\ \top Q}(P_{Miss}\ (Q,\theta)) + \beta\ average_{All-q\top Q}(P_{FA}\ (Q,\theta))\ ] \qquad \text{Equation3}$$

$AQWV(\theta)$ will be calculated separately for the Cross-language Information Retrieval (CLIR) aspects of the system (as described above) and for the full E2E system (see Section 3, below).

2.2 Area Under the Curve (AUC) as an Early-Stage Detection Metric

Because AQWV can go negative, it may not be a useful metric early in the development of a system. For that early stage of system development, performers may wish to use "Area Under a ROC curve" (AUC)

[5] One can similarly define Document Value and Actual Document Weighted Value metrics by considering individual documents rather than queries, but we do not plan to calculate it.

as a useful metric to optimize, and AUC can be a useful metric for communicating early-stage progress. NIST will not be using AUC in its evaluations. We assume that AUC would only need to be calculated on the Analysis Dataset and only in the earliest stages of system development (a system does not have to be very good before AQWV will always be positive).

3 Summaries and the Summarization Metric

When a system identifies a document as relevant to a query, it must then generate a textual summary in English of the document's content that is relevant to the query. As explained in the third paragraph of the introduction to this Evaluation Plan, research on summarization is in-scope for the MATERIAL program. The summary may make use of multiple colors of text, multiple alignments/rotations of text, multiple sizes of text, and spatial positioning of text. There are three main constraints on the summary:

1) it must be textual (no graphics such as lines, arrows, or bubbles),
2) it is limited to 100 words, and
3) it must be static (no animations).

Each summary is to be delivered as a separate file as an image (such as .jpg) with a suggested the image size be limited to 768 pixels tall by 1024 pixels wide, so that it can be viewed on a typical-size computer screen without being resized. The point of allowing an image format is to allow performers to control the visual presentation of their summaries[6].

Each summary will be evaluated by K judges[7] on how well summaries convey evidence of relevancy of the underlying document to the query. For each judgment by a judge, the judge will see the query (a query string plus a domain) and the summary for a document that a system deemed to be relevant to the query. Each judge, on the basis of only the query and the summary (without seeing the document itself), will (independently of all other judges) decide whether the document is (or is not) relevant to the query.

We will also have an answer key that gives the ground truth about the relevance of a document to a query, which we will use in scoring. The answer key will be generated by bilingual annotators who understand both the English query and the foreign-language document.

4 End-to-End Metric

In this section we explain the formulation of AQWV for E2E. We start by thinking about scoring a single document D_i processed by the Performer Team's System (PTS) in response to a specific query Q. For this Q-D_i pair, we have a CLIR contingency matrix that looks like:

		Performer System (CLIR/E2E)	
		R (Relevant)	N (Not Relevant)
Answer Key	R (Relevant)	X_1 (true positive)	X_2 (false negative/

[6] Detailed information regarding the summary is still under discussion.

[7] The actual number of judges is still under discussion.

			miss)
	N (Not Relevant)	X_3 (false positive/ false alarm)	X_4 (true negative)

Table 2: Contingency matrix

Only one of (X_1 X_2 X_3 X_4) can be 1 so we have four possible outcomes after CLIR for a given Q-D_i pair.

		PTS	
	$X_1=1$	R	N
Answer Key	R	1	0
	N	0	0

		PTS	
	$X_2=1$	R	N
Answer Key	R	0	1
	N	0	0

		PTS	
	$X_3=1$	R	N
Answer Key	R	0	0
	N	1	0

		PTS	
	$X_4=1$	R	N
Answer Key	R	0	0
	N	0	1

Table 3: Possible value for the contingency matrix for one document D_i, processed by a performer's system, in response to one query Q.

The PTS generates a summary if it deems the document is relevant (where $X_1=1$ and $X_3=1$). We will use human judges to assess the quality of the summary. If the assessment is made by K judges then we have two possible ways of using the judgments:

- Convert them into a single binary judgment. That is, take the set of K responses and under some decision rule annotate the corresponding document (D_i) as either Relevant (R) or Not Relevant (N).
- Use the individual responses directly. That is, annotate D_i as having some number of relevant judgments and some number of not relevant judgments.

In the case of binary judgment, the original CLIR contingency matrix either remains the same or changes to the $X_2=1$ case where $X_1=1$; and the original CLIR contingency matrix either remains the same or changes to the $X_4=1$ case where $X_3=1$.

In the case of raw judgment, the original CLIR contingency matrix becomes a mix of $X_1=1$ and $X_2=1$ cases where $X_1=1$; and the original CLIR contingency matrix becomes a mix of $X_3=1$ and $X_4=1$ cases.

If $X_2=1$ or $X_4=1$, the human assessment step does not apply; and the corresponding CLIR contingency matrices are used "as is" in the E2E AQWV computation.

If $X_1=1$ or $X_3=1$, the human assessment step can potentially modify the corresponding CLIR contingency matrices that are used in the E2E AQWV computation.

Looking at the set of documents D wrt to a specific Q, we can compute the aggregated contingency matrix by summing the $(X_1\ X_2\ X_3\ X_4)$ for each individual document D_i.

$$\widehat{X}_1 = \sum_{i=1}^{D} X_{1i} \qquad \widehat{X}_2 = \sum_{i=1}^{D} X_{2i} \qquad \widehat{X}_3 = \sum_{i=1}^{D} X_{3i} \qquad \widehat{X}_4 = \sum_{i=1}^{D} X_{4i}$$

So AQWV for E2E is $AQWV_{E2E} = 1 - \left(\frac{\widehat{X}_2}{\widehat{X}_1+\widehat{X}_2} + \beta\frac{\widehat{X}_3}{\widehat{X}_3+\widehat{X}_4}\right)$.

5 Domain and Language Identification Evaluation

To assist performers with developing their systems, we will evaluate two additional dimensions: Domain Identification and Language Identification. These will be an automated evaluation, where performers can submit their outputs repeatedly and automatically get back results.

The Domain Identification (DomainID) task consists of identifying the domains to which each document is relevant (a document can be relevant to more than one domain). We will score Domain Identification, separately for each domain, by counting the number of correct domain detections (X_1 in Table 2), the number of domains that the system fails to detect (misses, X_2 also in Table 2), the number of domains that the system incorrectly "detects" (false alarms, X_3), and the number of domains that the system correctly detects as not relevant (X_4). We will report those four numbers for each domain, as a percent of the ground truth number (i.e., of X_1 for a perfect system).

Similarly, the Language Identification (LangID) task consists of identifying which of the documents in the corpus corresponding to a given MATERIAL language are in fact in that language. We will score Language Identification similarly to Domain Identification by counting the number of documents the system correctly detects for a given language (X_1 in Table 2), the number of documents that the system fails to detect (misses, X_2 also in Table 2), the number of documents the system incorrectly "detects" (false alarms, X_3), and the number of documents that the system correctly detects as not belong to the given language (X_4). We will report those four numbers for each language, as a percent of the ground truth number (i.e., of X_1 for a perfect system).

6 Data Resources

At various time during the program period, data packs will be released for system development and testing. The data packs are described below while their distribution timeline is given in section 8.1.

6.1 Build Packs

Performers 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 800k words of bitext for MT training. Performers 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 [8] set, but performers are free to re-partition this data in any way desired
- Conversational audio: some in 8-bit a-law .sph (Sphere)[9] files and some in .wav files with 24-bit samples
- The 800k words of bitext (sentences in the language and corresponding English translations)
  - We anticipate providing source URLs but probably little or no other metadata

6.2 Domains

Domains in MATERIAL are broad-level subject categories. Generally a document is deemed relevant to a domain if a portion of it clearly shows sufficient evidence of that domain, that is, it is not simply a passing mention. Table 4 lists the domains that have been released to date.

Language	Release Stage[10]		
	X	Y	Z
SWA (1A)	GOV LIF	BUS LAW	SPO
TGL (1B)	GOV LIF	HEA MIL	SPO
SOM (1S)	GOV MIL	LAW BUS	REL

Table 4: MATERIAL Domains Released to Date

Detailed descriptions of the above domains are given in the the Appendix.

Domain names (e.g. `Government-And-Politics`) are to be used in CLIR and CLIR+S submissions as is currently described in Section 7.2 as well as, if desired, in communications between the teams and IARPA/T&E, while domain codes (e.g. `GOV`) are intended primarily as shorthand notation for communications (reports, presentations etc).

[8] Although somewhat similar in purpose, this DevTest set (designed specifically to test and tune ASR models) is different from the one described in Section 6.3.1 (designed to test and tune E2E systems).

[9] 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

[10] Per MATERIAL BAA (https://www.fbo.gov/utils/view?id=eae1f0b7afd4cbe6fa94e99117774121) Section 1.B.2.2

6.3 DOCUMENT PACKS

There are three types of document packs: *DevTest, Analysis,* and *Evaluation*. The document packs contain six genres of "documents" listed in Table 5.

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

Table 5: Genres of MATERIAL documents and their abbreviations

Some metadata including the genre information will be provided in the document packs.

Audio files will be in .wav file format (some will be .sph format in the build packs released around program kickoff), 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 expected to be substantially larger than the volume of speech. Perhaps ¾ of the documents will be text.

Because perhaps ¼ of the documents will be audio, performer teams will need ASR[11]. Likewise, performers' systems will have to adapt to new genres and new domains, which is a key challenge for the program.

Conversational Speech data will originate as two-channel audio and will be provided to performers 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 performers (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 or summarization.

Domains will be specific to the languages. Some domains will be used in multiple languages, others may not). Some documents in each dataset will be relevant to more than one domain of interest.

The subsections below give more detail about each document pack type. Figure 2 shows the relationship among the various document and query packs.

[11] Audio data in the build packs released at program kickoff and in the Analysis Dataset will come with transcriptions, but transcriptions will not be provided for the evaluation data. Performers' systems must ingest audio speech data automatically.

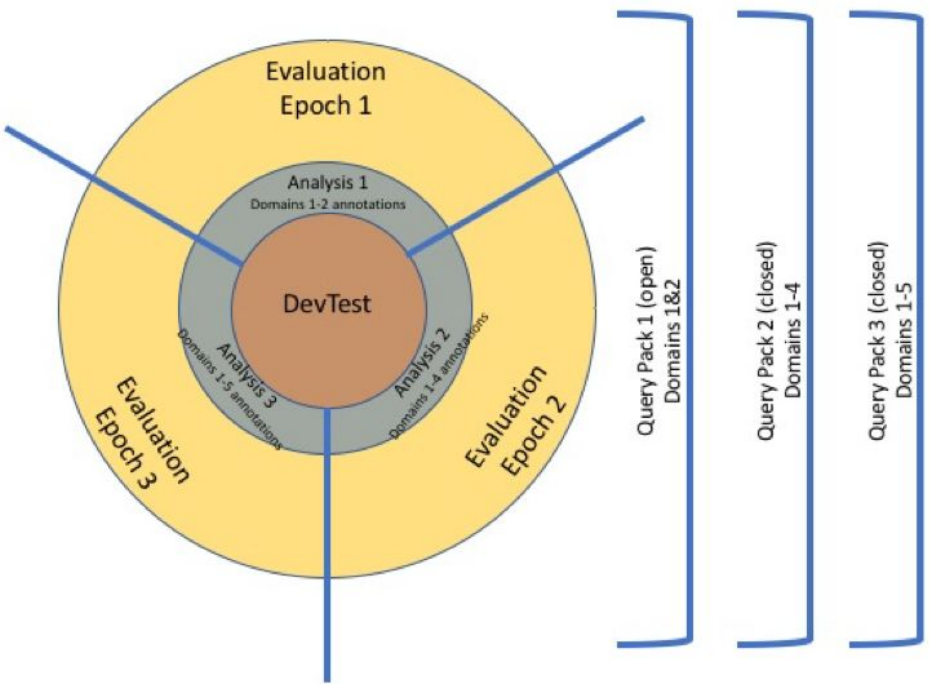


Figure 2: Relationship between document and query packs. Circles representing the various document packs are not drawn to scale, i.e., Analysis appears much bigger than DevTest to accommodate the text label but in reality they are closer in size.

6.3.1 DevTest

To assist performers with system development, we will provide to the performers some data that is similar to the Evaluation Dataset, which performers can use as a development test. The DevTest Dataset is intended for the performers to use only for internal testing purposes. The DevTest Dataset will consist of about 650 "documents" for each language and will be released in its entirety in a single DevTest pack.

6.3.2 Analysis

To assist performers with error analysis, we will provide an Analysis Dataset that will be released in three packs[12] reflecting the domains that have been announced (first pack reflecting the first two domains, etc.). English translations and transcriptions of the audio documents as well as domain annotations and query relevance[13] will be included in each pack. The Analysis Dataset will be roughly the same size of the DevTest Dataset and its composition will be similar to the DevTest (but will have more documents that are not relevant to the domains of interest).

[12] The three analysis packs together are also referred as the Analysis Dataset, and likewise the three evaluation packs are referred as the Evaluation Dataset.

[13] With the exception of the first analysis pack where it will only include domain annotation. Query relevance for the analysis pack 1 will be released when the first query pack is released.

6.3.3 EVALUATION

The Evaluation Dataset will also be released in three packs where each pack represents an "epoch". Each epoch contains ~5000 documents based on the date and time of origin[14] for each document, with the time interval of the epochs being specific to each genre in each language. Epoch 1 will be the earliest documents, and Epoch 3 will be the most recent. Some documents in the Evaluation Dataset may be in closely-related languages, to serve as "distractors" as well as documents that are not relevant any of the five domains.

While the DevTest and Analysis will have essentially equal numbers of documents in each of the five domains of interest, no such constraint will be imposed on the Evaluation Dataset.

6.4 QUERY PACKS

The program queries will be distributed to performers in three packs for each language under test. The first query pack will contain *open* queries where performers can conduct any automatic or manual exploration or data harvesting activities[15] on the open queries as long as they are documented and disclosed. The second and third query releases will contain *closed* queries where performers are only allowed to submit to NIST for scoring their results produced against the Analysis, DevTest, or Evaluation document packs. These results must be generated by their fully automatic E2E systems with no human in the loop.

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

Table 6 shows the minimum number of queries, per practice language, expected to be released at the three stages. We expect some additional queries will be collected.

	Query Pack 1	**Query Pack 2**	**Query Pack 3**	**Number of Queries**
Restrictions	**Open**	**Closed**	**Closed**	
Domains X	300+	200+	200+	700
Domains Y		200+	200+	400
Domain Z			200+	200
Total				**1300**

Table 6: Query release counts per language. +Queries for each pack and domain are the same across analysis, development, and evaluation sets.

6.5 DATA USAGE RESTRICTIONS

This section describes the rules for document and query use.

	Build	**DevTest**	**Analysis**	**Eval**

[14] Except for conversational (CS) documents that do not have timestamps associated with them and so were assigned to Evaluation packs at random.

[15] Some of these activities may not be allowed in a later program period.

Manually examine documents **before** the language is released	Yes	No	Yes	No
Manually examine documents **after** the language is released[16]	Yes	Yes	Yes	Yes
Manually examine Q1 and relevance annotations on <document set>	-	Yes	Yes	No
Manually examine Q2, Q3 and relevance annotations before E2E eval	-	No	No	No
Manually examine Q2, Q3 and relevance annotations after E2E eval[17]	-	Yes	Yes	Yes
Automatic processing of all queries (Q1, Q2, Q3)	-	Yes	Yes	Yes
Mine vocabulary from documents and queries for MT/ASR development	Yes	No	No	No
Train MT/ASR models on languages currently evaluated from <document set>	Yes	No	No	No
Automatically extract and process vocabulary from documents and queries for IR and Summarization	-	Yes	Yes	Yes
Parameter tuning	Yes	Yes	Yes	No
Index data for automated modeling and E2E component algorithms	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 7: Rules outlining what is allowable for query and document sets.

Performers should use the DevTest Dataset to test their systems (one does not want to test on one's training data) and can also use the DevTest Dataset as a held-out dataset to set the values of general system parameters.

Unlike the DevTest Dataset, performers 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 performers 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. Performers may use the Analysis 1 documents (i.e.

[16] As of Oct 30, 2018, only 1A has been released.

[17] Except for 1B which remains closed. Please note that examining relevance annotations does not include examining the underlying documents.

the first pack of Analysis documents) and the open query relevance annotations (i.e. for the queries from first Query release pack) for "glass-box" analysis and parameter tuning of E2E systems, or their components, that are trained using other data. Performers 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, performers may calculate ASR Word-Error-Rate scores and MT BLEU[18] scores on the Analysis Dataset.

Evaluation Dataset is to be treated as a blind test.

No domains will be analyzed for the Conversational Speech genre. The speakers in the Conversational Speech documents were not asked to converse about particular domain(s). For that reason, analyzing the domains in such data is not entirely valid. Also, a typical conversation can include more wide-ranging topics than is the case for the other genres.

Performer teams may mine the web for additional training and/or development test data. This paragraph is intended to clarify the restrictions mentioned at the top of page 11 of the BAA. Specifically, any such data harvested for training or development must be shared with the other teams after the end of the evaluation cycle in which it is first used (for example, after the CLIR evaluation, after the CLIR+S end-to-end evaluation, etc.). In contrast, if teams purchase data, it must be shared with the other teams immediately (see the first full paragraph on page 11 of the BAA). In either case, as stated in the first full paragraph on page 11 of the BAA, teams must not hire native speaker consultants for data acquisition, system development, or analysis. For example, it is forbidden to use native speaker consultants to find or post-process any such data.

Performer teams may not use third-party commercial software in any part of their pipeline (e.g., transcription, translation, retrieval, summarization, language ID, data harvesting). Teams may use web-based MT software for translating a few words or phrases from the Analysis documents as a potential way to understand errors in their systems.

Performer teams may use the open queries in any way they wish but must document their usage. Performer teams must treat the closed queries as part of the blind evaluation set (no examination, no probing, no human in the loop). All closed queries remain closed for the duration of the program unless T&E specifies otherwise.

While data crawling may continue during a program evaluation, models applied to Eval 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 Eval data. With a single exception[19], this rule does not preclude online learning/adaptation during Eval data processing during an evaluation so long as the adaptation information is not reused for subsequent runs of the evaluation collection. So, for example, any adaptation performed during a CLIR evaluation must be redone from scratch during the corresponding CLIR+S (E2E) evaluation. Performers must document the ways their online learning/adaptation approaches incorporates information extracted from the Eval corpus.

No data or annotations may be distributed outside of the MATERIAL Program.

[18] 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

[19] Performers are not allowed to use text Eval data for adaptation of their ASR models to the speech Eval data.

6.6 Structure of Datasets Released to Performers

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

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

<EvalPeriod> ::= { BASE | OPTION1 | OPTION2 | ... }

<LangID> ::= { 1A | 1B | 1S | ... }

<DatasetName> ::= { DEV | ANALYSIS1 | EVAL1 | ... }

<DocID> ::= MATERIAL_<EvalPeriod>-<LangID>_<DocumentNumber>

<DocumentNumber> is an uninformative 8-digit random number that we assigned to the document.

An example of a legal DocID would be `MATERIAL_BASE-1A_12345678`.

7 File Formats and Their Interpretation

NIST has implemented a scoring tool[20] to calculate scores for tasks listed in section 1. 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.

7.1 Query Format

A query will consist of a query string (a word string), followed by a colon, followed by a domain specification, with no extra punctuation, periods, spaces, or tabs.

Domain and subdomain names are alphabetic, with words separated by a hyphen, and with each word having an initial capital letter.

[20] NIST will make public the scoring tool for performers to use.

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

Domain = [-,a-zA-Z]

Query ::= QueryString:Domain

Here are two examples:

```
music:Lifestyle
```

```
ebola:Physical-And-Mental-Health
```

There are three types of queries:

- 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.

Refer to the MATERIAL Program Query Language Specification Document for a complete description of the query syntax including what is allowed and not allowed.

7.2 System Output Format

7.2.1 CLIR System Output Format

If the task is CLIR, there will be one file per query. The name of these files must match the name of the corresponding reference files. The NIST scoring server will name the reference files using the query ID:

<QueryID>.tsv

For example:

```
query00043.tsv
```

The file content will have one line for each document along with the hard decision and confidence factor that the system assigned to that document for the given query. Those lines will be formatted as follows:

<DocID><tb><HardDecision><tb><ConfidenceFactor[21]>

Assuming the dataset has 4 documents, a legal example of the `query000043.tsv` would be:

```
MATERIAL_BASE-1A_12345678  Y  0.85
MATERIAL_BASE-1A_23456789  Y  0.840
MATERIAL_BASE-1A_34567890  Y  0.840
MATERIAL_BASE-1A_45678901  N  0.5
```

7.2.2 Domain Identification System Output Format

If the Task is DomainID, there will be one file per domain. The name of these files must match the name of the corresponding reference files. The NIST scoring server will name the reference files using the domain ID:

<DomainID>.tsv

[21] Confidence factors are specified in more detail in a later section of this Evaluation Plan.

For example:

`GOV.tsv`

The file content will have one line for each document along with the hard decision and confidence factor that the system assigned to that document for the given domain. Those lines will be formatted as follows:

<DocID><tb><HardDecision><tb><ConfidenceFactor>

Assuming the dataset has 4 documents, a legal example of the `GOV.tsv` would be:

```
MATERIAL_BASE-1A_12345678  Y  0.85
MATERIAL_BASE-1A_23456789  Y  0.840
MATERIAL_BASE-1A_34567890  Y  0.840
MATERIAL_BASE-1A_45678901  N  0.5
```

Since CS documents are not scored, teams should not include them in the system output as these documents would cause to fail validation.

7.2.3 Language Identification System Output Format

If the Task is LangID, there will be one file per language. The name of these files must match the name of the corresponding reference files. The NIST scoring server will name the reference files using the language ID:

<LangID>.tsv

For example:

`1A.tsv`

The file content will have one line for each document along with the hard decision and confidence factor that the system assigned to that document for the given language. Those lines will be formatted as follows:

<DocID><tb><HardDecision><tb><ConfidenceFactor>

Assuming the dataset has 4 documents, a legal example of the `1A.tsv` would be:

```
MATERIAL_BASE-1A_12345678  Y  0.85
MATERIAL_BASE-1A_23456789  Y  0.840
MATERIAL_BASE-1A_34567890  Y  0.840
MATERIAL_BASE-1A_45678901  N  0.5
```

7.2.4 End-to-End System Output Format

If the task is E2E, for each query it will contain:

7.2.4.1 System Output File

This file has a similar format as the one used in the CLIR task but with one additional column containing links to the summary of the corresponding document. This will allow NIST to validate each relevant document has a corresponding summary.

<QueryID>.tsv

The content of this file will be:

<DocID><tb><HardDecision><tb><ConfidenceFactor><tb><Metadata File>

Where:

```
<Metadata File> ::= <TeamID>.<SysLabel>.<QueryID>.<DocID>.json
```

Example:

```
query123.tsv

MATERIAL_BASE-1A_12345678     Y     0.85       FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.json
MATERIAL_BASE-1A_23456789     Y     0.840      FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.json
MATERIAL_BASE-1A_34567890     Y     0.840      FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.json
MATERIAL_BASE-1A_45678901     N     0.5
```

7.2.4.2 Summary Metadata File

<TeamID>.<SysLabel>.<QueryID>.<DocID>.json

This file captures all the metadata of a summary. It is a JSON file and contains the following metadata.

`team_id`: the name of the Performer Team. One of[22]:

`FLAIR, QUICKSTIR, SARAL, SCRIPTS`

`sys_label`: an alphanumeric [a-zA-Z0-9] name that performers assigned to the submission

`uuid`: UUID of the summary

`query_id`: the ID of the query in the Query-Document pair

`document_id`: the ID of the document in the Query-Document pair

`run_name`: the name of the retrieval run

`run_date_time`: the date of the retrieval run

`image_filename`: the filename of the summary image

`content_list`: array of up to 100 words or snippets that correspond to the words and snippets in the summary image. Note that the total number of individual words must be 100 or less for consistency with Section 3.

`instructions`: optional field that contains instructions on how the given summary should be interpreted for the corresponding components in the query-document pair<QueryID>-<DocID>. This information is incorporated into the AMT HIT templates.

7.2.4.3 JSON Schema

The updated JSON schema is provided separately but is repeated here for completeness. In the case of inconsistencies, the separately provided json file is definitive.

```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "id": "http://localhost/summary_spec/v1.1",
  "title": "MATERIAL Summary Specification",
  "description": "JSON Schema for MATERIAL 1S Summaries. Updated: 2018-11-15",
  "type": "object",
  "properties": {
    "team_id": {
      "description": "The Performer Team name; case sensitive.",
      "type": "string",
```

[22] Case sensitive

```
      "enum": ["FLAIR", "QUICKSTIR", "SARAL", "SCRIPTS"]
    },
    "sys_label": {
        "description": "An alphanumeric [a-zA-Z0-9] name that Performer Team
assigns to the submission.",
      "type": "string",
      "pattern": "^[a-zA-Z0-9]+$"
    },
    "uuid": {
      "description": "The UUID of the summary.",
      "type": "string",
      "pattern":"^[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-
[a-fA-F0-9]{12}$"
    },
    "query_id": {
      "description": "The ID of the query in the Query-Document pair.",
      "type": "string"
    },
    "document_id": {
      "description": "The ID of the document in the Query-Document pair.",
      "type": "string"
    },
    "run_name": {
      "description": "The Performer Team assigned name of the retrieval run.",
      "type": "string"
    },
    "run_date_time": {
      "description": "The ISO-8601/RFC-3339 date-time of the retrieval run.",
      "type": "string",
      "format": "date-time"
    },
    "image_filename": {
       "description": "The filename of the domain summary image; one of jpg or
png.",
      "type": "string",
      "pattern": ".jpg|.png$"
    },
    "content_list": {
       "description": "The list of words and/or snippets in the summary. See
the MATERIAL Evaluation Plan for an explanation of the size limits and other
constraints on summary content.",
      "type": "array",
      "minItems": 1,
      "maxItems": 100,
      "items": {
        "type": "string"
       }
    },
    "instructions": {
       "description": "Optional instructions provided by the Performer Team.
See the MATERIAL Evaluation plan for guidance on the content of the
instructions.",
      "component_1": {"type": "string"},
      "component_2": {"type": "string"},
      "domain": {"type": "string"}
```

```
    }
  },
  "required": [
    "team_id",
    "sys_label",
    "uuid",
    "query_id",
    "document_id",
    "run_name",
    "run_date_time",
    "image_filename",
    "content_list"
  ],
  "additionalProperties": false
}
```

7.2.4.4 Notes and Examples

The corresponding summary image filename (i.e., the values in the `image_filename` field) must match the basename of the metadata file and has to be in the form:

<TeamID>.<SysLabel>.<QueryID>.<DocID>

So a valid metadata file for a conjunct query like:
`bribing+,"panel of judges" GOV`

would be named something like:
`FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.json`

and might look like:

```
{
  "team_id": "FLAIR",
  "sys_label": "MySystem1",
  "uuid": "dc91049e-d934-40d7-a175-5af32cbabbca",
  "query_id": "query123",
  "document_id": "MATERIAL_BASE-1A_12345678",
  "run_name": "dummy1",
  "run_date_time": "2018-03-27T13:51:00-05:00",
  "image_filename": "FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.jpg",
  "content_list": [
    "thing", "habit", "even", "will", "look", "group", "situation",
    "property", "easily", "usually", "behavior", "trivial",
    "people", "convoy", "police", "university", "school",
    "community", "student", "brawl", "bribery", "secret payment",
    "this is an example snippet contains the term panel of judges",
    "and another that contains panels of judges",
    "panel", "judges"
  ]
}
```

7.2.4.5 Summary Image

All actual document summaries, whatever presentation form the Performer Teams have settled on for themselves, should be provided as a jpg or png image. The image should be exactly 1024p wide and be at most 768p high.

By using an image we avoid any issues associated with rendering a Perform Team Summary in a browser window. The images might get uniformly scaled for display but they will not otherwise be transformed.

7.2.4.6 General Instructions for MQ Subtypes and Domain relevance

In addition to instructions specific to a particular query-summary pair described above, each team is allowed to submit general instructions for each MQ subtype and domain as described in the 07/17/2018 team Basecamp Project posts “HIT screenshots and request for instructions”. More specifically, general instructions can be submitted for each query type:

- Morphological
- Extended morphological (e.g. `"<wearing> slippers"` or `"coconut <trees>"`)
- Lexical
- Conceptual
- EXAMPLE_OF
- extended "EXAMPLE_OF" (e.g. `"tongue of EXAMPLE_OF(animal)"` or `"EXAMPLE_OF(fruit) juice"`)
- Domain

Those instructions will not be included in the JSON metadata files and will instead be submitted directly to IARPA/T&E via each team's Basecamp Project. They should be formatted as text with optional embedded HTML tags.

7.3 Reference Format

7.3.1 CLIR Reference Format

The reference files for the CLIR task on the scoring server 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 no confidence factor field.

Assuming the dataset has 4 documents, a legal example of the CLIR reference file for `query000043` would be:

```
MATERIAL_BASE-1A_12345678  Y
MATERIAL_BASE-1A_52763409  Y
MATERIAL_BASE-1A_32198765  Y
MATERIAL_BASE-1A_98765432  N
```

7.3.2 Domain Identification Reference Format

The reference files for the DomainID task on the scoring server will be named as:

<DomainID>.tsv

For example:

`GOV.tsv`

The format of the DomainID reference is similar to that of the DomainID system output format except no confidence factor field.

Assuming the dataset has 4 documents, a legal example of the DomainID reference file for domain `Government-And-Politics` would be:

```
MATERIAL_BASE-1A_12345678  Y
MATERIAL_BASE-1A_52763409  Y
MATERIAL_BASE-1A_32198765  Y
MATERIAL_BASE-1A_98765432  N
```

7.3.3 Language Identification Reference Format

The reference files for the LanguageID task follows a similar format as DomainID and will be named as:

<LangID>.tsv

For example:

`1A.tsv`

The format of the LangID reference is similar to that of the LangID system output format except no confidence factor field.

Assuming the dataset has 4 documents, a legal example of the LangID reference file for language `1A` would be:

```
MATERIAL_BASE-1A_12345678  Y
MATERIAL_BASE-1A_52763409  Y
MATERIAL_BASE-1A_32198765  Y
MATERIAL_BASE-1A_98765432  N
```

7.4 Confidence Factors

MATERIAL CLIR 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.

8 Evaluation Scoring Server

NIST will provide an automated scoring server for the MATERIAL evaluation. To make a submission, performers must sign up for an evaluation account via the instructions below:

1. Go to https://material.nist.gov and sign up (fill in the appropriate fields).
2. If you are a team/site PI, send an email to material_poc@nist.gov to say you are PI of <team/site>. We will verify and grant you permission to create team/site. After permission is granted, you can log back in to create site/team.
3. If you are **not** team/site PI, you can request to join an existing site/team (from the drop-down list). An email will be sent to the PI of the site/team you want to join to ask him/her to grant your approval. If the site/team does not exist yet (not in the drop-down list), you have to check back later. How later? Until your PI creates your team so it's best if you wait until your PI is done first.

8.1 Submission Limit and Data Release Schedule

Teams can submit their system output on the various datasets for scoring to help their system development. Each dataset, however, has a submission limit as given in Table 8. When a new document or query pack is released, the server will score only the superset. For example, if EVAL2 and QUERY2 are released, submission should be for EVAL1+EVAL2 on all the queries (QUERY1+QUERY2) released so far, not for EVAL1 alone or EVAL2 alone. Please refer to the base period schedule document that specifies the exact dates for the various cycles.

During the evaluation week, teams can submit up to 5 submissions where one must be designated as *primary*. Primary submissions will be used to compare across teams and assessed by human judges in the case of E2E task. Submissions made during the evaluation week will not receive any score feedback.

In the case of CLIR and E2E, there should be one primary E2E following E2E file format and up to four contrastive CLIR following CLIR file format. There is no need to submit a CLIR primary since the CLIR results will be computed from the E2E primary.

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 the sections below.

Period	Stage	Event/Date	Allowable Input	Number of Submissions Per Week	Results Displayed
BASE	Practice Language	**Program Kickoff** **ID of 1A/1B Release** **ID of Domains X Release** **Build Packs Release** **10/18/17**			
		Q1/A1/D/E1 Release **11/20/17**	DomainID X / Q1 on A1	200	yes

			DomainID X / Q1 on D	200	yes
			DomainID X / Q1 on (A1+D)	100	yes
			DomainID X / Q1 on E1	1	limited
		PM Site Visit 02/21/17			
		ID of Domains Y Release 03/09/18	DomainID XY / Q1 on A1	200	yes
			DomainID XY / Q1 on D	200	yes
			DomainID XY / Q1 on (A1+D)	100	yes
			DomainID XY / Q1 on E1	1	limited
		Q2/E2 Release 04/06/18	DomainID XY / Q2 on D	200	yes
			DomainID XY / Q2 on (E1+E2)	1	limited
		CLIR Eval Week 05/14/18 - 05/18/18	DomainID XY / Q2 on D	200	yes
			DomainID XY / Q2 on (E1+E2)	5 (4 + 1)	no[23]
		ID of Domain Z Release 05/18/18	DomainID XYZ / Q2 on D	200	yes
		A2 Release 05/22/18	DomainID XYZ / Q2 on D	200	yes
		CLIR/DomainID Results Release 05/25/18			
		Q3/E3 Release 07/05/18	LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
		CLIR+S Dry Run Decryption for E3 07/24/18 - 07/27/18	LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	1	no
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
		CLIR+S Eval Week 08/06/18 - 08/10/18	LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	5 (4 CLIR+1)	no
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	1	yes
		A3 Release 08/14/18	LangID / DomainID XYZ / (Q2+Q3) on (A1+A2+A3)	1	yes
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	1	yes
		MTurk Assessment Sept 2018			
		CLIR+S Results Release Oct 2018			
	Surprise Language	**ID of 1S Release ID of Domains X Release Build Pack Release Q1/D/A1 Release 09/05/18**	LangID / DomainID X / Q1 on A1	200	yes
			LangID / DomainID X / Q1 on D	200	yes
			LangID / DomainID X / Q1 on (A1+D)	100	yes
		1S Kickoff Sept 26-27, 2018			
		ID of Domains Y Release 10/03/18	LangID / DomainID XY / Q1 on D	200	yes

[23] Top level results will be released a few days after the CLIR Eval Week has ended. Detailed results (breakdown by mode, genre, query type, etc. minus the E2E results which requires MTurk assessment and takes longer) will be released a few days after the CLIR+S Eval Week has ended.

		Q2/E1/E2 Release 11/01/18	LangID / DomainID XY / Q2 on D	200	yes
		CLIR Sanity Check CLIR+S Dry Run (Validation only) 11/05/18 - 11/09/18	LangID / DomainID XY / Q2 on D	200	yes
			LangID / DomainID XY / Q2 on (E1+E2)	5	no
		A2 Release 11/13/18	LangID / DomainID XY / Q2 on (A1+A2)	1	yes
			LangID / DomainID XY / Q2 on D	200	yes
		ID of Domain Z Release 11/30/18	LangID / DomainID XYZ / Q2 on (A1+A2)	1	yes
			LangID / DomainID XYZ / Q2 on D	200	yes
		Q3/E3 Release 01/11/19	LangID / DomainID XYZ / Q2 on (A1+A2)	1	yes
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / Q2 on (A1+A2)	1	yes
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
		CLIR+S Eval Week 01/14/19 - 01/18/19	LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	5	no
			LangID / DomainID XYZ / (Q2+Q3) on (A1+A2)	1	yes
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	1	yes
		A3 Release 01/22/19	LangID / DomainID XYZ / (Q2+Q3) on (A1+A2+A3)	1	yes
			LangID / DomainID XYZ / (Q2+Q3) on D	200	yes
			LangID / DomainID XYZ / (Q2+Q3) on (E1+E2+E3)	1	yes
		MTurk Assessment Feb 2019			
		CLIR+S Results Release Mar 2019			
		PM Site Visit 03/18/19			
		BP Final Report and Deliverables Submitted 03/29/19			
OP1		**Notification of OP1 Award Apr 2019**			
		OP1 Kickoff Meeting May 2019			

Table 8: Base period schedule and submission quota by dataset depending on the timeline.

8.2 Evaluation Submission Format

Currently we have two scoring servers, each with a different front-end interface: web and Google Drive (GD). For E2E submissions, we ask that you submit via GD. For non-E2E submissions, you can submit to either. However, we ask that for a given task you pick one platform rather than splitting submissions across the two because currently they do not communicate with each other, e.g., 2 DomainID submissions to GD and 3 DomainID submissions to web.

If you are submitting via GD, you must rename your file to a particular naming convention so that the backend connecting to GD will know how to process your submission. If you are submitting via the web, the renaming is done for you by selecting the required attributes using the drop-down menu.

8.2.1 NON-E2E SUBMISSIONS

8.2.1.1 Submission via Web

If performers are submitting via the web, each submission will be an archive file named as follows:

<SysLabel>.tgz

<SysLabel> is an alphanumeric [a-zA-Z0-9] that performers assigned to the submission so they can keep track of which system output was submitted.

Prior to uploading the submission file to the scoring server, performers will be asked for information about the submission. The scoring server will attach these information to the submission file to uniquely identify the submission:

<TeamID> = [a-zA-Z0-9] obtained from login information

<Task> ::= { CLIR | E2E | DomainID }, selected from drop-down menu

<SubmissionType> ::= { primary | contrastive }, selected from drop-down menu

<TrainingCondition> ::= { unconstrained }, hard-coded[24]

<EvalPeriod> = see section 6.6, selected from drop-down menu

<LangID> = see section 6.6, selected from drop-down menu

<DatasetName> = see section 6.6, selected from drop-down menu

<Date> = <YYYYMMDD>, obtained from server at submission time

<Timestamp> = <HHMMSS>, obtained from server at submission time

<QuerysetID> ::= { QUERY1, QUERY2, QUERY2QUERY3 }

<DomainID> = automatically selected based on <LangID> and evaluation event date.

8.2.1.2 Submission via GD

If performers are submitting via GD, each submission will be an archive file named as described below. The renaming script distributed by NIST can be used to generate this filename.

<TeamID>_<Task>–<SubmissionType>–<TrainingCondition>–<DomainID|QuerysetID>–<SysLabel>_<EvalPeriod>–<LangID>–<DatasetName>_<Date>_<Timestamp>.tgz

The above fields are described in section 8.2.1.1. For <DomainID>, it will be a list of 3-letter domain ID separated by a hyphen (-)

For example:

```
NIST_DomainID-contrastive-unconstrained-GOV-MIL-BUS-LAW-mybestsystem_BASE-1S-EVAL1EVAL2_20181113_225652.tgz
```

[24] At the end of a period when teams have shared all data resources, teams may be asked to run a "constrained" training condition utilizing the same shared resources to allow algorithmic comparison.

8.2.1.3 Packing System Output into Submission File

System output files should be packed into a submission file. There should be no parent directory when the submission archive file is untarred. The tar command should be:

```
> tar <MySubmissionLabel>.tgz query*.tsv
```

The server will validate the submission file content to make sure the system output files conform to the format described in section 7.2.

8.2.2 E2E Submissions

Due to size limitation of the web, only GD can be used for E2E submissions (specifically to `E2E/input` directory).

A complete E2E submission will consist of a collection of individual directories each of which will contain all submission files corresponding to that query in a subfolder with the name <QueryID>, e.g.:

```
./query123/
      ./query123.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.jpg

./query45/
      ./query45.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_11223344.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_11223344.jpg
```

A single zipped TAR `<MySubmissionLabel>.tgz` that will contain all query subdirectories. The renaming script previously distributed by NIST can be used to generate `<MySubmissionLabel>`. The query-specific directories `<QueryID>)` will be collected together as follows:

```
tar zcvf <MySubmissionLabel>.tgz *
```

8.3 Reporting Scores

This section describes the analyses and scores that will be reported for the various kinds of evaluations[25].

8.3.1 Reporting Scores for CLIR and End-to-End Evaluations

In addition to overall results, results on various factors (e.g., genre, query type, etc.)[26] will also be reported. We expect such factors will include various characteristics of queries such as the domain, 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 teams will also get these breakdowns for the Eval datasets. However, once the

[25] The scoring server currently only calculates query weighted and document weighted AQWV for the entire dataset. Many of the analyses described in this section will be implemented at a later time.

[26] Other results dissections are possible.

development cycle ends and evaluation cycle starts, teams will only receive top level results on the Eval datasets.

Domains will not, however, be analyzed for the Conversational Speech genre (in effect, we will treat Conversational Speech documents as relevant to all domains).

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 $P_{FA} = 0$. NIST is also planning to compute another version of AQWV where queries with no relevant documents are removed prior to scoring.

8.3.2 Reporting Scores for Domain Identification

Identifying documents that are relevant to a query requires a system to identify two separate things: (1) the domain of a document, (2) the relevance of the query string. In order to give performers more detailed feedback, at the time of the official CLIR evaluation the performer's system is also to do a separate run and determine which domain(s) are relevant to each document in the dataset. Then for each of the domains in the evaluation, the system is to generate a *Domain Output File* that lists all the documents that the system deems to be relevant to that domain, along with a confidence factor (as in section 7.4) indicating how sure the system is that the document is relevant to the domain.

The scoring returned for each domain output file will be four numbers:

- the percentage of true positives (documents that are relevant to the domain that the system also deemed relevant)
- the percentage of misses (documents that are relevant to the domain but which the system deemed not relevant)
- the percentage of false alarms (documents that are not relevant to the domain but which the system deemed relevant)
- the percentage of true negatives (documents that are not relevant to the domain that the system also deemed not relevant)

This will only be run and scored at the time of the official CLIR evaluation. At no time will we tell the performers which fileIDs were in each of the four categories (true positive, miss, false alarm, true negative) for any dataset.

9 Appendix: Descriptions of Domains Released to Date

Government-And-Politics

Working Definition: Anything to do with local, regional, national or international government. Includes national level functions such as the provision of national or international infrastructure and capabilities.

In Scope	Out of Scope
Federal Government, Local Government, Lawmaking, Civil Rights, Government Corruption, Policies relating to national infrastructure (e.g., highway construction, electrical grid, bridge repairs), Taxation, Government Aid (e.g. aid to refugees), Activism, Non-violent protest, Elections, Politician at Work, Governance, Government Budgets, Government Protests, Regional and International Relations, Diplomacy, …	Law enforcement, Law and Judicial Systems, Criminal activity, Terrorism, Military engagements, Military equipment, Military personnel, Defense Spending, Cyber Warfare, Chemical and Biological Warfare, Labor, History, Technology, Human Trafficking, Geography, Finance.

Additional Feedback from the Annotation Teams:

- **Law and Judicial Systems** often come up in the context of government activities related to passing/changing laws, appointment of judges by government bodies/executives, and law-breaking by politicians or other people involved in government. Such sub-topics may have been marked as multiple domains.
- **Military engagements, Military Equipment, Military Personnel and Defense Spending** may arise in the context of parliamentary approval for military action and the role of diplomacy/government policy in international conflicts. Such sub-topics may have been marked as multiple domains.
- **Defense spending** may have come up in the context of taxation and general government spending and been marked as multiple domains.
- **Technology** may come up in the context of government policy relating to new technology, or in relation to infrastructure, and been marked as the Government-and-Politics domain.
- **History** frequently involves government activity, elections, civil rights, regional/international relations, diplomacy, etc. There were no guidelines in terms of differentiating between "current" and "historical" events.
- **Labor**: Issues of industrial action where politicians or government are involved may have been treated as in-scope for Government-and-Politics.
- **Terrorism**: This is treated as in-scope for Government-and-Politics if there was a substantial mention of government involvement.
- **Civil Rights**: There may be a grey area where gender is involved. For example, where a politician is advocating for girls to follow cultural practices on abstinence, or

commenting on education for girls in terms of cultural practices, this could be annotated Lifestyle or Government-and-Politics or both.

Lifestyle

Working Definition: Anything to do with the lives of families and individuals and the activities they engage in, as well as cultural values, norms, practices and expressions.

In Scope	Out of Scope
Daily activities, Celebrations (Weddings, Baptisms, Birthdays, etc.) Food and cooking, Parenting and childcare, Personal transportation, Gardening, Fashion, Personal work and employment, Family and friends, Leisure, Vacations and travel, Realization of cultural practices, Effect of cultural norms on daily life, Effects of poverty or pensions on daily life, …	Public health, Medical facilities, Medical conditions, Disability, Religion, Theology, Fine Arts, Literature, Organized sports (e.g., baseball, soccer), Student loans, Schools and teaching, Vocational training, Celebrities, Movies, Concerts, Video Games, Journalism and Media, Agriculture, Nutrition, Pharmacology, Philosophy, Natural Disasters, Pop Culture, Race, Social Issues, Predatory Lending, Real Estate Industry, Climate, Tourism Industry.

Additional Feedback from the Annotation Teams:

- **Medical conditions**, and **Disability** may arise in the context of their impact on family life or personal work/employment. They are likely to have been treated as in-scope for Lifestyle in this context, and potentially for other domains as well.
- When **Religion** arises in the context of attending church services or religious festivals as a family activity, or the impact of religious beliefs on family life and child care, it is likely to have been treated as in-scope for Lifestyle.
- **Student loans** and **Schools and teaching** are often mentioned in the context of personal work and employment as well as family life (e.g. taking on loans in the hopes of getting a better job, challenges of balancing education with work/family commitments, etc.), or cultural attitudes towards such things. In such contexts, multiple domains may have been selected, including Lifestyle.
- **Nutrition** may have been labeled Lifestyle where it is mentioned in the context of food and cooking.
- **Natural disasters** are often mentioned in the context of their impact on family life. In such a context, Lifestyle and other domains may apply.
- **Race** and **Social issues** are frequently intertwined with cultural values/practices, which have been treated as lifestyle topics. In this context, the Lifestyle domain may have been selected.

- **Personal work and employment** will often have been treated as out-of-scope for Lifestyle, rather than in-scope. For instance, a document about looking for a job overseas would have been treated as out-of-scope.
- Some **Social Issues** are in-scope when they appear in the context of how they affect daily lifestyles, e.g. poverty, pensions.

Business-And-Commerce

Working Definition: All activities and entities associated with economic endeavor.

In Scope	Out of Scope
Tourism industry, Real estate industry, Manufacturing, Retail, Wholesale, Labor, Finance, Employment, Investment, Financial markets, Commodities markets, Restaurants, Banking, Trade, International trade agreements …	Organized sports (e.g., baseball, soccer), Movies, Concerts, Video Games, Journalism and Media, Agriculture, Drug manufacturing, Medical insurance industry

Additional Feedback from the Annotation Teams:

- Organized sports may be marked as Business & Commerce where there is substantial discussion of topics such as players' salaries, team owners' profits, or costs associated with stadium construction.
- Movies or concerts may be marked within this domain when there is substantial discussion of topics like actors' salaries, studio financing, or production/promotion costs.
- Unlicensed enterprise activity such as market stalls or personal financial efforts may have been treated as in scope if there was sufficient discussion of the activity as an economic endeavor.
- Government documents discussing employment or tax with reference to employment may have been considered in scope even if not related to any specific business industry.
- Financial compensation and prizes were considered out of scope unless there was substantial discussion of corporate sponsorship.
- Donations of monetary amounts were considered out of scope unless there was substantial discussion of corporate involvement.
- Agricultural production may have been considered in scope if there was further discussion about the post-production enterprise activities.
- In general, other out of scope topics may also appear in this domain if there was substantial discussion of associated salaries, profits and/or costs.

Law-And-Order

Working Definition: Anything to do with crime, violence or the enforcement of local, regional and national laws.

In Scope	Out of Scope
Types of crime, Executions, Prisons, Prison sentences, Legal aid, Violent protests, Police, Law enforcement, Law and Judicial Systems, Courts and litigation, Criminals, Criminal activity, Terrorism, …	Lawmaking, War, Laws

Additional Feedback from the Annotation Teams:

- Lawmaking and laws were frequently considered in scope for both Law-and-Order and Government-and-Politics. The overlap between these domains is substantial, particularly for national and global-scale law and policy.
- Documents relating to war may have been considered in scope for both Military and Law-and-Order when terrorism and war crimes were discussed.
- Documents relating to law enforcement within the military, which often discussed war, may also have been considered in scope for Law-and-Order.

Physical-And-Mental-Health

Working Definition: Anything to do with the provision of health and wellbeing to a population, as well as causes and correlates that affect health and wellbeing, such as accidents and non-natural disasters. Includes community public health concerns.

In Scope	Out of Scope
Fields of medicine, Drugs and medication, Nutrition, Drug abuse, Prenatal care, Suicide, Cancer, Disability, Primary care, Hospitals and other medical facilities, Medical conditions, Treatments including non-traditional and folk remedies, Hygiene, Public health and safety, …	Natural disasters, Accidents, Death (unless a medical cause of death is specified)

Additional Feedback from the Annotation Teams:

- Natural disasters and accidents may have been included where there was substantial discussion of resulting injuries and/or mental health issues, for example, injuries caused by a car accident, depression or PTSD experienced by earthquake survivors.
- Documents discussing drug abuse and its possible treatments and/or rehabilitation would generally have been treated as in scope, while discussions of drug abuse as a social issue were not.
- Suicide as a social issue was not included in this domain, but if there was content related to the wellbeing of the affected person(s), then the document would generally be considered in scope.
- Documents relating to sports or physical exercise where benefits to physical health were emphasized would generally have been considered in scope.

Military

Working Definition: Anything to do with military capability, activity or entities.

In Scope	Out of Scope
Branches of the military, Military engagements, Military equipment, Military budget, Military training, Military personnel, Military recruitment, Strategies in war, Rescue operations involving military, Defense spending, Role of women in the military, Cyber Warfare, Chemical and Biological Warfare, ...	International security agreements, International peace treaties, police operations, police recruitment and personnel.

Additional Feedback from the Annotation Teams:

- Documents relating to police operations were subject to some initial erroneous annotation as Military. Police operations were considered out of scope for this domain unless there was clear evidence of the involvement of military in the matter, as a separate entity to the police.
- International security, international relations, and other geopolitical issues were not treated as in scope for Military, unless military budget and structure were addressed.

Sports

Working Definition: Anything to do with sports activities and entities.

In Scope	Out of Scope

Types of sport, Sports teams, University sports, Sportsmen/women, Sports competitions, Sports venues/stadia, Sports organizations, Sports equipment, Corruption in sports, Match fixing, ...	News tangentially related to sports tournaments; Oblique references to sports such as medical recommendations on exercise or passing mention of attendance at a sports event

Additional Feedback from the Annotation Teams:

- Incidental/tangential mentions of the Olympic games, or other large-scale sports tournaments, that did not discuss the sports events in detail, were considered out of scope.
- A "detailed description" of a sports event was taken to be one which described a sports match, scores, and/or sports players.
- Documents which discussed sports team supporters or sports stadium structure were considered in scope.
- In general, the presence of this domain in a document was reasonably clear and unlikely to be misinterpreted.

Religion

Working Definition: All aspects of personal and organizational belief systems and practices that relate humanity to what the adherents of that religion consider to be ultimate reality.

In Scope	Out of Scope
Theology, religious beliefs/practices/ constraints, individual religions, religious festivals, religious texts, religious places/buildings, religious education, history of religion.	Terrorism, religious artwork, legal concerns involving religious figures, religious constraints on health-related issues.

Additional Feedback from the Annotation Teams:

- A religion does not necessarily include a belief in the existence of a God or Gods. Theology is included in this domain and covers the study of religious beliefs.
- **Terrorism** was subject to some initial erroneous annotation as Religion. It was decided that documents mentioning terrorism were only in scope if there was a substantial mention of religious influencers, or any form of military conflict with a basis in religious conflict.
- **Legal regulations** affecting religious practices or observance in public, and religious individuals or communities in conflict with law enforcement personnel were likely to be considered in scope for both Law & Order and Religion, provided there were predominant mentions of religious persons and their involvement.

- The relationship between **government** and religious organisations was considered in scope as long as there was comprehensive discussion of the religious aspect.
- The impact of religion on **lifestyle choices** at a family or community level were considered in scope if religion was more than an incidental mention. These activities include religious constraints on cooking, driving cars, or other activities that may be constrained by a religious holiday or festival.
- Religion in the context of **family activities**, or the impact of religious beliefs on family life and child care, was treated as in scope for both the Lifestyle and Religion domains.
- **Race** and **Social issues** are frequently intertwined with cultural values/practices, which have been treated as Lifestyle domain topics. In this context, religion was only considered in scope if there was detailed and specific discussion of religious issues.
- The impact of religious belief on **physical health**, for example, objections to certain medical procedures or treatments based on religious beliefs were considered out of scope, unless there were several mentions or significant discussion of religion.
- **Performances** or **events** related to religious holidays or festivals were only considered in scope if the document contained in-depth discussion of the religious aspects.
- The **history of religion** was considered in scope unless religion was deemed an incidental mention. Conversely, **religious artwork** was only considered in scope if there were substantial references to religion or religious symbols.