GUIDELINES for PARTICIPANTS

DOCUMENT COLLECTIONS

Registered participants can download the corpora from the CLEF website (registration form and end-user agreement must be first filled in).
Here below you find the document collections that will be used for each target language:

TARGET LANGUAGE	COLLECTION	PERIOD
Bulgarian (BG)	Sega	2002
	Standart	2002
Dutch (NL)	NRC Handelsblad	1994/1995
	Algemeen Dagblad	1994/1995
English (EN)	Los Angeles Times	1994
	Glasgow Herald	1995
French (FR)	Le Monde	1994
	Le Monde	1995
	French SDA	1994
	French SDA	1995
Germany (DE)	Frankfurter Rundschau	1994
	Der Spiegel	1994/1995
	German SDA	1994
	German SDA	1995
Italian (IT)	La Stampa	1994
	Italian SDA	1994
	Italian SDA	1995
Portuguese (PT)	Público	1994
	Público	1995
	Folha de São Paulo	1994
	Folha de São Paulo	1995
Spanish (ES)	EFE	1994
	EFE	1995

QUESTIONS

This year we propose four types of questions:

1. **factoid questions:** fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.
Examples:

- *Who was Lisa Marie Presley's father?*
- *What year did the Second World War finish?*
- *What is the capital of Japan?*
- *What party did Hitler belong to?*
- *How many monotheistic religions are there in the world?*
- *What is the most-read Italian daily?*

2. **definition questions:** questions like *"What/Who is X?".* This year definition questions won't concern only people and organisations, but also objects, natural phenomena, legal procedures etc.
Examples:

- *Who is Lisa Marie Presley?*
- *What is Amnesty International?*
- *What is the FDA?*
- *What is Swiss army knife?*
- *What is a router?*
- *What is a tsunami?*
- *What is DSL?*
- *What is impeachment?*

3. **temporally restricted questions:** three different types of temporal restriction may be applied to any kind of questions, i.e.:

- restriction by **DATE**, e.g. *Who was the US president in 1962;*
- restriction by **PERIOD**, e.g. *How many cars were sold in Spain between 1980 and 1995?*
- restriction by **EVENT,** e.g. *Where did Michael Milken study before enrolling in the university of Pennsylvania?*

4. **list questions:** questions which require a list of items as answers, for example:

- *Name works by Tolstoy.*
- *Which European cities have hosted the Olympic Games?*

Test sets will be made up of 200 questions, most of which will be temporally unrestricted factoids. There will be around 40 definitions and 40 temporally restricted questions.
Some questions may even have no answer in the document collection, and in this case the correct response is a blank string with docid "NIL". A question is assumed to have no right answer when neither human assessors nor participating systems can find one.
NB: this year the question type will not be provided to the systems, which will have to deal with the questions without this piece of information.

QUESTION FORMAT

Test sets will be formatted as plain text files (UTF-8 encoded), with one question per line.
Each line will be structured as follows:

- Source language
- Target language
- Question number (4 digits)
- Question (UTF-8 encoded string)

i.e.:

BG|DE|EN|ES|FI|FR|IN|IT|NL|PT|RO	BG|DE|EN|ES|FI|FR|IT|NL|PT	4 digits	QUESTIONS

- The task is described in the first and second column.
- Question number, indicated in the third column, ranges between 0001 and 0200.
- Columns are separated by a tab.

Example:
The first three questions in the EN-ES test set – i.e. English questions that hit a Spanish document collection - might be represented as follows:

EN ES 0001 In which country is the Cape of Good Hope?
EN ES 0002 What is Amtrack?
EN ES 0003 Who was the president of the United States in 1961?

ANSWER FORMAT

Each participating group will be allowed to participate in any task. Anyhow, we encourage participants (especially "veterans") to consider questions and target languages other than their own language and English.

Participating teams must return **one or more (up to a maximum of 10) exact answers per question**, and **up to two runs.** All questions must be answered and no partial submissions will be accepted.

As a novelty, **one or more text snippets (up to a maximum of 10) are required to support the answer.**

An answer is basically structured as a {*DOCID/ answer-string/ text snippets*} sequence, where:

- the *document-id* is the unique identifier of a document that supports the answer.
- the *answer-string* contains **nothing more than a complete and exact answer** (the minimum amount of information needed). If there are more than one answer to the same questions, they will be put in different answer lines, listed one below the other, according to their confidence score.
- the *text snippets* will be used as a justification by human assessors in order to determine the correctness of the answers. There can be more than one text snippet per question (up to a maximum of 10), and they will be placed in the same answer line, one after the other, and separated by a tab.

There are **no particular restrictions** on the length of an **answer-string** (which is normally very short), but **unnecessary pieces of information will be penalized**, since the answer will be marked as **non-exact**.
Because **definition questions** may have long strings as answers, assessors will be less demanding in judging their exactness: assessors will mainly focus on their responsiveness and usefulness.

RUN SUBMISSION

Each submitted run is a single file. Each answer is **a single UTF-8 encoded line**, carrying the following information:

–question number
–team ID+year (last two digits, i.e. 06)
–the identifier of the run (1 or 2)
–task identifier (source language id+target language ID, e.g. ENES)
–confidence score
–docid
–exact answer-string
–justification snippets

Each piece of information is given in different fields (columns separated by a tab), as in the following example:

Field 1	*Field 2*	*Field 3*	*Field 4*	*Field 5*	*Field 6*	*Field 7-n*
0001	irst061enes	0.861	EFE19940427-16057	Sudáfrica	La historia escrita de Sudáfrica comenzó el 6 de abril, de 1652, cuando van Riebeeck estableció una puesto de avituallamiento en el cabo de Buena Esperanza.	

where:

- in the *first field* the **question number** is indicated: answers must be returned in the same ascending (increasing) order in which questions appear in the test set, i.e. from 0001 to 0200.

- the *second field* describes, in one single string:

  ~ the **name of the participating team** (arbitrary sequence of four ASCII characters)
  ~ the **current year** (06 stands for 2006)
  ~ the **number of the run** (1 if it is the first one, or 2 if it is the second one)

~ the **task identifier** (including both source and target languages, as in the test set).

Clearly, the content of this field never changes within the same submission file. Each submission file must be named after this column, with a .txt extension. For instance, the lines in the examples above would be part of the file *irst061enes.txt*.

- in the *third field* the **confidence score** is indicated. This is a mandatory integer or floating point value (maximum length is 8 characters) that can range between 0 and 1, inclusive, where 0 means that the system has no evidence of the correctness of the answer, and 1 means that the system is absolutely confident about the correctness of the answer. If a system returns integer or floating point confidence values that are higher than 1, it must normalize them for each response. If a system does not produce any score number, it must return a default score equal to 0 (zero). Score value will be used in a second, additional evaluation (the main measure is accuracy) in order to test systems' self-evaluation ability.

- the *fourth field* contains the **DOCID of the document** that supports the text snippets. Some questions may not have any known response in the document collection: in that case the correct answer would be a blank answer-string, and the string NIL would replace the docid.

- the *fifth field* contains the **answer-string**, that is blank if the DOCID is NIL. There can be no line breaks in the answer-string.

  Human assessors will judge both correctness and exactness of the submitted answers, so unnecessary additional information will be marked as non-exact. The same judgment will be assigned to responses that lack significant bits of information.

- From the *sixth field on*, the **text snippets supporting the exact answer** are given. The text snippets are put one next to the other, separated by a tab. As a consequence, they cannot contain any tab or line break. Snippets are required to be substrings of the specified documents, subject only to whitespaces conversion. In whitespace conversion all contiguous sequences of one ore more whitespace characters (CR, LF, SPACE or TAB) within the snippet may be converted to a single SPACE.

  Snippets should provide enough context to justify the exact answer suggested. Systems may return **up to 10** snippets. Snippets for a given response should be **a set of sentences up to 500 bytes in total.** Unnecessarily long snippets, i.e. those that do not meet this requirement, might be judged as not supporting their answers - this decision it up to the assessor.

Each **answer must have at least one supporting snippet**, up to a maximum of 10. If the value in the docid field is NIL, the sixth field (and following) will be empty.

Generally speaking, participants should follow the "no-empty-slot" rule, except when the value in the DOCID field is NIL. There must be **one tab between fields** and a **line break after each response line.**

If there are more than one answer to a question, they are listed one after the other, according to their confidence score. In this case, the first and second field do not change, meanwhile DOCID, exact answer and supporting snippets are different.

EVALUATION

The files submitted by participants in all tasks will be manually judged by native speaking assessors. Each language coordination group will guarantee the evaluation of **at least three answers.**

Further details about evaluation will be provided later.

IMPORTANT DATES

Question sets for all tasks will be released on the **June 5th**.

Participants will have one week to process them and return their results (by **June 12th**).

Instructions concerning the results submission procedure will be given when the track will start. Individual results will be released to each participating group from July 17th