# CLEF 2003 Question Answering Track:
# Guidelines for the monolingual and bilingual tasks

## Summary

The coordinators of the four tasks (monolingual Italian, monolingual Spanish, monolingual Dutch and cross-language bilingual) will provide 200 questions that seek short, fact-based answers. Each participant may produce up to two runs, and all submitted runs will be judged. The whole process must be completely automatic: your system must retrieve answers without any sort of manual intervention. **No changes** can be made to any component of your system or any resource used by it between the time you download the questions and you submit your results. All questions must be processed from the same initial state (i. e., your system may not adapt to test questions that have already been processed).

You may return up to three responses per question, and answers should be ordered by confidence, which is expressed in the 'answer rank' field and optionally in the 'score' field of the response unit. A response is either a [answer-string, docid] pair or the string "NIL". The "NIL" string will be judged correct if there is no answer known to exist in the document collection; otherwise it will be judged as incorrect. Unlike last year's TREC competition, where the answer-string had to contain nothing other than the answer, we accept two kinds of response: the exact answer or a 50 bytes long string that must contain the exact answer. Participants can decide which criterion to use for each run.
The pair will be marked as follows:

- incorrect: the answer-string does not contain a correct answer or the answer is not responsive;
- unsupported: the answer-string contains a correct answer but the document returned does not support that answer;
- non-exact: the answer-string contains a correct answer and the document supports that answer, but the string is missing bits of the answer or contains more than just the answer (just for the exact answer runs);
- correct: the answer-string consists of exactly a correct answer (or contains the correct answer within the 50 bytes long string) and that answer is supported by the document returned.

All judgments are in the opinion of human accessors.

## Document set

The documents that constitute the database for the tasks are drawn from a collection of newspaper and news agency articles of the year 1994.
The only portions of text you should consider are those tagged with <DOCID>, <TITLE>, <TEXT> and <TABLE>. We strongly require not to consider other tags such as <PERSON>, <LOCATION>, <AUTHOR> or whatever, because they could give additional information to your system and thus facilitate the search for answers.

In order to have access to the test collection, a registration form and the relevant data release forms must be first compiled, signed and sent (by express mail) to the CLEF coordinator Carol Peters (see the section HOW TO PARTICIPATE in the main CLEF web site for further details).

## Questions

Systems should retrieve answers in response to 200 fact-based questions. That means they can search for a particular quantity, name, date, place, object and so on.
Organizers, who generated the questions, avoided 'subjective' (e. g. "Who is the most important person in Italy"), 'definitional' (e. g. "What is a car?"), 'double' (e. g. "Which Prime Ministers visited the poorest country in the world in 1994?") and multiple instance questions (e. g. "Name at least three titles of Miller's plays"). The latter category constituted a particular 'list task' in last year's TREC, which our competition lacks. Actually, it would have been quite difficoult to define how a system could extract several instances and put them together into an exact answer-string. Some questions may not have a (known) answer in the document set, so the correct response for them is an empty string (with docid "NIL"). A question will be assumed to have no correct answer in the collection if our assessors do not find an answer during the answer verification phase **and** no participant returns a correct response.

The questions for the task will be available for downloading from the CLEF web site on May 7. Remember that you must freeze your system before downloading the questions. Downloadable resources for each task could be available in this site, including a variety of data that you may use to develop and train your QA system.

## Response Unit

You may return up to two runs and up to three responses per question.
Answers should be ordered by confidence, which is expressed in the 'answer rank' field and optionally in the "score" field of the response unit. A response is either a [answer-string, docid] pair or the string "NIL".
Unlike last year's TREC competition, where the answer-string had to contain nothing other than the answer, we accept two kinds of response: the exact answer or a 50 bytes long string that must contain the exact answer. Participants can decide which criterion to use for each run. For example, if you submit two runs you can produce your responses using two different methods, or the same one for both.

Participants will have one week between the time the questions are released and the results are due back. The primary week will be May 7- May 15. However there may be participants for whom that week will not work. In this case, by prior arrangement with each task coordinator, participant may choose another one week period. Participants who will be using an earlier week should communicate that as soon as possible.

The basic unit of a response for the task is the [answer-string, docid] pair. As an example, consider the question "What is the capital of Spain?". If the submitted run has retrieved just **exact answers** (and nothing else in the answer-string), the following can be considered correct answers:

1. Madrid
2. madrid

while none of the following are correct, exact answers:

a) with 3,000,000 inhabitants, Madrid
b) beautiful Madrid
c) she told him: Madrid is a
d) Madri
e) Sevilla

On the other hand, if the submitted run has retrieved **50 bytes long strings as answers**, responses a), b) and c) must be considered correct.

An answer is considered "responsive" when,for example, it includes units for quantitative responses (e.g., $20 instead of 20) and when it refers to a famous entity itself rather than its replicas or imitations.
As regards locations, assessors will base their judgment on the conventional names that appear in atlases. For example, some locations in Italian require a specific name (e. g. "Mar Nero", and not only "Nero", which would be incorrect), while others can have different forms (e. g. "Mar Tirreno", and also "Tirreno").
As regards the date of specific events that ended in the past, both day and year are normally required (unless the question refers only to the year), but if the day cannot be retrieved, the year is sufficient. For example, if a system answers the question "When did Napoleon die?" returning "5$^{th}$ May", it will be judged as incorrect. On the other hand, both "May 5, 1821" and "1821" could be correct exact answers.
As with correctness, exactness will be in the opinion of human assessors. In case of doubtful evaluation, assessors will discuss to find an agreement.

The interpretation of the [answer-string, docid] pair is that answer-string is an answer to the question and docid is the number of a document that justifies that answer. Responses will be judged by human assessors who will assign one of four possible judgments to a response:

- incorrect: the answer-string does not contain a correct answer or the answer is not responsive;
- unsupported: the answer-string contains a correct answer but the document returned does not support that answer;
- non-exact: the answer-string contains a correct answer and the document supports that answer, but the string is missing bits of the answer or contains more than just the answer (just for the exact answer runs);
- correct: the answer-string consists of exactly a correct answer (or contains the correct answer within the 50 bytes long string) and that answer is supported by the document returned.

## Particular cases:

If a wrong answer is given, it could be assessed as correct if the document that correctly supports that answer contains the same mistake, i. e. if the document is itself in error. For example, answer e) above would be considered correct if in the document that supports that answer Sevilla is regarded as the capital of Spain.

If the same document contains the requested answer in different contexts (e. g. Madrid is present once as the capital of Spain, once as a beautiful city, and twice as a polluted place) the answer supported by that document will be considered as correct.

**ANSWERS FORMAT:**

Each submission is a single file. Questions must be returned unranked, in the same order (from 1 to 200) as they have been downloaded. On the other hand, answers to each question must be ranked so that the most confident response is at the first place in the ranking. **Ranking value must be in ascending (increasing) order and must be 1,2, or 3**. Score number, on the other hand, is not compulsory, and if your system does not produce any score number, it must set this field to 0. What counts is the order of the answers. The response unit for each question must have the following format:

qid system run-tag answer rank (score) docid answer-string

where

| | |
|---|---|
| qid | is the question number; |
| system run-tag | is the run id; |
| answer rank | shows that the answers are ordered by confidence, and that the system places the most sure response in the first position; |
| score | is an integer or floating point number. The "score" field is not compulsory, but if a system does not submit any score number, answers will be marked considering the place they occupy in the ranking; |
| docid | is the id of the supporting document or the string "NIL" (no quotes) if no answer is in the collection or if the system cannot find one; |
| answer-string | is the exact answer (a text string with no embedded newlines), or a 50 bytes long string (which must contain the exact answer). If the docid field is "NIL", this column should be empty |

Any amount of white space may be used to separate columns, as long as there is some white space between columns and every column is present (modulo empty answer-string when docid is NIL). The 'system run tag' should be a unique identifier for your group AND for the method used. That is, each run should have a different tag that identifies the group and the method that produced the run (in the example below we propose the letters "ex" to show that the submitted run contains exact answers, but a formal decision has not been taken, yet).
Answer-string cannot contain any line breaks, but should be immediately followed by exactly one line break. Other white space is allowed in answer-string.

Example:

| | | | | | |
|---|---|---|---|---|---|
| 1 | irstex03 | 1 | 4057 | LASTAMPA19941102 | new york |
| 1 | irstex03 | 2 | 3166 | SDA19941407 | Boston |
| 1 | irstex03 | 3 | 233 | LASTAMPA19940506 | chicago |
| 2 | irstex03 | 1 | 1244 | NIL | |
| 2 | irstex03 | 2 | 981 | LASTAMPA19942203 | tom cruise |

3 irstex03 1 4429 SDA19942211 Animal Farm
[.....]

Answer ranking replaces the procedure of last year's TREC, where questions had to be ordered by confidence in the response. If a system submits score numbers, they must be in descending (non increasing) order and must not equal 0. If a system does not submit any "score" number in the response unit (and it is not compulsory to do that), the score assigned to each question will be the reciprocal of the rank for the first response to be judged correct, or 0 if no response is judged correct.

## Evaluation

The exact answer run and the 50 bytes long string answer run constitute **two different tasks** of the competition, and each participant can decide whether to participate or not in both tasks. Note that you can submit up to two runs for the QA Track, and that it is not allowed to use both the answering criteria in the same run. For example, if you participate with two runs, you can submit either one run of exact answers and the other one with 50 bytes long strings, or both the runs with exact answers, or both the runs with 50 bytes long strings.
Each task will be assessed separately, and CLEF assessors will build two separate result classifications.
The total score will be the mean score over all questions.

## Restrictions

No manual processing of questions, answers or any other part of a system is allowed in this track: all processing must be **fully automatic**. No part of the system can be changed in any matter after you have downloaded the questions from the web site. Your system must process **exclusively** the parts of the document set tagged as we have outlined above. Each question must be processed starting from the same initial state (i.e, no learning from previous questions, since that confuses evaluation results).

## Timetable

Registration open: 15th January 2003

Data Release: 30th January 2003

Question Release: 7th May 2003

Submission of runs by participants: 15th May 2003

Release of Individual results: 1st July 2003

Submission of papers for Working Notes: 20th July 2003

Workshop: 21st – 22nd August 2003