Elements of a traditional laboratory IR evaluation experiment (after Cranfield):

- A system:
  that is, a set of methods and procedures
  (whether human or machine)
  for indexing and searching;
- A collection of documents;
- A collection of requests representing information needs;
- An experimental design;
- A basic output evaluation process:
  providing *relevance* judgements
  on individual documents in relation to request/needs;
- Performance measures derived from the relevance information.

Portable test collections

- Collections of test material
  (documents, requests, relevance judgements)
  made available for use by other researchers
  (prime example: the Cranfield collection)
- ... so that tests could be conducted in laboratories
  without the necessity for finding real documents
  — or real users!
- These collections used far outside the conditions for which they were originally designed.

'Ideal' test collection

- Proposed in the 1970s by the UK IR research community (Sparck Jones *et al.*)
- ... as a large, general-purpose, portable test collection.
- But abandoned as too expensive.

TREC — the Text REtrieval Conference

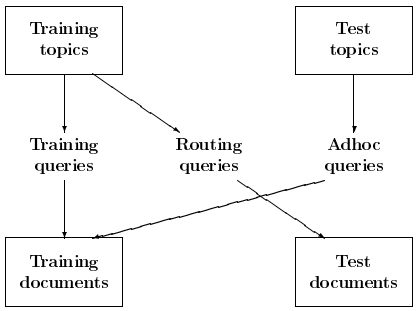
- Competition to drive up the level of performance (à la MUC)
- ... with regular annual rounds.
- Design based on 'ideal' collection design.

Basic TREC methods

- Accumulating collection of documents, with additional set every year
  (typically $\frac{3}{4}$m documents ≈ 2Gb for each test);
- ditto requests ('topics' — 50 per test).
- Old topics/old documents have relevance judgements from previous rounds.
- 'Adhoc' test of new topics against old documents;
- 'routing' test of old topics against new documents.
- Relevance judgements on pooled output from participants;
- various standard performance measures (of the recall–precision type) based on ranked output.

TREC: Basic design for each round

(In TREC terminology, the original specification of an information need is called a *topic*. A *query* is something like a search formulation.)



TREC and interaction

Some simple manual intervention in early TRECs.

How to do genuine interactive searching in lab tests?

Some conflicts:

- In a lab test, we try to separate the different factors...
- × ... but in interactive searching, the user has access to a portfolio of interactive mechanisms.
- In a lab test, we try to keep the user outside the system...
- × ... but in interactive searching, the user/searcher is inside (part of) the system.
- In a lab test, we can repeat an experiment, with variations, any number of times...
- × ... but in interactive searching, repetition is difficult and expensive and unlikely to produce identical results.

For interactive experiments in TREC:

- Searchers are not the 'real' users
  ('real' users = the assessors, who formulate the topics and assess relevance).
- The searcher may act as an intermediary for the remote user...
- ... or may be given a task to encourage user-like behaviour.
- Either way, the task must be carefully formulated,
- as must the specification of how the outcome is to be evaluated.

TREC–3 interactive track:

- Based on the routing task; searcher acts as intermediary.
- Searcher takes routing topic, searches old database,
- with access to relevance judgements on retrieved items
  (as if the end-user were looking over her shoulder and making judgements).
- Task is to arrive at a good search formulation which can be used on the new document collection.

TREC–4 interactive track:

- Adhoc task: to find as many relevant documents as possible, without too much rubbish,
  in limited time (≈ 30 minutes);
- (subsidiary task: to formulate query for offline searching).
- Wide variety of experiments tried.
- Evaluation measures different from main tasks
  (set recall, set precision, time).
- But the evaluation method was not satisfactory.

Possibilities for TREC–5?

- Abandon attempt to make any part of the interactive task compatible with the main TREC tasks.
- Choose topics of a particular type, and define the task for the searchers accordingly.
- Ask the assessors to make different kinds of evaluation,
- and define appropriate measures.

Experimental design considerations:

- Main effects: system, topic, searcher.
- Pairwise interaction effects: system/topic, system/searcher, topic/searcher.
- Three-way interaction effect.

Some issues and problems

- Setting appropriate tasks.
- Assessment
  e.g. single documents or entire output?
- Evaluation measures.
- Experimental design.
- Outcome *versus* process.