REPORT ON A DESIGN STUDY FOR THE 'IDEAL' INFORMATION RETRIEVAL TEST COLLECTION

K. Sparck Jones
R. G. Bates

Computer Laboratory
University of Cambridge
Corn Exchange Street
Cambridge

October 1977

Summary

This Report presents the findings of a Design Study for the 'Ideal' Information Retrieval Test Collection.

Part A, Design, covers a detailed collection specification, an investigation of sources of document material and methods of obtaining requests and relevance assessments, and estimates of the costs of building various versions of the collection. The conclusion is that a collection consisting of a main set of 30,000 scientific documents and 750 requests with adequate relevance assessments, plus a supporting set of 3000 social science documents with 250 requests and assessments, providing a range of characterisations for the documents and requests but not citation data, could be provided in a two year building programme for about £85K. A similar collection with citation data for the documents could cost £94K. One with more supporting document sets and their own requests and assessments, but without citations, could cost £109K, and a collection with this range of sets and citations could cost £123K.

In Part B, Uses, information on possible uses of the 'ideal' Collection collected as part of the Design Study work is summarised, to allow some evaluation of the proposed collection as a tool for worthwhile research and teaching.

The collection design and costings were considered by the Study Project's Advisory Panel, and in Part C, Discussion, the main comments made by the Panel are noted. Consequent changes to the initial design and costings are indicated, and a possible cut-price collection costing perhaps £55K is presented. Professor Cleverdon has questioned the need for a specially-built test collection, and more specifically argues that sufficiently useful test material may be obtained as a byproduct of operational system investigations. We therefore examine the relationship between a purpose-built 'ideal' collection and an 'incidental' one, in the context of the types of information retrieval research which may by desirable; and conclude that unless very severe restrictions are placed on the kind of research done, an 'ideal' collection, if only a cut-price one, is needed for effective and efficient research.

Acknowledgement

The Design Study was carried out under British Library Research and Development Department Grant SI/G/231. We are very grateful to the many people who provided information for the Study.

Contents

PART A : DESIGN

PART B : USES

PART C : DISCUSSION

References

Appendices