SUMMARY

The present report is the ninth in a series covering research conducted at the Computation Laboratory of Harvard University in automatic information storage and retrieval. An initial version of the fully-automatic SMART document retrieval system was described in Information Storage and Retrieval, Report No. 7, dated June 1964. Search results and retrieval evaluations obtained with this running system were included in ISR-8, dated December 1964. The present report, ISR-9, is divided into two main parts: a new "extended" SMART system is described in Sec. I through XVII, and search results and evaluation procedures developed in conjunction with the original system are reported in the remaining Sections XVIII through XXIV.

The organization of the original SMART system reflects two principal purposes: early computer implementation during the design stages, and fast dictionary look-up and search algorithms. Because of the emphasis on speed, many of the original programs make use of internal (core) processing only, and eschew external (tape) procedures as far as possible. The internal processing is responsible in large part for a variety of size restrictions which limit the application of the system to relatively small, experimental document collections (500 documents, 1000 words per document, 1000 thesaurus classes). In order to ensure the applicability of the SMART

procedures and evaluation techniques to more realistic situations, a decision was made to reprogram the system and thereby to remove the original limitations.

The extended system described in the present report can process document collections of over 250,000 documents, and admit over 250,000 different concept classes as document identifiers. The permissible length of each document is effectively unrestricted, except for the limitation on the total number of concepts which can be produced. The size of the various dictionaries used as intellectual aids is also effectively unrestricted. Theoretically, the extended SMART system should thus be applicable to almost all search conditions which can be expected to be of interest in the foreseeable future; in practice, the system may not of course be operationally useful in every situation because of the inherent time limitations.

A short progress report of the SMART project is presented in Section I of this report by G. Salton. Also included in this section is a list of possible extensions of the system and future tasks which are presently under consideration. Sections II and III by Tom Evslin contain, respectively, a general, relatively nontechnical, description of the SMART system, and a more technical, detailed discussion of the extended SMART system. Section IV by Michael Lesk includes a detailed listing of all input specifications which are applicable to the new system.

The actual programming details of the system are described in Sections V through XVII of the report. Sections V and VI by Mark Cane contain the efficient dictionary (thesaurus) look-up procedures incorporated into the system, and the thesaurus set-up and updating procedures. Section VII by George Shapiro covers the statistical phrase process which makes it possible to group individual concepts assigned to documents into larger units. Arthur Priver and Michael Lesk describe the statistical clustering procedures in Section VIII; the clustering techniques are used to form concept, as well as document, groups, based on statistical co-occurrence criteria.

Analysis procedures using syntactic relations between terms are covered in Sections IX through XII. The formats and updating procedures of the syntactic "criterion" tree file are described in Section IX by Guy Hochgesang. The main syntactic procedures, including the methods used for the recognition of syntactic phrases are examined in Section X by Alan Lemmon. In Section XI, James Prowse covers the procedures used for the analysis of incomplete English sentences, such as titles of documents. Finally, Section XII by Michael Razar describes the tree matching algorithm used to compare syntactically analyzed English text with the pre-stored dictionary of "criterion" trees.

The methods used to correlate analyzed documents and search requests are discussed in Section XIII by Tom Evslin and Guy Hochgesang. Section XIV by Mark Cane and George Shapiro covers the term-term and document-document correlations used as a part of many of the statistical association procedures. The construction and utilization of the concept hierarchy, including the list processing methods available to perform all hierarchical expansion, are

described in Section XV by Michael Razar and George Shapiro. Output formats are similarly described in Section XVI by Guy Hochgesang, and the automatic evaluation routines which terminate a SMART run are detailed in Section XVII by Michael Lesk.

The remaining sections of the report describe evaluation methods and results obtained during the first half of 1965. The programs used to evaluate the iterative search process incorporated into SMART are covered in Section XVIII by Michael Lesk. Section XIX by Joseph Rocchio examines the results obtain by performing a number of iterative searches, using the so-called "merged" methods. In the fall of 1964, a number of students registered in a graduate course at Harvard were asked to submit search request to the SMART system without knowledge of systems operations; the results from this experiment are described in Section XX by Claudine Harris.

A number of special problems concerning retrieval systems evaluation are covered by Joseph Rocchio in Section XXI. Sylvia Sillers in Section XXII reports on work performed to determine a reasonable cut-off procedure which could distinguish retrieved from nonretrieved documents. Finally, a variety of iterative, user-controlled search procedures are described in Section XXIII by Joseph Rocchio, and in Section XXIV by Joseph Rocchio and Gerard Salton. Preliminary results indicate that these new iterative search methods are generally more powerful than simple one-shot procedures. Additional experminents are planned in this area.