OPTICAL CHARACTER RECOGNITION

1. Optical character recognition machines convert printed information into computer-readable codes. The two most common forms of character recognition scanners are photocell arrays and flying spots. Standard OCR fonts are OCR-A developed by National Standards Institute with emphasis on machine reading performance and OCR-B developed by European Computer Manufacturers Association with emphasis on conventional appearance. However, the advance of machines which can not only read a variety of common machine fonts but also handprinted numerals and some special symbols are responsible for the rapid expansion of OCR.

2. The scope of applicability for OCR is currently limited to a particular variety of fonts, to fixed document formats. Poor performance on handwritten documents, the lack of standardization within the industry, problems with recognizing noisy and degraded characters and the limitation imposed on their speed (as well as increased cost) by associated paper-handling devices are problems which are currently being researched by the industry.

3. Presently available equipment ranges in cost from $11,000 purchase for a bare, hand-fed page reader to $15,000 monthly rental for more sophisticated readers and associated equipment.

4. The speed of reading is presently as high as 2,000 characters per second or approximately 1 page per second, limited by the speed of page-feeding. Error rates vary from 1 to 10%, depending on source document control.

5. An initial attempt to solve the problem of noisy, degraded and otherwise poorly readable characters has been to introduce mixed systems. This means, for example, an integration of OCR and key-to-disks where an operator, sitting at a console, can immediately verify unreadable letters or numbers if the OCR device detects them.

6. MOCR, Microfilm Optical Character Recognition, is a new technique designed, in part, to get around the paper-handling problem. It also claims to be able to deal with an unlimited number of character fonts. It uses a flying spot scanner together with a microfilm transport unit. The disadvantage is its cost, approximately $1 million.

7. The current state of the art in OCR still leaves much to be desired other than for routine, massive operations involving large volumes of simple, format-controlled information which typically utilize special purpose, standardized fonts and character sets. All successful operations to date exercise strict control over the preparation of documents to be scanned. Well structured departmental material such as memoranda to the Minister and Cabinet submissions would be of sufficiently standard format as to be amenable to OCR. Much greater formatting discipline would be nesessary before other departmental material would be suitable.

8. Within the next ten to fifteen years it is quite certain that OCR will compete with data encoding as the primary method of converting material to machine-readable form.