



IDRC-TS3

Optical Character Recognition

**Use of OCR Techniques
in Decentralized Data
Collection for
Bibliographic
Information Systems**

H. W. Groenewegen and J. Marshall

International Atomic Energy Agency, Vienna, Austria

Optical Character Recognition

Use of OCR Techniques in Decentralized Data Collection for Bibliographic Information Systems

H. W. Groenewegen and J. Marshall

International Atomic Energy Agency, Vienna, Austria

This work, conducted by the International Atomic Energy Agency, was carried out under a contract with the International Development Research Centre, Ottawa, Canada. The views expressed are those of the authors and do not necessarily represent the views of the Centre.

*The IDRC Technical Studies series consists of papers designed
for rapid dissemination among a specialized readership*

© 1976
International Development Research Centre

Postal Address: Box 8500
Ottawa, Canada
K1G 3H9
Head Office: 60 Queen Street, Ottawa

ISBN: 0-88936-097-9
UDC: 681.327.1
Microfiche Edition \$1

Contents

Foreword	5
Introduction	7
Chap. 1. Preparation for the Experiment	9
Choice of equipment	9
OCR readers	9
The typewriters	10
Fonts	10
Familiarization	11
Chap. 2. Stage I of the Experiment: In-House Tests . .	13
Preparation of initial typescript input; testing the instructions	13
Testing the equipment	14
Computer programing	14
Summary of the results of Stage I	15
Chap. 3. Progress of the Experiment from This Point On .	16
Chap. 4. Preparation of AGRIS Input in a Form Suitable for OCR Processing on a Decentralized Basis . .	18
General	18
The participants	18
The materials	19
Processing of test data	19
The results	19
The Indian Centre	19
Costa Rica Centre	20
Compuscan with Perry 199	20
Autoreader with OCR-B	24
Philippines Centre	25
Compuscan with Perry 199	25
Autoreader with OCR-B	28
Summary and conclusions on Stage II of the experiment	29
Chap. 5. Production of AGRIS Input in a Form Suitable for OCR Processing: Volume Tests	31
Introduction	31
Results	31
Detailed comment	32
OCR-B	32
Courier 12	33

Acceptability of OCR input	33
Summary	35
Chap. 6. Processing of Experimental INIS Input (Abstracts) in a Form Suitable for OCR Processing on a Decentralized Basis by National INIS Centres	
	37
Introduction	37
Results	37
Danish INIS Centre	37
Netherlands INIS Centre	38
Israeli INIS Centre	38
Romanian INIS Centre	39
Supplementary remarks	39
Summary and conclusions	39
Supplementary note	40
Chap. 7. Summary Review of the Results of the Experiment	
	41
Introduction	41
Summary of results	41
Operator instructions	41
The typists	42
The equipment	42
Effects of climate, handling, etc.	43
Choice of OCR equipment	44
Final comments	45
Appendices	
I. Proposal for the conduct of an experiment in optical character recognition for the processing of AGRIS input	47
II. Machine specifications	53
III. IBM Selectric Typewriters	55
IV. Fonts and coding systems	57
V. Letter sent to AGRIS centres participating in the OCR experiments	60
VI. Specifications for an OCR machine suitable for bibliographic data processing (INIS and AGRIS).	93
VII. Integration of OCR techniques into INIS and AGRIS processing cycle	95

FOREWORD

The collection, selection, recording, and indexing of material for large information systems can either be done in one place (centralized), or can be shared among a network of participating centres (decentralized). Two major decentralized systems are now in operation:

- INIS, the International Nuclear Information System, managed by the International Atomic Energy Agency (IAEA)
- AGRIS, the International Information System for Agricultural Sciences and Technology, managed by the Food and Agriculture Organization of the United Nations (FAO).

The formula of decentralization is one under which the governments of individual countries undertake to report the relevant new information issued in their respective territories. In return each receives the complete data base that is compiled by merging the inputs from all participating countries. The formula is one that has also been proposed for new information systems in the future, such as DEVSIS (Development Sciences Information System).

All these systems, however, can handle the large volume of relevant material only if the processing is carried out by modern computer techniques. The capture of the relevant material in computer-readable form is an operation that requires skills and equipment. Because many participating centres, particularly in developing countries, do not possess adequate facilities for converting their inputs to computer-readable form, the operators of INIS and AGRIS have accepted inputs typed on "worksheets" and have converted these to magnetic tape or other computer-compatible media at their central processing facilities in Europe. But this is costly for the international organizations, and it hampers the development of the full benefits of decentralization.

In this report, the authors describe experiments that successfully demonstrated a technique for automatic reading of typed worksheets. At IDRC we believe that this demonstration has important implications for the international systems. But,

more important, we believe that any technical achievement that facilitates a developing country's participation in such systems contributes eventually to the development of the country itself. For it helps the country to organize the information that has been produced within its own territory, to exchange this with other developing countries, and indeed to secure access to a comprehensive file of similar information generated throughout the world. The resources of the international agencies can be tapped to assist a developing country to acquire the necessary skills to exploit these benefits; the more the merely technical problems can be minimized, the more these resources can be applied to immediately beneficial activities.

IDRC is happy to have been associated with IAEA in this endeavour, and I offer my congratulations to the Agency's Division of Scientific and Technical Information (and particularly to Messrs Groenewegen and Marshall) for making yet another contribution to the state of the art of international information systems.

John E. Woolston
*Director, Information Sciences
International Development
Research Centre*

INTRODUCTION

In 1973 the International Atomic Energy Agency (IAEA) commenced some experimental work with optical character recognition (OCR). The reason for its interest was the proposal that machine-readable abstracts would be included amongst the data collected and disseminated by the International Nuclear Information System (INIS). It was recognized that the inclusion of machine-readable abstracts would approximately double the amount of data to be processed, both in the national INIS centres and at the IAEA headquarters. Many INIS centres are not equipped to handle this increased input preparation load. OCR has been used successfully in natural language data processing (e.g., in processing magazine and newspaper texts for computerized photocomposition) and requires relatively simple and inexpensive data preparation equipment (i.e., IBM electric typewriters). It therefore seemed to be a potentially useful means of obtaining the increased input from the INIS centres, particularly those in developing countries, without a commensurate increase in the data conversion load at IAEA headquarters.

Accordingly tests of OCR techniques were conducted by the IAEA during 1973 with input prepared in Vienna. The results were encouraging. They showed that these techniques can be used effectively to prepare INIS abstracts in machine-readable form, at relatively low cost. At this time specific attention was paid to the problems of encoding the rather complicated mathematical and scientific formulae that frequently appear in the abstracts of documents within the INIS subject scope. The experiment did not pursue to any extent the particular problems that might occur when the OCR input is prepared on a decentralized basis and mailed to Vienna for processing.

Conceivably the following problems could occur if this procedure were adopted:

- (i) Difficulties in understanding the instructions, particularly in centres in which the national language is not English;
- (ii) Difficulties in procuring the required equipment and materials;
- (iii) Difficulties in maintaining the equipment to ensure continued quality in the output;

- (iv) Problems of quality control under conditions that vary from centre to centre (e.g., climatic variations, variations in experience in data preparation for machine processing, variations in working conditions, variations in quality of staff available, etc.);
- (v) Problems caused by careless handling of input sheets, including damage sustained in the mail.

It is true that a first step was made toward investigating these problems. The IAEA invited the INIS Liaison Officers to prepare some test input for OCR processing. A set of rudimentary instructions was dispatched to them in August 1974, but the response was very poor; only two centres submitted any test input.

In December 1974 a demonstration of OCR equipment was arranged for the members of the Second Advisory Committee for INIS, which was meeting in Vienna at that time. One of the members of the Committee, John Woolston, Director of the Information Sciences Division of the International Development Research Centre, Ottawa, Canada, expressed considerable interest in OCR techniques, particularly as a method of preparing input for AGRIS and DEVSIS, two international information systems in which IDRC is closely interested. Subsequently IDRC offered the IAEA a contract to conduct an experiment with OCR to establish the possibility of using this technique for decentralized information systems, particularly those in which a major proportion of the input is prepared in developing countries.

In view of the fact that this was an aspect of the experimental work that still needed to be carried out also for INIS, the IAEA accepted the offer and produced a proposal for the conduct of an experiment (Appendix I). The results of the experiment are contained in this report. The authors acknowledge the assistance they have received in their work from a number of people and organizations, particularly: F. Ettenauer and A. Harle (Crosfield Electronics, Vienna and London); J. Hödl (OCR Datenverarbeitung, Vienna); P. Clague and P. Simmons (INSPEC, United Kingdom); A. Pijpers and H. Vissers (ECRM, Netherlands); H. Priegl (Firma Rode, Vienna); Dr W. Richter and H. Frömmel (IBM Austria); A. Chepkasov and L. Warner (IAEA, Vienna); H. Dierickx and H. Schmid (FAO-AGRIS Input Unit, Vienna). All these gave generously of their experience and some made valuable equipment available.

CHAPTER 1

PREPARATION FOR THE EXPERIMENT

Choice of Equipment

OCR readers

A large range of optical character readers is now available on the market. However, it was decided to confine the experiment to two makes of equipment only, namely the ECRM Model 5200 Autoreader (hereinafter referred to as the "Autoreader") and the Compuscan 170 (hereinafter referred to as the "Compuscan"). Detailed specifications of these two machines are found in Appendix II. The choice of the equipment was influenced by the fact that at the time the experiment was begun, they were the only two OCR machines available in Vienna that were specially designed for language data processing. As such they permit:

- (i) input of free format data;
- (ii) use of an extensive character set, including upper and lower case Latin alphabets, with diacritical marks;
- (iii) considerable possibilities for editing and correcting of data, both prior to and during the processing of the input.

The two machines are representative of the kind of medium-priced OCR equipment now being manufactured and have found widespread acceptance in the USA and Western Europe.

The possibility of using OCR equipment manufactured by IBM was also contemplated, as this would have been accessible to the IAEA. However, IBM does not as yet market OCR equipment suitable for the INIS/AGRIS application. For example, neither the IBM 1288 nor the IBM 3886 offers a facility for processing lower case characters in the Latin alphabet. Both accept only a limited range of special characters. On-line editing and correction is not possible. The equipment is basically designed

for commercial data processing operations, such as processing of invoices, salary records, stock records, etc. This it does very quickly and reading speeds of up to 1000 characters/second can be achieved. But because of the other limitations mentioned, IBM OCR equipment was excluded from the experiment.

Arrangements were made with the representatives of the manufacturers of Compuscan to conduct a series of preliminary tests on a demonstration machine installed in Vienna. When it became clear that the machine would be available for a limited period only, a contract was let to INSPEC in the United Kingdom to process further test material on its Compuscan. A contract was also made with a private firm in Vienna that owned an Auto-reader to permit the IAEA to process some input through its machine. Subsequently arrangements were made with ECRM, the Netherlands, for the installation of an Autoreader loan machine in the IAEA headquarters in Vienna. This machine was provided free of charge, other than transportation costs. In the final event this Autoreader remained on the IAEA premises for approximately 8 months so that a unique opportunity was available to test this equipment thoroughly.

The typewriters

The manufacturers of both the Compuscan and the Autoreader recommend the use of IBM Selectric typewriters for the preparation of input. In discussions, representatives of Compuscan indicated that certain other makes of electric typewriters had been used successfully for input preparation for their equipment. On the other hand, ECRM, the makers of the Autoreader, stated quite firmly that its equipment was designed specifically for the processing of manuscripts prepared on IBM Selectric typewriters and that they therefore could not make any guarantees about obtaining satisfactory results if other makes of typewriters were to be used for input preparation. In view of these statements and as no other typewriters with the required font were available in any case, only IBM Selectric typewriters were used for the experiment. A Selectric II model was made available for loan to the IAEA by IBM Austria, for the duration of the experimental period. Subsequently a number of IBM Selectric I and II typewriters were purchased by the IAEA and by the AGRIS Input Unit in Vienna,¹ so that, in all, a number of different machines were tested.

Fonts

Both makes of OCR equipment tested are capable of

¹The IBM Selectric typewriters are marketed under a variety of model designations. See Appendix III for more details.

processing typescripts prepared in a number of different fonts, although neither can process fonts interchangeably. In other words, in the case of each machine, a new recognition program must be loaded if typescripts prepared with a different font are to be read.

Each machine was initially tested with two fonts. In each instance one of the fonts selected was the one most commonly used with that make of machine, i.e., the font with which the manufacturers could be expected to have the greatest amount of experience and for which they had presumably developed the most effective recognition program. In the case of the Compuscan this was Perry 199. For the Autoreader it was Courier 12. In addition each machine was tested for its ability to read OCR-B. OCR-B is² now an international standard for optical character recognition. With the Compuscan only the German version of OCR-B was tested; with the Autoreader the U.K. version was tested.

None of the fonts tested provided the full AGRIS character set. For example, the < (less than) and > (greater than) signs, as well as the opening and closing square brackets were missing from the OCR-B and Perry 199 typewriter elements tested. The < and > signs were also not available on the Courier 12.³ However, for both the Compuscan and the Autoreader machines it is possible to code the "missing" characters on the typescript input by the use of combinations of available characters, or by the use of alternative characters (see also Appendix IV).

For INIS, the IAEA was particularly interested in the possibility of processing the Cyrillic alphabet by OCR. Enquiries have revealed that a Compuscan machine has been installed in the Soviet Union on which text in Cyrillic is processed experimentally. Autoreader is not available with a Cyrillic recognition program as yet, but the IAEA has now commissioned ECRM to develop such a program.

Familiarization

A period of familiarization was required to give the staff involved in the experiment an opportunity to learn the basic requirements for preparing input for the equipment.

²European Computer Manufacturers' Association. Standard ECMA-11 for the alphanumerical character set OCR-B for optical character recognition. Third ed. Geneva, ECMA, 1975.

³It should be noted that ECMA-11 standard OCR-B consists of 121 characters. These include all the AGRIS characters, so that it should be possible to commission an OCR-B typing element containing all AGRIS characters.

During this period a small amount of test material was prepared and processed. The results were carefully analyzed and this helped to clear away many basic misconceptions and misunderstandings about the use of the equipment. As a result of the original familiarization tests, it was possible to prepare a set of preliminary draft instructions for the typist preparing OCR input for Stage I of the experiment.

CHAPTER 2

STAGE I OF THE EXPERIMENT: IN-HOUSE TESTS

Preparation of Initial Typescript Input; Testing the Instructions

Following the completion of basic orientation with both Compuscan and Autoreader equipment and the preparation of an initial set of instructions, a typist was engaged to prepare test input.

The purpose of further in-house testing was:

- (a) to test the completeness and comprehensibility of the instructions;
- (b) to provide test data for the testing of the computer programs designed to process the OCR output (paper tape) and to dump the contents of the paper tape for visual comparison checking;
- (c) to determine which fonts, ribbon, and quality paper gave the best results in Vienna, the intention being that only those materials that gave the best results in Vienna would be used in Stage II of the experiment.

The typist preparing the test input was especially recruited for the job. She had no previous experience in documentation work, a limited command of the English language (the language in which the instructions were written), and average typing ability. The reason for selecting this typist, rather than using one of the typists already employed in the INIS section of the IAEA, was that it was felt that she might be more representative of the type of person available in input centres in various countries. In the event it turned out that the choice was a wise one, insofar as she needed to ask a considerable number of questions and to obtain much more detailed information than had been provided in the basic instructions. As a result, areas in which additional explanations should be included in the draft instructions could be identified and the instructions could be expanded accordingly.

The training period required before the typist was ready

to work reasonably independently was approximately 2 weeks. Later experience with typists with a better command of English indicates that this training period was well above average. In fact, most typists learned how to prepare input for OCR in less than 1 week.

Testing the Equipment

Typescripts were prepared for processing on the Compuscan using both Perry 199 and OCR-B (German) fonts and for the Autoreader, using Courier 12 font. The Autoreader available for testing at this time was not equipped with the software required to process OCR-B.

Encouraging results were obtained in the Compuscan tests. The OCR-B recognition error rate was less than 1 error in 2000 characters read. The recognition was even better with Perry 199: 1 recognition error in every 5600 characters read. The main error occurring, with both fonts, was the reading of a lower case letter as if it were upper case; this occurred occasionally with the "w", the "v", and the "i". In addition the "4" was sometimes not read.

By contrast the Autoreader results were less favourable, with an error rate of 1 in 150 characters. Many of the errors included misreading of punctuation marks, particularly the full stop; the small "m" was frequently read as an "n" and the closing square bracket "]" was often read as an "1". It is not clear why these errors occurred so frequently. The typescripts seemed to be reasonably good. It may have been a function of the particular machine that was used and its condition at the time of the tests. Subsequently the IAEA obtained an Autoreader on loan and this made it possible to repeat the Courier 12 tests and to conduct tests with OCR-B. The results with both fonts were much better than they had been with the machine used in the earlier tests. For example the error rate for Courier 12 decreased to 1 error in approximately 2700 characters read; the error rate for OCR-B was approximately 1 error in every 2400 characters read. (See also Chap. 6, "Results," page 37.)

Meanwhile, however, it was decided to conduct Stage II tests for the Autoreader with OCR-B, as Courier 12 results up to this stage had been so poor. For the Stage II Compuscan tests the Perry 199 font was chosen.

Following the completion of these tests further amendments and improvements were made to the draft instructions in an effort to clarify points that had led to problems during the processing of test material.

Computer Programming

During this phase the computer programs were developed

to process the paper tape output from the OCR machines. These programs were mainly designed to convert the paper tape codes produced by the OCR devices under control⁴ of their dictionary programs into standard IBM EBCDIC codes.

The programs also formatted the data so that it could be dumped on the IBM 1403 line printer for visual comparison checking.

At a later stage the programs were developed further to format the OCR input for processing through the AGRIS and INIS input processing programs and error check routines.

Summary of the Results of Stage I

At the end of Stage I the following results had been achieved:

- (i) a set of detailed instructions for the preparation of AGRIS input in a form suitable for OCR processing had been prepared and tested for comprehensiveness and comprehensibility;
- (ii) tests had been conducted with the Compuscan and the Autoreader and the most effective⁵ font had been determined for each machine;
- (iii) computer programs had been written to process the output produced by the OCR equipment and to dump it in the form of computer printouts for comparison checking.

⁴The dictionary programs determine what code configuration is punched by the OCR device for each typewritten character that it reads. As the companies processing the test data were not able to amend these programs for the IAEA's use, it was necessary to accept nonstandard coding configurations in the early stages of the experiment. Accordingly the IAEA computer had to be programed to convert these nonstandard codes to IBM EBCDIC codes. It was only when the IAEA obtained an Autoreader loan machine for its exclusive use that dictionary programs could be changed to produce immediately EBCDIC code configurations as output from the paper tape punch. At that time, too, the dictionary programs for the Autoreader could be amended to accept coded representations of the more esoteric characters, not available on the typewriter elements, particularly special INIS characters. (See also Appendix IV)

⁵At this stage the Compuscan had performed markedly better than the Autoreader and the choice of font for the latter became basically a matter of avoiding the use of Courier 12.

CHAPTER 3

PROGRESS OF THE EXPERIMENT FROM THIS POINT ON

Up to the end of Stage I the experiment had proceeded according to schedule. However, now a certain number of new factors were added to the original design of the experiment. Although this resulted in a significant delay in the completion of the experiment, it is considered that those additional factors enhanced its value considerably.

They altered the character of the experiment from being a rather limited preliminary feasibility study to that of an open-ended trial that, by the end of 1975, led to the introduction of OCR processing of both AGRIS and INIS input on a production basis.

The following components of this new test situation can be identified:

- (i) Production of AGRIS input in a form suitable for OCR processing on a decentralized basis in three developing countries (Costa Rica, India, the Philippines). This was the original Stage II of the experiment and is discussed in Chap. 4 of this report.
- (ii) Production of AGRIS input in a form suitable for OCR processing on a production basis at the IAEA. In this case all input prepared was "live" data and was finally included in the AGRIS data base (discussed in Chap. 5).
- (iii) Production of experimental INIS input (abstracts) in a form suitable for OCR processing on a decentralized basis by national INIS centres, (discussed in Chap. 6 of this report).

These three components were conducted more or less simultaneously. Subsequently two further steps were taken:

- (iv) Preparation of specifications for an OCR machine to be purchased by the IAEA in 1975 (see Appendix VI).
- (v) Incorporation of OCR processing into the AGRIS and INIS input streams (see Appendix VII).

CHAPTER 4

PREPARATION OF AGRIS INPUT IN A FORM SUITABLE FOR OCR PROCESSING ON A DECENTRALIZED BASIS

General

The participants

On the advice of the officer-in-charge, AGRIS Input Unit, four AGRIS centres were selected for this part of the experiment and were approached with the request to participate. The centres were:

Centro Interamericano de Documentacion e Informacion
Agricola
IICA-CIDIA
Turrialba, Costa Rica

The Indian Council of Agricultural Research
Research Project Unit
Krishi Bhavan
Dr. Rajendra Prasad Road
New Delhi, India

Agricultural Information Bank of Asia
South East Asian Regional Centre for
Graduate Study and Research in Agriculture
College
Laguna, Philippines

Centre de Documentation pour le Programme de
Developpement du Bassin du Fleuve Senegal
B.P. 383
Saint Louis, Senegal

Each centre was asked whether it had access to an IBM Selectric typewriter model I or II capable of typing 10 characters to the inch. The centres were also told that they would need to

have available a typist with some knowledge of English who could work on the project for approximately 5 days.

Three of the centres replied that they would like to participate. Only the centre in Senegal was unable to take part in the experiment.

The materials

The three participating centres were each sent a complete OCR kit, consisting of the following:

- (a) two IBM typewriter elements: OCR-B (English) and Perry 199;
- (b) two sets of 100 sheets each of input paper;
- (c) two sets of instructions for the preparation of input for OCR: the first set was labeled "Compuscan," the second "ECRM" (i.e., Autoreader);
- (d) one set of 50 completed AGRIS worksheets (test data);
- (e) one IBM carbon film ribbon

A copy of the letter, under which cover the instructions were sent, and a copy of the actual instructions (Autoreader only) are attached as Appendix V.

Processing of test data

The procedure followed in the processing of the data and in the checking of the results was as follows.

All input was processed through the relevant OCR equipment. The output was then processed by the IAEA's computer and dumped on printouts. These were visually compared with the input sheets and the recognition errors tallied.

Processing of the input through the AGRIS checking programs would not have been adequate. In the first place these programs cannot identify errors in free text data, such as author's names, titles, etc. In the second place they would flag errors in formatted data that might not be due to recognition failures, but simply to typing errors.

The Results

The Indian Centre

Unfortunately the input received from the Indian

Council of Agricultural Research could not be processed because the correct typewriter was not available to them after all. A 12-pitch model IBM typewriter was used instead. This typewriter could not accept the once-only carbon ribbon that is required. Thus two basic requirements for OCR input preparation were not met (Fig. 1). The centre sent some samples of locally produced paper available to them. This, too, would cause problems. In particular, numerous small specks were noted that are caused by unbleached or partially bleached particles of wood pulp remaining in the paper. It is clear from the above that if this centre were to prepare AGRIS input in a form suitable for OCR processing it should be supplied with a suitable typewriter and acceptable paper.

On the positive side it may be said that the quality of the typing was good and that the typist had understood the instructions completely. The input sheets did not appear to have suffered from being sent through the mail and would have caused no physical problems in processing.

Costa Rica Centre

A full batch of test data was received from this centre. The results are as follows.

Compuscan with Perry 199

Results: Recognition errors on IAEA-supplied paper: 1 in 70 characters. Recognition errors on locally supplied paper: 1 in 130 characters.

Detailed comments: On first inspection these results appear to be very poor indeed. However, the actual characters causing recognition problems were few. The main problem was caused by the deletion character. The typist had made many errors and corrected them, in accordance with the instructions, with the deletion symbol. For Perry 199 on Compuscan this is the open box (☐).

Unfortunately the typewriter that was used required some adjustment on precisely that key and the symbol printed very poorly (Fig. 2). On 50 worksheets the deletion symbol was used on 217 occasions. However, it was sufficiently clearly typed to be recognized as such by the Compuscan on only 15 occasions. Other characters that printed poorly, and for that reason were frequently not recognized, were the slash (/), which was read as an apostrophe 22 times on the 50 worksheets analyzed, and the "4", which, when poorly typed, was not read at all. This

INSTRUCTIONS (NOTES) ← ALIGN FIRST CHARACTER UNDER THIS ARROW

1	001!IN1190028
2	002!1/2
3	004!N
4	008!FOO:L30/B/AM/KEV
	009!A
5	100!Hrishi, N. (Central Tuber Crops Research Inst., Trivandrum (India)
6	200!Problems and prospects in cassava production in India
7	210!Interdisciplinary Workshop on Cassava Processing and Storage
8	211!Pattaya (Thailand)
	213!17 Apr 1974
9	300!IDRC--031e
10	600!(En)
11	610!Also in microfiche. Summary (En, Fr). *IDRC, Ottawa (Canada)
12	620!1540!9350/6635
13	002!2/2
	009!M
14	100!Araullo, E.V.; Nestel, B.; Campbell, M. (eds.)
15	110!International Development Research Centre, Ottawa (Canada)
16	200!Cassava processing and storage
17	201!Proceedings of an interdisciplinary workshop
18	320!ISBN 0088936036
	401!Ottawa (Canada)
19	402!IDRC
20	403!1974
21	500!p. 59-62
22	*RT

Fig. 1. Input from Indian Centre (Perry 199). Incorrect spacing of characters (12 pitch instead of mandatory 10 pitch); smudging ribbon.

001!IF229))5)#####001!IF2290050
 002!1/2
 004!N
 008!FOO!R10/B/AM/KEV
 009!A
 100!Wijeratne, W.B. (Department for Development of Marketing,
 Colombo (Sri Lanka. Fruit and Vegetable Utilisation Lab.)
 200!Cultivation, processing and utilization of cassava in
 Sri Lanka
 210!Interdisciplinary Workshop on Cassava Processing and
 Storage
 211!Pattaya (Thailand)
 213!17 Apr 1974
 310!IDRC--031e
 600!(En)
 610!Also in microfiche, Summary (En, Fr). *IDRC, Ottawa
 Canada#####(Canada)
 620!1540!9230!9290/6744
 002!2/2
 009!M
 100!Araullo, E.V.; Nestel, B.; Campbell, M. (eds.)
 110!International Development Research Centre, Ottawa
 (Canada)
 201!Proceedings of an interdisciplinary workshop #####
 320!ISBN 0088936036 #####
 401!Ottawa (Canada) #####

Fig. 2. Input from Costa Rica Centre (Perry 199). Poor impression of deletion symbol.

happened 12 times. Finally, the "R" was read as a "P" on seven occasions. These four characters accounted for 91% of the recognition errors. In the tests at the IAEA only the recognition of the "4" has given similar problems to those experienced with the Costa Rica input, indicating that there might be a flaw in the recognition program here. The other three characters had given no trouble in the IAEA tests.

In addition to these recurrent errors, there were a number of nonrecurrent recognition failures. The quality of the typing was only fair; some characters were badly smudged but even so they were frequently still recognized. The number of times the typist had to delete data she had typed gives a fair indication of the overall quality of the typing. Probably more practice would have resulted in improvement in a number of instances, including those that now caused recognition errors. There was some evidence that the typist had misunderstood some details of the correction instructions.

The fact that the error rate was slightly lower on the locally supplied paper means nothing. It simply indicates that the typist made fewer errors in this batch (probably due to the fact that by then she had had more practice) and therefore used the troublesome deletion symbol less frequently.

Although the paper had traveled long distances (in the case of the IAEA paper, from Vienna to Costa Rica to Vienna, and then to the U.K. for processing), no processing problems occurred.

One point became very obvious: the Compuscan will not read typing that occurs outside the defined reading zone. It is very inflexible in this regard. On the locally supplied paper the reading zone was not defined for the typist by preprinted margins. As a result, 17 lines in a total of 50 worksheets were dropped by the Compuscan because they had been typed too close to the top or bottom of the input sheet.

In summary:

- (a) Results would have been much better with a properly adjusted typewriter. At least 91% of the recognition errors occurred with the same four poorly typed characters. In fact, 75% of the total errors were caused by the poor impression made by the deletion character. A simple typewriter adjustment would have produced much better results.
- (b) The quality of the local paper supplied was satisfactory for OCR.
- (c) The Compuscan was reasonably tolerant to smudged characters.
- (d) The Compuscan was very inflexible regarding reading

zones. Paper with a preprinted reading zone is essential when Compuscan is used.

- (e) No trouble in recognition or physical processing of the input could be traced to the handling of the paper at the input centre or in transit.

Conclusion: The results appear to be much poorer than they really are. Simple typewriter adjustment on four keys would have decreased the recognition error rate already to approximately 1 in 1200 characters.

Autoreader with OCR-B

Results: Recognition errors on IAEA-supplied paper: 1 in 6600 characters. Recognition errors on locally supplied paper: 1 in 1300 characters.

Detailed comments: Both the quality of the input and the recognition were excellent. Most common problems occurred with the recognition of the zero (0), which was misread three times on the IAEA-supplied paper. The zero was not recognized at all (i.e., it was missing from the output) seven times on the local paper. The "j" was not read on four occasions when it occurred in input prepared on local paper. Other than this, some nonrecurrent errors were noted with the local paper.

The Autoreader was much more sensitive than the Compuscan to the presence of small dark specks in the paper. Altogether it detected two specks on 50 input sheets typed on IAEA-supplied paper and interpreted these as full stops. On the local paper 10 specks were recognized. The INIS and AGRIS software can cope with extraneous data, such as the occurrence of stray full stops between fields and records. At worst the records need to be updated to eliminate such data.

The typing was very good. In fact, in one batch of 63 input sheets not a single typing error was found. Some problems were experienced by the typist in interpreting instructions, including a major misunderstanding resulting in incorrect encoding of the tag delimiter. This could be programmed around so that the data could still be computer-processed for checking.

The Autoreader is very tolerant about reading data that falls outside the defined "reading zone." No lines were lost in processing, even though some were typed quite high or low on the page. Printed margins would therefore not be essential, provided the typist set the margins on the typewriter.

In summary:

- (a) Results were excellent for OCR-B using the IAEA-supplied paper. Recognition problems with the zero (0) and "j" have been drawn to the attention of the manufacturers who believe the recognition software can be improved for these characters.
- (b) The Autoreader was more sensitive to paper quality than the Compuscan. Results with IAEA-supplied paper were markedly better than with local paper.
- (c) No problems were experienced with typescripts being outside of the reading zone.
- (d) No trouble in recognition or processing of the input could be traced to the handling of the paper at the input centre or in transit.

Conclusion: The results were excellent. This centre should be able to produce OCR input for the Autoreader without any difficulty.

Philippines Centre

A full batch of test data was received from the centre. The results are as follows.

Compuscan with Perry 199

Results: Recognition errors on IAEA-supplied paper: 1 in 2200 characters. Recognition errors on locally supplied paper: 1 in 1050 characters.⁶

Detailed comments: The most common recognition error was the zero (0) being read as a "C". This happened twice in the batch typed on IAEA-supplied paper but 35 times on the batch typed on local paper. This single problem accounted for 69% of the total recognition errors. It was difficult to detect with the naked eye the difference between the zeros that were recognized as such and the zeros that were recognized as a "C". Certainly, the left-hand side of the character was usually more sharply printed than the right-hand side, indicating some flaw in the typewriter adjustment on the key, but, even so, recognition failures occurred

⁶This was the best result for this batch, achieved on approximately half of the input sheets submitted. On the other half the recognition errors were much more frequent because a second, poorly adjusted typewriter had been used to type this second half of the batch (Fig. 3).

001!PH2200453

002!1/1

004!N

008!F25/J/AS/KE

009!A

100!Novoa, F.V.; Nunez, R. (Escuela Nacional de Agricultura, Chapinigo (Mexico). Colegio de Postgraduados)

200!Efficiency of five phosphate fertilizer sources in soils with different phosphate fixing capacities

210!Caribbean and Tropical American Soil Science Conference

211!St. Augustine (Trinidad and Tobago)

213!7 Jan 1973

600!(En)

610!Graphs, tables; 9 ref. Summary (En)

230!Tropical Agriculture (Trinidad and Tobago)

403!(Apr 1974)

500!v. 51(2) p. 235-245

*RT

Fig. 3. Input from Philippine Centre (Perry 199). Bad character alignment. Poor impression of certain characters, e.g., 9, 4, N, M, etc.

erratically. Other recurring problems were failures to recognize the "4" and difficulties in reading the letter "M".

The quality of the typing was generally good, except for one batch of input sheets typed on locally supplied paper. This had been prepared on a badly adjusted typewriter and caused considerable processing difficulties and a high error rate.

The locally supplied paper was foolscap size. This the Compuscan would not accept: the pages had to be cut to A-4 size. The Compuscan also had some trouble with the paper feed for these pages because the left-hand edges of the sheets had become crumpled in transit. Because input is fed into the Compuscan

001!PH3390049	
002!1/2	
004!N	
008!F00;E30;Q10/B/AM/KEV	
009!A	
100!Castillo, L.S. (University of the Philippines, Los Banos, Coll. of Agriculture)	
200!The cassava industry of the Philippines	
210!Interdisciplinary Workshop on Cassava Processing and Storage	
211!Pattaya (Thailand)	
213!17 Apr 1974	
310!IDRC-031e	
600!(En)	
610!Also in microfiche. Tables. 18 ref. Summary (En, Fr).	
*IDRC, Ottawa (Canada)	
620!1540;1580;9350/6732	
002!2/2	
009!M	
100!Araullo, E.V.; Nestel, B.; Campbell, M. (eds.)	
110!International Development Research Centre, Ottawa (Canada)	
200!Cassava processing and storage	
201!Proceedings of an interdisciplinary workshop	
320!ISBN 0083936036	
401!Ottawa (Canada)	
402!IDRC	
403!1974	
500!p. 63-71	

Philippines,



Fig. 4. Input from Philippine Centre (OCR-B). Bad character alignment. Poor impression of certain characters (e.g., 8). "Bounce" on right-hand side of the page (e.g., around the words "Philippines" in Tag 100; "Processing and Storage" in Tag 210, etc.) (see insert).

with the left edge leading, processing of these pages gave some difficulties.

In summary: Some typewriter problems and some paper problems interfered with the successful reading of the batch. Both problems should be correctable. It should be noted that part of the batch that was typed on a properly adjusted typewriter using IAEA-supplied paper gave an acceptable result.

Autoreader with OCR-B

Results: Recognition errors on IAEA-supplied paper: 1 in 120 characters. Recognition errors on locally supplied paper: 1 in 80 characters.

Detailed comments: Recognition on both batches was extremely poor. When the input sheets were inspected, it was immediately obvious that the bulk of these had been typed on a very badly adjusted typewriter. (Probably this was the same typewriter that caused such poor results with part of the material prepared for the Compuscan) (Fig. 4). A few sheets were obviously typed on another typewriter and it is easy to see the difference (Fig. 5). On the typescript prepared on the poorly adjusted typewriter the characters are badly out of adjustment and irregularly spaced. It is obvious that many pages were typed with the wrong ribbon (i.e., a fabric ribbon).

Finally, but fatally for recognition purposes, it is clear that the right-hand side of a number of sheets had not been properly held back against the typewriter platen, perhaps because the metal bar used for this purpose was either broken or not being used. As a result the pages "bounced" each time they were struck by the typewriter element. This caused small marks to be printed all around the characters and these, of course, interfered severely with the recognition.

For all these reasons it is impossible to make any detailed comments about this test input. However, the poor results obtained in this test should not lead to the conclusion that this centre could not prepare good input. The Autoreader input was poor because it was prepared on a typewriter that was in poor condition and partly with the wrong ribbon. When it used a good typewriter the centre had no difficulty in preparing satisfactory input, as can be seen from the results achieved with Compuscan.

001!PH3390040	
002!1/1	
004!N	
008!P10;E50/B/M/ZEV	
009!M	
100!White, A.; Sevfour, C. (Colorado Univ., Boulder (USA). Inst of Behavioural Science)	
110!International Development Reser@arch Centre, Ottawa (Canada); Assistance Technique Suisse	
200!Rural water supply and sanitation in less-developed countries	
201!A selected annotated bibliography	
310!IDRC--028e	
320!ISBN 0088936033	
401!Ottawa (Canada)	
402!International Development Research Centre	
403!1974	
500!81 p.	
600!(En)	
610!Also in microfiche. Summary (En, Fr). *IDRC, Ottawa (Canada)	
620!/GZ63	
*RT	

Fig. 5. Input from Philippine Centre (OCR-B). Compare with Fig. 4. Note correct alignment of characters; clear impression of all characters; no "bounce" effect.

Summary and Conclusions on Stage II of the Experiment

(a) Three centres submitted input but the input of only two of the centres was processible; the input prepared by the third centre could not be processed because of failure to use the correct equipment and materials.

(b) The results of the test were quite varied. One centre produced very good input for the Autoreader, but poor input for the Compuscan; the other centre produced acceptable input for the Autoreader. The variations in quality were easily traced to the degree of adequacy of the equipment used.

(c) None of the centres seem to have had major difficulties in following the instructions. Some minor misunderstandings did occur but they would be relatively easy to clear up.

(d) Two of the centres managed to obtain local paper that could satisfactorily be used for OCR. Given that a reasonable amount of care in packing of the input is exercised, there should be no difficulty in processing material that has been sent through the mail, even for long distances.

(e) There was no evidence that climatic or other local conditions under which the input was prepared interfered with the recognition.

(f) The tests demonstrate clearly the basic problem that could arise with decentralized data preparation for OCR, namely maintenance of the equipment. A decision to invite OCR input from any country should take into consideration whether there is available in the country a competent service organization that could maintain the typewriters. It is vital that the typewriters on which the input is prepared be in good condition and properly adjusted. Even typescripts that may at first glance appear to be perfectly acceptable can give difficulties in recognition.

(g) Awareness on the part of inputting centres of the recognition problems that are caused by the use of poor materials will also help to avoid these problems. For example it is clear that the Philippines centre did not recognize the problems that the use of a fabric ribbon would cause.

(h) Despite the poor results obtained in part of the tests, we are convinced that these centres could produce good input for OCR processing and they should be encouraged to do so. If necessary they should be equipped with the materials that they will need.

CHAPTER 5

PRODUCTION OF AGRIS INPUT IN A FORM

SUITABLE FOR OCR PROCESSING: VOLUME TESTS

Introduction

Whilst awaiting receipt of the material prepared by the centres participating in Stage II of the experiment, advantage was taken of the fact that an Autoreader was available at IAEA headquarters to process more OCR input. The main aim was to obtain additional experience; a secondary consideration was to assist the AGRIS Input Unit in processing a backlog of AGRIS worksheets that had accumulated at that time.

During the months of June and July 1975, some 700 OCR input sheets were processed. Most of the items were published in *Agrindex*, volume 1, no. 8 and 9, August and September 1975. Approximately 400 of the OCR worksheets were typed with the OCR-B typing element; the remainder were typed with Courier 12 font.

Subsequently, commencing in November 1975 routine preparation of AGRIS input by means of OCR was commenced. During the months of November and December some 2000 AGRIS input sheets were typed with the OCR-B typing element, and processed for inclusion in *Agrindex*.

Results

In the period June-July: Recognition errors with OCR-B typing element: 1 in 2400 characters. Recognition errors with Courier 12 typing element: 1 in 2700 characters.⁷

⁷ One AGRIS worksheet contains approximately 430 characters. Thus there was an average of 1 recognition error in every 5.6 worksheets for OCR-B, and 1 recognition error in every 6.3 worksheets for Courier 12.

Checking the actual sequences of worksheets in which no

In the period November-December: Recognition errors with OCR-B typing element decreased to less than 1 in 22 000 characters.

Detailed Comment

OCR-B

In the test conducted during June and July the error rate remained fairly consistent over the entire batch of 400 worksheets, typed over a period of 3 weeks. The most common error was the zero being read as the capital "O". This occurred 16 times. Other common errors were the lower case "u" being read as upper case "U" and the lower case "j" not being read at all. These three errors accounted for 43% of the recognition failures. They appear to have been caused by bugs in the OCR-B recognition program, which at that time had only just been released by the manufacturers of the Autoreader. Other errors, which occurred only once or twice, such as the "g" being read as "a" or "o" can be explained by poor character impressions on the typescript or by the presence of dirt on the scanning mirrors.

When full-scale processing of AGRIS input resumed in November, many of these problems had been overcome. Software improvements appear to have eliminated problems with the lower case "u" and the lower case "j". Some difficulties in regard to the recognition of the zero still persist, but misreadings have become much less frequent during the period the Agency started working with new typewriters, and some care was taken to ensure that they were properly adjusted. The accumulation of experience on the part of the operators played a major role in improving the quality of the typing and maintaining the Autoreader in good condition.

Throughout the early part of the test (June and July) two variations of paper were used. One had a smooth surface, the other a rougher surfaced bond paper. Both gave equally good results from the point of recognition, although the smoother paper caused some feed problems with two or more sheets occasionally feeding through simultaneously. Because of this relatively

errors occurred, the longest sequence was 26 worksheets without errors for Courier 12 (an error rate of better than 1 error per 11 000 characters read). For OCR-B the best result was no error in a sequence of 19 worksheets (error rate better than 1 error in 8000 characters read). However, for Courier 12 the average number of worksheets that were read before an error occurred was a little less than 4, whilst for OCR-B the average error-free sequence consisted of a little more than five worksheets.

minor problem it was decided to continue working with the bond paper only.

Courier 12

The results with Courier 12 achieved during June and July were more variable throughout the batch. Although one batch of approximately 130 worksheets gave good results (error rate: 1 in 2700 characters), results with another batch of worksheets were disappointing (error rate: 1 error in 700 characters). The variation was due entirely to the variation in quality of the typed input. The poorly typed batch was prepared on a typewriter with a dirty element. This caused some smudging around the characters, particularly around the bottom of the zero (0), which consequently was read as a "U" 11 times on 50 worksheets. Some other characters also posed consistent problems, the main ones being "8", which was frequently recognized as an "S", and the ")", which was frequently recognized as a "]". Curiously, the "(" was never confused with the "[".

The experience with Courier 12 is that under ideal conditions recognition of that font is as good as, if not better than, OCR-B on the Autoreader. This is to be expected, as the manufacturers of the equipment have had long experience with Courier 12. However, Courier 12 recognition deteriorates quite quickly once small imperfections appear in the typescript. This appears to be due to the fact that Courier 12 is quite a fine font. With OCR-B, which is a much bolder font, comparatively much greater deterioration from first quality in the typescript can be tolerated before the recognition rate is significantly affected. Because of this, Courier 12 was not used in the November-December production runs.

As with the OCR-B, there was no noticeable effect on the recognition produced by the use of different qualities of paper.

Acceptability of OCR Input

The experiment with a large volume of AGRIS input proved that OCR is a feasible method of input preparation for a bibliographic information system. The error rate, already acceptable during the June-July tests, kept decreasing as staff gained familiarity with the requirements for OCR processing.

The accuracy of recognition was finally so good that, compared to the formal errors and the typing errors that are detected in the data after they have passed the input checking

program, the recognition errors can now be called negligible at less than 1 error in 50 worksheets, approximately.

The most serious problems that can occur are recognition errors in the Temporary Reference Number (TRN); less serious, but still annoying, are errors in the tag numbers on the worksheets. Errors in these data normally cause rejection by the computer of all or part of the record. If the input is punched on paper tape (by means of the Flexowriter) or encoded on magnetic tape (by means of a magnetic tape encoder), rejection of input by the computer means that the rejected data must be repunched. With OCR this is normally not necessary. All that needs to be done is to feed the worksheets through the machine again, taking particular care (by monitoring on the display screen) that the characters are correctly recognized this time. The data are then ready for reprocessing at a considerable saving in operator time.

This first experience with OCR processing on a larger scale brought home the considerable advantage in day-to-day operation resulting from this facility of having what amounts to "eye-readable machine-readable form." A further advantage is that each record can be typed on a separate sheet or a series of separate sheets and therefore forms a discrete unit. This, and the eye-legibility of the sheets, makes sorting, reprocessing, correcting, etc., of the input very simple.

It is a matter of judgment as to whether there is advantage in proofreading the worksheets before they are processed. Certainly, proofreading can detect basic errors that can then be corrected by means of the quite sophisticated correction procedures available on both types of OCR equipment tested. However, there is no absolute certainty that what has been typed will be correctly read every time. Thus it seems preferable to subject the data to the complete checking programs first and then to proofread the error lists. This will permit recognition errors to be detected as well as formal errors and typing mistakes.

It was noted that there was no need to type the input on specially ruled paper. Left- and right-hand margins set on the typewriter would be sufficient to keep the typing within the "reading zone." The Autoreader was very tolerant and read characters up to less than 1 centimetre from the right-hand margin.

The possibility of using special AGRIS OCR-sheets preprinted with tag numbers in the chosen font was considered. There seems to be little advantage in doing so; as the typist would waste time

⁸Correction procedures permit insertion and deletion of entire lines as well as single words and characters at any place in the typescript, provided sufficient space has been left between lines.

aligning the data with the appropriate tag, she would find it quicker just to type the tag. In addition, although the computer is programed to ignore tag numbers in the input that are not followed by data, intolerable delays in processing such sheets on the OCR would occur, as the machine must scan along each blank line after having "read" the nonapplicable tag number on that line.

In practice it would seem best for a descriptive cataloguer first to complete a standard AGRIS worksheet and then hand the complete worksheet to a typist who transfers the data on a sheet for OCR processing. Preparation of an OCR input sheet by a descriptive cataloguer working direct from the document seems to be fraught with danger, particularly as he would move back and forth between the lines, completing data fields and correcting data already recorded on the sheet. In this respect OCR cannot be expected to give advantage over the preparation of input by means of more conventional input preparation equipment, such as the Flexowriter and the magnetic tape encoder, which also presupposes that the operator copies data from a completed worksheet.

For bibliographic systems such as INIS and AGRIS, OCR has one disadvantage over the other input devices mentioned. This is that typewriters do not, of course, have a "playback" feature by which recurrent information can be punched automatically. Nor can they be programed to "prompt" the operator.

In an effort to overcome the first disadvantage, experiments were made in which recurrent information for a group of items was typed on a separate sheet of paper and fed through the OCR machine a number of times, interleaved with the nonrecurrent information for each item. This was reasonably successful; however, recognition does deteriorate after a sheet of paper has been fed through the OCR more than some 10 or 15 times and in any case there can be no guarantee that recognition was perfect each time, so that the repeated data for each item must still be proofread. This procedure also introduced some possibility of error on the part of the operator who must feed through the sheets in the correct sequence. Thus this solution was a compromise one only.

Finally it may be said that this stage of the experiment gave useful experience in the programing of the Autoreader, particularly in the amendment of the dictionary program to ensure that the machine punched the required EBCDIC codes. It also provided experience in the integration of OCR input into the AGRIS data processing and checking cycles.

Summary

It is considered that the large-scale experiment in the

middle of 1975 and subsequent experience in late 1975 have proved without doubt the suitability of OCR techniques for the preparation of input to bibliographic information systems. Of the two fonts used, OCR-B showed itself to be more tolerant to imperfection in typing quality and is therefore the preferred font for future work, also on a decentralized basis. OCR-B recognition programs proved themselves to be very reliable in the later part of the tests, despite the fact that they have been developed only recently.

There seemed to be little to choose between two types of paper, as both gave good results.

It was possible to develop routines for integrating OCR input into the AGRIS processing cycle. The kinds of errors that might be expected with OCR data preparation became clearer. Many technical problems were solved and very useful experience was gained. This enabled the IAEA to define specifications for an OCR reader suitable for processing both INIS and AGRIS input (see Appendix VI) and to integrate OCR input preparation into the activities of the AGRIS Input Unit and the INIS Bibliographic Control Unit (see Appendix VII).

CHAPTER 6

PROCESSING OF EXPERIMENTAL INIS INPUT (ABSTRACTS) IN A FORM SUITABLE FOR OCR PROCESSING ON A DECENTRALIZED BASIS BY NATIONAL INIS CENTRES

Introduction

As mentioned in the Introduction, the IAEA had a special interest in OCR, because it might facilitate the provision by certain national INIS centres of INIS abstracts in a machine-readable form.

To test this possibility further, it was decided to invite national INIS centres to submit some abstracts in a form suitable for OCR processing. An INIS Technical Note, containing the necessary instructions, was prepared. The text was based extensively on the text of the instructions prepared for AGRIS centres for Stage II of the experiment. This Technical Note was then sent to all national INIS centres under cover of a letter inviting them to submit some test input. Although none of the centres was provided with any materials for this test, the instructions did of course explain what materials were required.

Test input was received from the INIS centres in Denmark, Israel, the Netherlands, and Romania.

Results

Danish INIS Centre

This centre sent input sheets prepared with the Courier 12 typing element. On the whole, recognition was satisfactory. Some minor problems were experienced with the recognition of the "9", the "i", and the "l", and the symbols "@", and "&". These problems were due to slightly poorer impressions of these symbols on the typescript, probably caused by the fact that the typing element was

still new. It had already been noted in the tests conducted in Vienna that a new typing element seems to need a short period of "wearing in" before all the characters produce a good sharp impression on the typescript. In addition some slight adjustment of the typewriter could have helped.

The paper supplied by the Danish centre was satisfactory and caused no reading problems. The sheets were packed securely between two pieces of cardboard, which effectively prevented them from being damaged in the mail.

Although not enough input was received to prepare meaningful statistics, the test material sent indicated clearly that no problems would be experienced in processing OCR input from this centre.

Netherlands INIS Centre

This centre also prepared input with the Courier 12 typing element. The results were exceptionally good. The recognition rate was better than 1 error in each 10 000 characters read. The centre used paper supplied by the IAEA, identical to that supplied to the AGRIS centres for Autoreader tests. The test sheets were hand-carried to Vienna, i.e., they were not subjected to mail handling. It would, however, be wrong to attribute the excellent results to this fact; even test sheets prepared in Vienna did not give such good results.

A more likely explanation for the excellent results obtained was given by the manufacturers of the Autoreader. As the European headquarters of this company are located in the Netherlands, the equipment is generally tuned to the IBM typewriters that are manufactured in that country. Of course the Dutch INIS centre had access to precisely such typewriters. It is likely that this was the reason for the outstanding result. This does not mean that results as good as these could not be obtained with typewriters manufactured elsewhere. It only proves that for the best results the Autoreader and the typewriters must be tuned to each other. When input is prepared on a decentralized basis, such "tuning" can clearly only be a compromise designed to cope reasonably with many typewriters manufactured in different parts of the world.

Israeli INIS Centre

This centre also prepared input using a Courier 12 typing element, but unfortunately, the standard version was used, not the one specially adapted for the Autoreader. The main differences lie in the shape of the exclamation mark (!), the comma (,) and the zero (0). Therefore the Autoreader failed

to recognize the first two characters entirely, whenever they occurred in the typescript, and had great difficulty with the third. Blurred typing of the "g" and the "s", and to some extent of the "w", caused consistent recognition problems. This blurring was probably due to a dirty typing element being used. The paper supplied by the Israeli centre was satisfactory for recognition purposes, although slightly thin.

Again, not enough input was prepared to calculate meaningful statistics. However, it would seem that the centre would have no difficulty preparing input for OCR processing, provided the correct typewriter element was used and kept in clean condition.

Romanian INIS Centre

As for the Israeli centre, here too an unsuitable version of the Courier 12 typing element was used, so that the exclamation mark (!) was not recognized. Apart from this a major problem was caused by the use of a fabric ribbon instead of the once-only carbon ribbon. As a result much of the input was not recognizable.

Supplementary Remarks

This test once again brought home the essential role played by the materials. Two of the centres did not recognize the vital importance of this and substituted other materials for those stipulated. The result was largely a waste of effort as the input could only be processed with the greatest difficulty. When centres do have access to the correct materials the results are good.

The tests demonstrated that the instructions were not too complicated and could be followed easily by typists even in those centres where the national language is not English.

Little or no trouble was experienced with a variety of papers supplied. Reasonable care was exercised by all centres in the mailing of their input so that in every case it was received in a condition that caused no difficulties in processing. Thus the instructions issued in that respect proved to be adequate.

Summary and Conclusions

(a) The test further broadened the experience with decentralized OCR input, including as it did, input received from both developed and developing countries in temperate climates and from a country with a hot and dry climate. The climatic conditions do not appear to have influenced the results significantly.

(b) The test demonstrated once again the extent to which

successful preparation of input for OCR processing depends on the use of the correct equipment and materials.

(c) Typists appear to have had little difficulty in observing the relatively simple rules for OCR preparation. Few problems were caused by failure to follow them.

(d) On the basis of the results obtained, the IAEA proposes to continue to encourage certain INIS centres to submit INIS input in a form suitable for OCR processing. If necessary, the IAEA would assist those centres in obtaining some of the materials and equipment they require to enable them to do this.

Supplementary Note

Late in 1975 the Czech INIS Centre began to submit INIS abstracts entirely in a form suitable for OCR processing. It was the first INIS centre to do so. The abstracts were typed with the OCR-B typewriter element. The first batch submitted had a recognition error rate of 1 error in 2300 characters read, i.e., an average of 1 error in every six abstracts. All errors were easily corrected on the Autoreader screen having been signaled as uncertain recognitions by the machine. Subsequent results with input from this centre have been equally satisfactory.

An indication of the timesaving OCR represents to the INIS Secretariat is given by the fact that 10 abstracts can be read, and if necessary corrected, on the Autoreader in the time that it would take to punch one abstract.

CHAPTER 7

SUMMARY REVIEW OF THE RESULTS OF THE EXPERIMENT

Introduction

In the proposal for the conduct of an experiment in optical character recognition (OCR) for the processing of AGRIS input that was submitted to IDRC (Appendix I), the IAEA set out specifically to determine:

- (a) what operator instructions are necessary for successful decentralized input preparation for OCR;
- (b) what criteria should be observed in the selection of input preparation staff, with respect to language ability, typing skills, etc.;
- (c) what are the most suitable equipment and materials for input preparation;
- (d) what are the effects of climatic conditions, physical handling, transportation, etc., on the input;
- (e) what is the most suitable OCR equipment from the point of view of range of fonts that can be read, tolerance to various qualities of paper, error correction software, purchase price.

This chapter summarizes our conclusions for each item. Some final comments about the use of OCR for bibliographic data processing are also given.

Summary of Results

Operator instructions

The instructions attached as Appendix V were tested in an actual working environment both at the IAEA headquarters and by three AGRIS centres, in Costa Rica, India, and the Philippines. They proved to be comprehensible and comprehensive enough to

enable all centres but one to prepare input that completely met the formal requirements for OCR. One centre misunderstood one part of the instructions. Appropriate action has been taken to clarify the offending section.

The typists

Experience at the IAEA headquarters demonstrates that an average typist can produce satisfactory input, after a short training period ranging from 2 to 10 days, depending on the individual.

When preparing input for a bibliographic system like INIS or AGRIS, the typist spends most of her time copy-typing. Accuracy is essential. Occasionally some simple decisions are necessary about whether preprinted data should or should not be copied from the worksheet. In addition some characters (very few for AGRIS) need to be encoded.

One of the most important considerations is that the typist (and supervisors) should be aware of the standards that must be met for OCR and that are to some extent different from the standards for correspondence typing. For example, the typescript need not be faultless; errors may be included, provided they are cancelled by means of the error correction routines. On the other hand, even slightly smudged or unevenly typed characters can be fatal for recognition. Thus the typist must be trained to inspect her output from time to time to ensure that the typewriter still produces acceptable quality copy. The simplicity of the equipment used (i.e., a normal typewriter) and the availability of comprehensive correction features greatly facilitate the training of input preparation typists. There are no requirements for language ability additional to those generally required from staff preparing input for AGRIS.

The equipment

Use of an IBM Selectric typewriter is essential for the Autoreader and recommended for the Compuscan. In practice the IBM Selectric II (IBM 82 Series) was more pleasant to work with and produced results superior to those produced by the IBM Selectric I (IBM 72 Series). The model 833, as opposed to others in the IBM 82 Series, has standard 10-pitch spacing that is also essential for OCR. It permits typing of up to 110 characters per line. Accordingly this typewriter is recommended. All tests indicated the essential need to maintain the typewriters in good condition. For the Compuscan, both fonts OCR-B (German) and Perry 199 gave satisfactory results when tested in Vienna. Perry 199 performed acceptably when tested with input prepared by one overseas centre; with input prepared by another overseas

centre the results were poor, but this was mainly due to bad adjustment of one key on the typewriter.

For the Autoreader, Courier 12 initially gave somewhat better results than OCR-B under optimal conditions. However, the results with Courier 12 were much more variable than with OCR-B and Courier 12 produced exceptionally high error rates when the quality of the input was below standard. The results with OCR-B were relatively much more constant and this font is therefore recommended for use in decentralized input preparation. Subsequent experience with OCR-B has proved the wisdom of this decision.

Use of the appropriate once-only carbon ribbons proved to be essential. The IBM 82 Series typewriters use a carbon ribbon that has a much greater output in terms of the number of characters typed relative to the cost of the ribbon than the IBM 72 Series.

The Autoreader was more tolerant to the use of different qualities and sizes of paper than the Compuscan. The Compuscan has a faster and more efficient feed mechanism but provides this at the cost of being quite finicky about the weight and size of the paper being used. The Compuscan was more tolerant than the Autoreader of small specks that might appear in the paper. The Autoreader tried to recognize these more often than the Compuscan did. However, in general the tests proved that the quality of the paper was not a critical factor for Autoreader provided the basic requirements regarding size, weight, and colour stipulated by the manufacturer were met.

Effects of climate, handling, etc.

No recognition problems could be traced to climatic effects or variations in the conditions existing in the centres in which they were prepared. Mailing of OCR input sheets did not adversely affect their readability, provided they were packaged with reasonable care and according to some simple instructions. However, for the Compuscan it was necessary to take special care to protect the left-hand edges of the sheets against dog-earing or wear, as this is the edge with which the sheets are fed into the reader.

Neither the Autoreader nor the Compuscan have great difficulty processing slightly crumpled sheets. Input sheets

⁹The manufacturers' specifications state that "any smooth-surfaced, good quality, non-rag content white bond paper is satisfactory as long as it produces clear, well-defined typed characters. Weight should be between 71-75 grams per square metre (15 to 24 pounds). Standard DIN-A-4 sheets are recommended."

could be processed up to 20 times through the Autoreader before there was any noticeable deterioration in recognition. There was no opportunity to test this on the Compuscan. Multiple handling of the sheets did not cause any deleterious effects on the sheets, provided the letters were not smudged in doing so.

Choice of OCR equipment

Manufacturers' specifications for both machines are included in Appendix II, and the technical details may be compared there.

In the tests the performance of both machines was fairly similar. Perhaps the Compuscan performed slightly better overall than the Autoreader; particularly in the early tests the Autoreader performed quite poorly but the later results exonerated the machine and most recent results are excellent.

The Compuscan was faster than the Autoreader model 5200 with which the tests were conducted. It took 25-30 seconds to process a worksheet through the Compuscan, whereas on the average 1 minute to 80 seconds were required to process a sheet through the Autoreader. There is, however, an Autoreader model available (the 5300) whose rated output speed is twice that of the 5200. This faster model was not available for testing. It should perform at close to the same speed as Compuscan.

Both machines offer a series of editing features that were found to be invaluable in the preparation of input for INIS and AGRIS. The most useful are the deletion symbols, used at the typewriter level to cancel out erroneous characters, words, or lines. Somewhat more care must be exercised in the manual deletion of characters or words (by striking them through with a black felt pen) at the copy-editing level and in the insertion of changes or additions between lines. By the hand-deletion method it is possible to delete accidentally more than what was intended. In inserting data between lines care must be taken to ensure that the typescript is lined up exactly and that the inserted data do not touch existing lines. For decentralized data preparation these methods of input correction are therefore not recommended wholeheartedly. They should only be used by centres that have demonstrated their ability to prepare good input.

Compuscan offers some additional editing features that are not available on the Autoreader. They are intended for insertion of large slabs of additional data at specified points in the original typescript. These features are not very applicable to the purpose for which the equipment was used in the tests.

In routine operation during the tests neither machine reached the rated error rate. It is likely that after a longer period of experience with OCR input preparation, centres would provide progressively better input and the error rate would go down. Certainly this was the experience in the IAEA. However, it may be unrealistic to expect to obtain error rates with decentralized input as low as those that could be achieved when input preparation is centralized and carefully controlled.

During the 8-month period that the Autoreader was on loan to the IAEA only one breakdown occurred but this was repaired quickly by the replacement of a circuit board in the computer. Thus the machine can be said to be reliable in operation. No first-hand assessment of the reliability of the Compuscan could be made. The cost of the Autoreader compares favourably with that of the Compuscan. The price of the Autoreader 5300 is approximately 70% of the price of the Compuscan 170, an approximate price difference of \$30,000.

Final Comments

The experiment provided an excellent opportunity to obtain familiarity with OCR techniques. The results have shown the feasibility of using OCR as a relatively simple technique for the preparation of input to computerized bibliographic information systems on a decentralized basis by countries that do not have the facility for preparing magnetic tape input. The main consideration is that centres preparing input should have access to sound equipment and the correct materials. However, the experiment showed that even flawed input could be processed and corrected very quickly and at a fraction of what it would cost to punch the input centrally.

Appendix I

Proposal for the Conduct of an Experiment in Optical Character Recognition for the Processing of AGRIS Input

1. Objectives

To test the feasibility of creating on a decentralized basis AGRIS input suitable for OCR processing at a central location.

Specifically to determine:

- (a) What operator instructions are necessary for successful decentralized input preparation for OCR;
- (b) What criteria should be observed in the selection of input preparation staff, with respect to language ability, typing skills, etc.;
- (c) What are the most suitable equipment and materials for input preparation (choice of typewriter, font, paper quality, etc);
- (d) What are the effects of climatic conditions, physical handling, transportation, etc. on the input preparation;
- (e) What is the most suitable OCR equipment for input processing from the point of view of range of fonts that can be read, tolerances to various qualities of paper, error correction software, purchase price.

2. Methodology

2.1 Preparation of data for OCR

- (a) Complete standard AGRIS worksheet for a number of bibliographic items;
- (b) Transfer the data from the worksheets to sheets of plain paper using a special typewriter and according to certain specified rules in respect of

- (i) Choice of spacing and alignment of data;
 - (ii) Coding conventions for characters that are not part of the character set supplied on the typewriter's font;
 - (iii) Error correction.
- (c) Therefore no special worksheets need to be designed, but a set of instructions for OCR input preparation must be drawn up before work can commence.

2.2 *Internal testing of OCR input under simulated conditions such as may occur in an average inputting centre*

(a) Use a good average typist whose native language is not English but who has a fair working knowledge of the language.

(b) Instruct the typist, limiting the instructions to those prepared under 2.1 (c).

(c) Typist to transfer data from 50 worksheets to plain paper for OCR input, as in 2.1 (b) above. Note all questions she asks that are not covered by written instructions.

(d) Process the OCR input sheets. Note all read problems and errors. Evaluate results.

(e) Note causes of errors and if necessary adjust instructions accordingly. Also take into consideration questions asked under 2.2 (c) above.

(f) Repeat with different types of typewriter (IBM Selectric 72 and 82), and fonts, different qualities of paper, tabulating results for each experiment. Then repeat again for second OCR device.

Total number of input sheets prepared is $12 \times 50 = 600$ input sheets (see Table I).

2.3 *External testing of OCR input under real conditions in three AGRIS input centres.*

(a) Seek cooperation of three AGRIS centres. They must have access to IBM Selectric typewriter (72 or 82 models). It would be desirable but not essential if two different typists could be employed in each centre, to avoid possible confusion between the input preparation rules for the two different OCR devices that will be used in the experiment.

(b) Mail to these centres:

- (i) 50 worksheets already filled out (data);
- (ii) Supply of plain paper for inputting of the kind that gave best results under 2.2;
- (iii) Typewriter ball fonts;
- (iv) Instructions for preparing input for OCR;
- (v) Instructions for mailing input back to Vienna.

(c) Each centre to complete 100 input sheets on the paper

supplied from Vienna and another 100 input sheets on locally acquired paper. In each case 50 of the input sheets should be prepared according to the instructions and standards required for the ECRM Autoreader, and 50 of the input sheets should be prepared according to the instructions and standards required for the COMPUSCAN. Total input per centre = 200 sheets.

2.4 *Evaluating the results of the decentralized input preparation experiment*

- (a) Process OCR input sheets on two different OCR devices; determine what difficulties, if any, are encountered with each device; determine costs of processing.
- (b) Process tapes by computer.
- (c) Determine causes of operator/hardware errors revealed by computer processing of the data.
- (d) Draw up a final set of instructions and repeat the experiment with one centre, using the final instructions.
- (e) Report on experiment. Final set of instructions to be an appendix to the report.

3. *Estimated requirements for the conduct of the experiment*

3.1 *Staff*

No.	Grade	Duties	Time* (in weeks)	Cost
1	P-4	Supervision of conduct of experiment; write report and final instructions	8-12 (half-time)	\$ 2600 (f.t. rate = \$ 5200)
1	GS-6/7	Assist in conduct of experiment; assist in writing instructions; supervise input preparation; assist in mailing results, tabulating causes of errors, if necessary	8-12 (half-time)	\$ 1500 (f.t. rate = \$ 3000)
1	P-3	Programing support	4	\$ 2000
1	GS-4/5	Preparation of input during internal testing; necessary correction of externally prepared input to provide feedback; some clerical assistance during experiment; typing of correspondence, instructions, final report	8-14	\$ 2100

* It is expected that none of the staff will be occupied full time with the experiment throughout its duration.

3.2 Equipment - purchase

6 only golf-ball fonts for IBM Selectric 82 or 72 typewriters (Courier 12, Perry, OCR-B)	\$	150
--	----	-----

(These become the property of IDRC at the conclusion of the experiment)

3.3 Equipment - rental

Pass some 800 sheets through COMPUSCAN 170	\$	850
Pass some 600 sheets through ECRM 5300	\$	650

(Cost calculated at AS1000 per hour, 60 sheets/hour)

Computer processing (CPU time)

Processing and error checking of data (2000 records) at different times (estimated 15 secs/record) = 10 hours		
Testing and developing of programs	<u>6 hours</u>	
	<u>16 hours</u>	\$ 1500

3.4 Materials

<i>Paper</i> 4000 sheets (A-4 size) various qualities paper	\$	150
<i>Sundries</i> Postage, stationary, airfreighting of worksheets, paper, etc., back and forth from Vienna to input centres, mailing envelopes, cardboard backing etc.	\$	300

3.5 Travel

Possible visit of an INIS staff member to view operational COMPUSCAN operation, e.g., at INSPEC or some other centre in Europe	\$	1000
--	----	------

3.6 Contingencies	\$	700
-------------------	----	-----

Total cost	\$	13,500
------------	----	--------

4. Timetable

A proposed 15-week timetable for the experiment is attached (Table II). This assumes that it will be necessary to repeat the experiment with one external centre for 50 sheets to test the effectiveness of the final instructions.

Table I

VIENNA												OTHER CENTRES											
												CENTRE A		CENTRE B		CENTRE C							
ECRM Autoreader				Compuscan 170								ECRM	Compuscan		ECRM	Compuscan		ECRM	Compuscan				
IBM Selectric 72C		IBM Selectric 82C		IBM Selectric 72C		IBM Selectric 82C				IBM Selectric		IBM Selectric		IBM Selectric									
Courier 12		Courier 12		Perry	OCRB		Perry	OCRB		Courier 12	Perry	Courier 12	OCRB	Courier 12	Perry								
Paper "A"	Paper "B"	Paper "A"	Paper "B"	Paper "A"	Paper "B"	Paper "A"	Paper "B"	Paper "A"	Paper "B"	Paper "A"	Paper "B"	Paper "X"	Paper "C"	Paper "X"	Paper "C"	Paper "X"	Paper "D"	Paper "X"	Paper "D"	Paper "X"	Paper "E"	Paper "X"	Paper "E"
01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

Table II

	March 1975		April 1975				May 1975				June 1975				
	EASTER														
	3-7	10-14	17-21	24-28	31-4	7-11	14-18	21-25	28-2	5-9	12-16	19-23	26-30	2-6	9-13
Preliminary drafting of instructions for typing															
<u>Internal input preparation</u>															
(i) Preliminary test of instructions															
(ii) Preparation of 12x50 worksheets															
(iii) Evaluating results															
<u>External input preparation</u>															
(i) Contact centres re cooperation															
(ii) Despatch material to centres															
(iii) Processing by centres															
(iv) Receipt of results															
Processing results of external data preparation															
Evaluate results															
Prepare and test final instructions															
Prepare final report															

Appendix II

Machine Specifications

(Manufacturer Supplied)

1. Compuscan 170

The Compuscan series 170 Optical Page Readers convert typewritten text in an upper and lower case font to computer-readable form. Damaged characters can be entered from a key-board using the mini-viewer and the CRT display. Alternately, lines containing rejected characters can be identified by a line marker. These pages drop into a hopper.

The key-board and the two display may be used to enter edits on-line at any point in the text by use of a stop character pencil marked on the original.

Character set	Perry 199 OCRB-ECMAII French, German, Scandinavian OCRA, COURIER 12.
Character Pitch	Ten characters per inch.
Throughput	100 characters per second.
Paper size	Maximum 28 x 30cm Minimum 15 x 20cm
Paper weight	77g/m ² .
Line spacing	2½, 3, 5 and 6 lines/inch.
Margins	Top: Adjustable, minimum 1 inch Bottom: Adjustable, minimum 1 inch. Right: Adjustable, minimum ½ inch Left: 1¼ inches.
Non read print	Light blue or red pen marks will not be read by the machine.
Physical dimensions	Length: 190.50cm Width: 75.00cm Height: 98.00cm
Power requirement	118 volts, 30 amperes (not including peripherals) — or 220-240 volts, 15 amperes.
Gross weight	Compuscan 548kg Teletype 51kg

2. ECRM 5200 Autoreader

BASIC SYSTEM (STANDARD)

Scanner	Yes
Computer/Core	PDP-8, 20K
Paper tape reader	250 cps, 6 and 8 level
Paper tape punch	75 cps
Software	Standard software incl. (other optional)
On-line Display Terminal	Optional
Installation, basic training	Yes
On-site Applications training	Yes

2. ECRM - Continued from page 53

THROUGHPUT SPEED (ENGLISH TEXT)*	
Single spaced (60 lines/page)	up to 500 words/minute
Double spaced (30 lines/page)	up to 450 words/minute
Triple spaced (20 lines/page)	up to 400 words/minute
Skip rate	1 inch/second
ERROR RATE	
English text	.01 %
INPUT REQUIREMENTS	
Paper size (stack fed)	8.5 x 6 to 13"
Paper size (roll fed)	8.5 x 6 to 48"
Paper weight	16 lb. to 24 lb.
Paper colour	White or red
Skew tolerance	1/6" over 7"
Stack feeder capacity	20 pages
Typewriter	10 pitch, carbon ribbon IBM Selectric
Fonts read (standard)	Courier 12, OCR-A (numeric) Other fonts are available including several European fonts
Characters read	Upper and lower case alphanumerics, punctuation and symbols
Margins (all sides)	1/2"
Characters/line	0-75
EDITING (STANDARD)	
Cancel character, word or line with single keystroke (typewriter level)	Yes
Horizontal mark deletion by hand for multiple characters and changes (copy editing level)	Yes
Vertical mark deletion by hand for single characters and changes (copy editing level)	Yes
Insertion of changes and additions between lines (requires double or triple spacing)	Yes
Selective scanning of up to 8 separate horizontal areas per page selected via header sheet	Yes
Dropout ink	Red
Editing pen	Red
STANDARD INTERFACE OPTIONS	
PHYSICAL REQUIREMENTS	
Power	Height: 40", Width: 40.5" Depth: 27", Weight: 575 lbs Volts: 115/230 vac Amps: 25/13, Phase: single Line frequency: 60 or 50 Hz
* Application dependent	

Appendix III

IBM Selectric Typewriters

The IBM Selectric typewriters in two model series were found to be suitable for the preparation of OCR input.

These are:

- (1) Model Series 72 Also known as Selectric I
- (2) Model Series 82 Also known as Selectric II

Within these model series, the recommended typewriters are as follows:

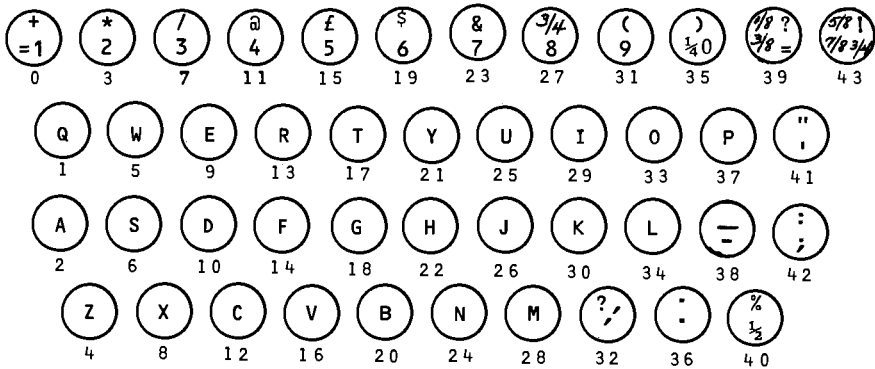
- (1) *Series 72*: Recommended model 713
N.B. A 10-pitch version of the model is required. It will permit typing of 11-inch lines. This is useful, as the rubber rollers that are attached to the metal bar that holds the paper against the platen can then be spaced out wide enough to avoid smearing of already typed text. Recommended carbon film ribbons for this model typewriter are marketed by IBM with Part no. 1136 108.
- (2) *Series 82*: Recommended model 833 (designated 83302 in USA)
N.B. Dual pitch (10 and 12 pitch, interchangeable) versions are available. However, a 10-pitch only model is recommended. This typewriter is also available with the correcting feature. This should not be purchased if the typewriter is used only for preparation of OCR input. Model 833 has a 13.5-inch paper capacity. The film carbon ribbons for the typewriter are marketed by IBM with Part no. 1136 390.

A recommended keyboard for use with the OCR-B (English) golf ball is shown in the attached diagram. It can be obtained by special order modification of the standard "English 984" keyboard, which is the keyboard available ex-stock that most closely resembles the special keyboard required.

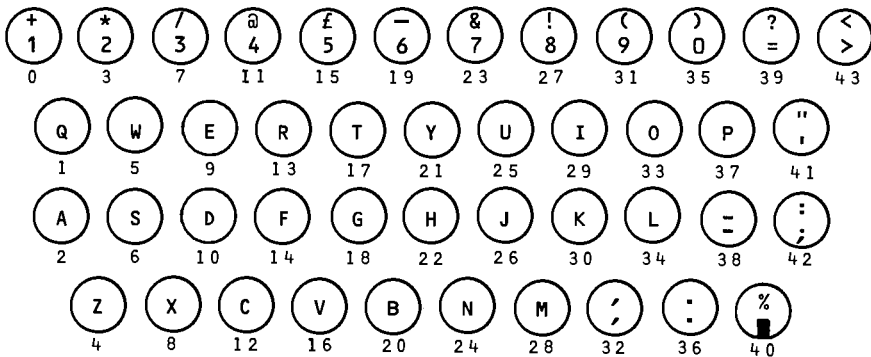
Exact pattern, as typed with the OCR-B (English) golfball
 (IBM part no. 6522598) with a machine wired like England 984

first row (shifted)	+ * / @ £ - & ! () ? <
(unshifted)	1 2 3 4 5 6 7 8 9 0 = >
second row (shifted)	Q W E R T Y U I O P "
(unshifted)	q w e r t y u i o p '.
third row (shifted)	A S D F G H J K L _ :
(unshifted)	a s d f g h j k l - ;
fourth row (shifted)	Z X C V B N M , . %
(unshifted)	z x c v b n m , . ■

England 984 from factory



England 984 after modification



Appendix IV

Fonts and Coding Systems

Golf ball samples

Each row of keys typed shifted and then unshifted on a machine wired like England 984.

Perry 199

Courier 12

<p>! ¶ @ \$ % / & * () } Δ 1 2 3 4 5 6 7 8 9 0 { ▣ Q W E R T Y U I O P + q w e r t y u i o p = A S D F G H J K L - : a s d f g h j k l - ; Z X C V B N M , . ? z x c v b n m , . /</p>	<p>! @ # \$ % & * () + [1 2 3 4 5 6 7 8 9 0 =] Q W E R T Y U I O P " q w e r t y u i o p ' A S D F G H J K L Δ : a s d f g h j k l - ; Z X C V B N M , . ? z x c v b n m , . /</p>
--	--

OCR-B English

OCR-B German/Swedish

<p>+ * / @ £ - & ! () ? < 1 2 3 4 5 6 7 8 9 0 = > Q W E R T Y U I O P " q w e r t y u i o p ' A S D F G H J K L _ : a s d f g h j k l - ; Z X C V B N M , . % z x c v b n m , . ▣</p>	<p>° + # = * ? ↑ () / ö ▣ 1 2 3 4 5 6 7 8 9 0 ö \$ Q W E R T Y U I O P Ä q w e r t y u i o p ä A S D F G H J K L _ `. a s d f g h j k l - `. Z X C V B N M " : Å z x c v b n m , . å</p>
---	--

OCR Coding system developed to cope with the
120-character INIS character set :


<u>INIS</u>	<u>OCR-B</u>	<u>Courier 12</u>	<u>INIS</u>	<u>OCR-B</u>	<u>Courier 12</u>
A-Z	A-Z	A-Z	∫	∫integral∫	#integral#
a-z	a-z	a-z	√	∫root∫	#root#
0-9	0-9	0-9	>	∫gt∫	#gt#
\$	∫dollar∫	\$	<	∫lt∫	#lt#
=	=	=	∫HI∫	∫XI∫	#XI#
,	,	,	∫Ω∫	∫OMEGA∫	#OMEGA#
(((%	%	%
)))	0	*H0	*H0
[*([1	*H1	*H1
]	*)]	2	*H2	*H2
;	;	;	3	*H3	*H3
:	:	:	4	*H4	*H4
/	/	/	5	*H5	*H5
-	-	-	6	*H6	*H6
,	,	,	7	*H7	*H7
.	.	.	8	*H8	*H8
*	∫ast∫	#ast#	9	*H9	*H9
+	+	+	+	*H+	*H+
α	*a	*a	-	*H-	*H-
β	*b	*b	0	*L0	*L0
γ	*g	*g	1	*L1	*L1
μ	*m	*m	2	*L2	*L2
ν	*n	*n	3	*L3	*L3
π	*p	*p	4	*L4	*L4
ω	*o	*o	5	*L5	*L5
Δ	∫DELTA∫	#DELTA#	6	*L6	*L6
Λ	∫LAMBDA∫	#LAMBDA#	7	*L7	*L7
Σ	∫SIGMA∫	#SIGMA#	8	*L8	*L8
♀	∫female∫	#female#	9	*L9	*L9
♂	∫male∫	#male#			
→	∫yields∫	#yields#			

Codes developed for Stage II of the experiment
 (subsequently altered and improved)

<u>AGRIS</u>	<u>OCR-B</u>	<u>PERRY 199</u>
A - Z	A - Z	A - Z
a - z	a - z	a - z
0 - 9	0 - 9	0 - 9
.	.	.
,	,	,
:	:	:
;	;	;
/	/	/
=	=	=
+	+	+
-	-	-
\$	<i>£</i> dollar <i>£</i>	<i>£</i>
%	%	%
*	<i>£</i> ast <i>£</i>	*
^	<i>£</i> lt <i>£</i>	*LT
v	<i>£</i> gt <i>£</i>	*GT
u	*([
U	*)]
(((
)))
:	<i>£</i> ex <i>£</i>	:
?	?	?
' (apostrophe)	'	'
" (quotes)	"	@

Appendix V

Letter Sent to AGRIS Centres Participating
in the OCR Experiments

	INTERNATIONAL ATOMIC ENERGY AGENCY AGENCE INTERNATIONALE DE L'ENERGIE ATOMIQUE МЕЖДУНАРОДНОЕ АГЕНТСТВО ПО АТОМНОЙ ЭНЕРГИИ ORGANISMO INTERNACIONAL DE ENERGIA ATOMICA	TELEPHONE: 52 45 11 52 45 25 TELEX: 1-2645 CABLE: INATOM VIENNA
KÄRNTNER RING 11, P.O. BOX 590, A-1011 VIENNA, AUSTRIA		
IN REPLY PLEASE REFER TO: PRIERE DE RAPPELER LA REFERENCE:		
Dear		
<p>Thank you for agreeing to assist us with our experiment in optical character recognition (OCR). This experiment is being conducted by the INIS Section of the International Atomic Energy Agency under contract to the International Development Research Centre (IDRC), Ottawa, Canada. Its purpose is to determine the feasibility of creating on a decentralized basis AGRIS input suitable for OCR processing at a central location.</p>		
<p>I am sending you under separate cover by first class air mail the following:</p>		
<ul style="list-style-type: none">- Two IBM golf-ball typing elements for use in your IBM Selectric typewriter (Model I or II). The fonts which we are sending you are OCR-B (English) and Perry 199;- Two sets of 100 sheets each of input paper;- Two sets of instructions for the preparation of input for OCR. The first set is labelled "Compuscan", the second set is labelled "ECRM"(*)- One set of 50 completed AGRIS worksheets (Xerox copies)- One only IBM carbon film ribbon.		
<p>(*) Note that whilst there are great similarities between the instructions for the Compuscan and those for the ECRM, there are some significant differences, particularly in the correction procedures. Ideally two typists should be used: one to prepare input for the Compuscan and the other for the ECRM. If this is impossible and only one typist is available, then the differences in the instructions should at least be drawn to her attention.</p>		

I would be grateful if you would use these materials in the preparation of four batches of input for OCR processing, as follows:

- (1) Batch AA: typed for the Compuscan 170 OCR processor on paper provided by us;
- (2) Batch AB: typed for the Compuscan 170 OCR processor on paper provided by you from local stocks;
- (3) Batch BA: typed for the ECRM 5200 Autoreader on paper provided by us;
- (4) Batch BB: typed for the ECRM 5200 Autoreader on paper provided by you.

The attached matrix shows which of the materials should be used for the preparation of each batch. The following points may be noted:

N.B.1: The Perry 199 typing element can be easily recognized by the fact that the number "199" is engraved on it in the small space to the left and below the black lever used to insert the element.

N.B.2: The OCR-B typing element can be recognized by the characters "OCR-B ECMA 11", which appear in white writing on the top of the ball.

N.B.3: For Batch AA you should use the paper which has the words "Compuscan Training Form R-1" printed on it. For Batch BA you should use the other variety of paper that we have provided.

N.B.4: For Batches AB and BB use locally produced paper according to the following specifications: Any smooth-surfaced, good quality, non-rag content white bond paper is satisfactory as long as it produces clear, well-defined typed characters. Weight should be between 71-75 grams per square meter (15 to 24 pounds). Standard DIN A-4 sheets are recommended. Use the sheets provided by us as a guide to the paper quality required. Instead of using preprinted margins ensure that the typist follows the instructions for setting the margin as given (instructions vary for Batches AB and BB!).

In the preparation of the various batches the typist should be able to follow the instructions as given in the relevant manuals. They have been written in a non-technical way and it should therefore be possible to proceed with the minimum of verbal instruction and supervision. However, there are two things which the typist will need to be told:

- 3 -

(1) Numbering the pages. In the instructions the typist is asked to number each page with a batch number. She will be asking you to tell her what the batch number is. Please provide her with the information as given above.

(2) Tag 001 (TRN field) The typist has been instructed not to copy the information in this field from the AGRIS worksheets but to come to you for instructions. Please tell her that she should substitute the following for the first four characters of the TRN as found on the worksheet:

Characters 1-2: The appropriate AGRIS code for your centre.
E.g. IN for India, PH for the Philippines, etc.

Characters 3-4: She should write
11 for Batch AA
22 for Batch AB
33 for Batch BA
44 for Batch BB

The remaining 5 characters of the TRN should be copied precisely as they are on the AGRIS worksheet.

For example: The TRN of the first AGRIS worksheet you have been sent is FD7490001. If your Centre's code is IN (for India) then you should number the first OCR sheet in Batch AA IN1190001.

Please send us your completed OCR input sheets as soon as possible and in any case so that they reach us by no later than 21st July 1975. Care should be taken to ensure that the OCR input sheets are not damaged in transit. It is therefore recommended that a stack of sheets be enclosed between two sheets of stiff cardboard, slightly larger all-round than the input sheets, before being securely wrapped into a parcel. Please enclose with the input sheets the IBM typewriter golf-balls, which we would like to have back. It is not necessary to return any of the other material that was sent.

Please address your parcel to me at the INIS Section of the International Atomic Energy Agency, P.O. Box 590, A-1011 Vienna, Austria. The parcel should be sent by air.

May I thank you in advance for your assistance? Please do not hesitate to contact me if you have any difficulties.

Yours sincerely,

Hans W. Groenewegen
INIS Section
Division of Scientific and
Technical Information

Batch	Typewriter	Font	Ribbon	Paper	Instructions	Data
AA	IBM Selectric	Perry 199 (1)	Carbon Film	As supplied (3)	Labelled "Compuscan"	50 AGRIS worksheets
AB	IBM Selectric	Perry 199 (1)	Carbon Film	Local (4)	Labelled "Compuscan"	50 AGRIS worksheets
BA	IBM Selectric	OCR-P (2)	Carbon Film	As supplied (3)	Labelled "ECRM"	50 AGRIS worksheets
BB	IBM Selectric	OCR-B (2)	Carbon Film	Local (4)	Labelled "ECRM"	50 AGRIS worksheets

- (1) See N.B.1 in the text
- (2) See N.B.2 in the text
- (3) See N.B.3 in the text
- (4) See N.B.4 in the text

Typist's Instructions for OCR Input (Rev.1)

Introduction

You are now going to learn how to produce AGRIS input in a form that looks like a normal typewritten page but that can immediately be read by a computer, in other words without the need to repunch the information into paper tape or into 80-column cards. Computer reading of typewritten pages (typescripts) is possible because of the invention of a technique called "optical character recognition" (OCR). Machines known as optical character readers have now been developed that can recognize certain letters, figures, and symbols and that can convert them automatically into codes. The machines then punch those codes into paper tape or record them on magnetic tape. The paper tapes or magnetic tapes can then be processed by the computer.

If you have ever been associated with the preparation of work for computers you will know that they are very stupid machines that need to be told exactly what to do. Unlike human beings they cannot reason and they cannot deal with information unless it comes to them in a precisely defined predetermined form. This is also true for OCR.

Type fonts

For example, you know that almost everybody writes letters and numbers in their own individual way. Yet most people can easily recognize the letters and numbers written by other people no matter how they have been written. Here are examples of a variety of ways in which we could print the Latin alphabet.

```
; "=%&()'$/:~ QWERTZUIOPÜ ASDFGHJKLÖÄ YXCVBNM?!  
1234567890ß qwertzuioüp asdfghjklöä yxcvbnm,.-
```

```
; "=%&()'$/:~ QWERTZUIOPÜ ASDFGHJKLÖÄ YXCVBNM?!  
1234567890ß qwertzuioüp asdfghjklöä yxcvbnm,.-
```

```
; "=%&()'$/:~ QWERTZUIOPÜ ASDFGHJKLÖÄ YXCVBNM?!  
123456789+ß qwertzuioüp asdfghjklöä yxcvbnm,.-
```

```
+*/@f-8!()%: QWERTYUIOP< ASDFGHJKL?" ZXCVCNM,.-  
1234567890■; qwertyuiop> asdfghjkl=' zxcvbnm,.-
```

You will notice that the way in which each letter is written varies from group to group. You may also notice that the letters in each group are designed in such a way that they are alike in overall design, e.g., the letters in the first group have a square look about them, the letters in the second group slope

forward, the letters in the third group have little hooks (serifs) on them, which the letters of the fourth group do not have. Each group is known as a "font." Surely everybody who can read the Latin alphabet will recognize the letters in each of these fonts, even though they are made quite differently. However, the computer can only recognize the letters in the last font; the others it cannot recognize at all.

The first thing you need to do therefore when you produce information for OCR is to be sure that you have the correct font in your typewriter. The people who design fonts give them a name, e.g., the fonts shown above are called Polygo Pica 10, Light Italic 12, Delegate 10, OCR-B.

The font that we want you to use is called OCR-B.* Make sure therefore that this is the font that you have in your typewriter. If you are not sure please check with your supervisor or a representative of the manufacturer of the typewriter you are using.

The typewriter

There are two typewriters that we know from experience will give good results for OCR. The two typewriters we recommend are both manufactured by IBM and are known as the "IBM 72" and the "IBM 82." Both are so-called golf-ball typewriters because the type is stored on a little round ball approximately the size of a golf ball and not on the end of a set of levers as is the case with most other typewriters.

A special feature of the IBM 72 and 82 typewriters is that it is a simple matter to exchange the golf ball that contains one font for another golf ball with a different font. Thus, if you have access to an IBM 72 or 82 typewriter you can buy for it a golf ball that has fonts suitable for OCR.

If you are using an IBM 72 typewriter, make sure that it only types 10 characters to the inch. Some typewriters type 12 characters to the inch. They are not suitable for OCR work. On the IBM 82 typewriter it is possible to change the setting for 12 characters to the inch to 10 characters to the inch and vice versa. So if you are using an IBM 82 typewriter make sure it is set to type 10 characters to the inch.

There are some other things about the typewriter that you must have before you can start preparing input for OCR. One of the most important things about OCR is to ensure that the characters that you type are clear and distinct. You therefore cannot type with a worn-out ribbon and still expect the OCR to read what you

* English version, as shown in bottom left-hand example on page 57.

have typed. To avoid any danger of your ribbon being worn-out use a once-only carbon ribbon in your typewriter. Not all typewriters are equipped to take such ribbons. You must be certain that you have a typewriter that can.

Another factor that helps you to type clear and distinct letters is to ensure that the type on the typewriter is kept clean. If you have not already acquired a habit of cleaning your typewriter each day according to the manufacturer's instructions get into the habit now. Also if you do a lot of typing during the day take time off to clean the type from time to time. It will only take a few minutes and the results will be worth it.

Getting familiar with the typewriter

It is essential that you should know your typewriter intimately if you are going to produce good typescripts. This is even more important when you are preparing typescript for OCR. Your supervisor will know this and will therefore give you plenty of time to get used to your typewriter before you are asked to start typing in earnest. Here are some things you should do to get to know the machine.

(a) Read the instruction book that comes with the typewriter carefully. Be sure you understand everything in it. If you cannot understand some things ask your supervisor or the representative of the manufacturer to explain them to you.

(b) Be sure that the characters reproduced on the keys are in fact those that will be typed when you strike those keys. This will not always be true when you have an IBM golf-ball machine. You may have a golf ball that has some characters that are different to those shown on the keyboard.

Test the keyboard by striking each key first with the upper case shift off and then with the upper case shift on. Are the two characters that you have just typed the same as those shown on the key in the lower and upper position? If not, make a simple label showing what characters you have actually typed and stick it on the key. Repeat this for each key on the keyboard until you have exact correspondence between the characters on the keyboard and the characters produced on the typescript when you strike those keys.

(c) Now try out the machine with some exercises. Ask your supervisor to give you some material to copy and see whether you can produce an acceptable typescript. Keep going until you are thoroughly familiar with the typewriter and its capabilities. Get someone to check your work and show you your mistakes. Ask yourself whether the mistakes were caused by your lack of experience with the machine. If so, note the reason for your mistakes and satisfy yourself that you can now avoid them.

Getting ready to prepare some input for OCR

There are certain rules that you must follow in preparation of a typescript for OCR regardless of what information you happen to be typing. We will go through these rules first to make sure you understand the general principles. It may help you if you look at the attached OCR input form that has already been completed so that you can see exactly what we mean by each rule.

(a) *Inserting the paper in the typewriter*

Notice that in the sample the typewritten lines run exactly horizontally across the page. They do not slope up or down even a little bit. It is very important that your typewritten lines should also be exactly horizontal because the OCR reader will have difficulty in reading lines that slope even a little and may not be able to read badly sloping lines at all. Therefore, ensure that your paper is inserted squarely in the machine.

(b) *Setting the margins*

Your left- and right-hand margins should be set at least 2.5 cm. You should leave at least 2 cm at the top of the page before you start typing and leave at least 2 cm at the bottom. In the sample you can see how these distances have been marked all around the page by red lines forming a rectangular frame. You should only type within this frame. Note that there are dotted vertical and horizontal lines on the page. You may type over them. They are there simply as a warning that you are reaching the right-hand margin and the bottom of the page, respectively.

(c) *Setting the typewriter controls*

If you have a multiple copy control on your typewriter (which you will have if you are using an IBM 72 or 82) set it for single copy typing. On the IBM 72 or 82 that means setting it in the most forward position. Set your impression controls at the highest possible point, i.e., to provide the strongest possible impression of the type on the paper. If you are using an IBM 72 or 82 set the little lever that you will find at the right of the golf-ball typing element to the position marked 3. Finally you should set the vertical spacing control to produce vertical spacing of 3 lines to the inch.

(d) *Handling the typescript*

You should be very careful in handling your typescript so that it does not become smudged, stained, torn, or wrinkled. We want the OCR reader to read what you have typed accurately and without problems and you should therefore make sure that your typescript looks clean and neat when you have finished it. Also make sure that there are no rubber stamps or other unnecessary

information on it.

(e) *Laying out your typescript*

Start typing as close as possible to the left-hand margin of the frame of the sheet (see example). Never type outside the frame.

(f) *Errors*

Everybody makes mistakes and so will you. A little later in this manual we will tell you what to do when you make a mistake in your typing. Here we want to tell you about the things that you should not do when you make an error.

- (i) Do not backspace and type over the error;*
- (ii) Do not "x-out" the error (i.e., do not type a row of x-es across the incorrect word or words);
- (iii) Do not try to erase your error by using a typewriter eraser, razor blade, etc.;
- (iv) Do not try to cover over your error by using some white correcting solution or by pasting a slip of white paper over it.

(g) *Take care in typing certain characters*

In normal typing typists often use the small letter "l" (lower case "l") to stand for the numeral one and the upper or lower case letter "o" to stand for the numeral "0" (zero). You are not permitted to do this for OCR. You will find that you will have on your typewriter keys the numerals 0 and 1. Please take extreme care to strike those keys when you wish to type those numerals.

(h) *Underlining*

You should never underline any words or characters when you prepare input for OCR.

(i) *Breaking words at the end of a line*

If you reach the end of the line in the middle of typing a word you have nothing to worry about. Simply type an "=" (equal sign) and continue typing. Note that you don't need to observe any of the usual rules for breaking words between lines such as you would observe in normal typing. For OCR you may break a word anywhere you like because the computer that will process the OCR output will delete the equal sign and join the two words together again without a gap. Note, however, that if you wish to break a hyphenated compound word at the hyphen, you do not need to add the equal sign. When the last text character of a line is hyphen (-)

* Certain exceptions to this are explained in correction procedures (see Supplement).

the computer will preserve the hyphen and join the first word on the next line directly to the hyphen. If you do not want the word on the next line joined, you must leave a space after the hyphen and then type an equal sign. Note also that if you want to preserve an actual sign at the end of a line, two equal signs are required since one will be deleted by the computer.

Study the examples below and you will soon understand the principles.

Examples:

- (a) *Lines typed as follows:*
Manitoba Univ., Winni=
peg (Canada);
will be output as follows:
Manitoba Univ., Winnipeg (Canada);
- (b) *Lines typed as follows:*
Manitoba Univ., Winnipeg (=
Canada);
will be output as follows:
Manitoba Univ., Winnipeg (Canada);
- (c) *Lines typed as follows:*
The use of a computer=
driven photocomposition device is contemplated.
will be output as follows:
The use of a computerdriven photocomposition device is
contemplated.
- (d) *Lines typed as follows:*
The use of a computer-
driven photocomposition device is contemplated.
will be output as follows:
The use of a computer-driven photocomposition device is
contemplated.
- (e) *Lines typed as follows:*
methyl, ethyl, n- =
and isopropyl
will be output as follows:
methyl, ethyl, n- and isopropyl
- (f) *Lines typed as follows:*
Cross sections for A=242 and v==
2200 m/s
will be output as follows:
Cross sections for A=242 and v=2200 m/s

Error correction

You have already been told about all the things that you should not do when you notice that you have made a mistake during typing. If you make a mistake in the line that you are currently typing, then there are easy ways of correcting the mistake by using one of the three deletion symbols: @, &, /. The use of these symbols and other possibilities for correction of errors are explained in the attached supplement.

Supplement

Error Correction Procedures for OCR Input

1. Correcting normal characters

(all characters except @ & / and space)

It often happens that you strike the wrong key and immediately realize your mistake. To correct your error simply do the following: type a single @ ("at" symbol) immediately after the wrong character that you really meant to type and go on typing. If you have typed more than one character wrong in the same word, you may use as many deletion symbols as the characters which you wish to eliminate. For short words, it is probably easier to eliminate the whole word (explained later), but for long words, and complicated formulae, multiple character corrections can be very useful.

The deletion symbol will not delete a space in this manner. If a @ is typed after a space, the last typed character will be eliminated, but the space will remain (see example f below). It is therefore important to remember not to put a space between the character(s) which you want to delete and the deletion symbol. (However no harm is done if character(s) are eliminated at the end of a word; in this case multiple spaces after the word are eliminated by the OCR software (see example g below)).

Examples:

a) A line typed as follows:

Cente@ro Internacional di@e Mejoramiento de Maiz
will be output by the OCR device as follows:
Centro Internacional de Majoramiento de Maiz

b) A line typed as follows:

500!p.5@ 5-7
will be output as follows:
500!p. 5-7

c) A line typed as follows:

401!!@ottawa (Canada)
will be output as follows:
401!ottawa (Canada)

d) A line typed as follows:

200!The relationshoaa@ip between DNA syntka@thesis
will be output as follows:

200!the relationship between DNA synthesis

e) A line typed as follows:

200!Extraction of 1-phenyl-3-methyl-5@4-benyoza@@zoyl
will be output as follows;

200!Extraction of 1-phenyl-3-methyl-4-benzoyl
(Note: the "o" in -benyqz@@zoyl was correct, but
also eliminated in order to correct the "y").

f) A line typed as follows:

401!Ottawa (K @Canada)

will be output incorrectly as follows:

401!Ottawa (Canada)

g) A line typed as follows:

401!Ottawaoo @@ (Canada)

will be output as follows:

401!Ottawa (Canada)

II. Correcting or deleting words

Sometimes you may want to cancel an entire word or many words which you have typed and start afresh. A "word" in this context is defined as a group of characters (letters and/or numbers) preceded by a space. Thus, in this context, the following groups of characters each represent a word:

001!IA7500001

401!Ottawa

(Nigeria);

1-phenyl-3-methyl-4-benzoyl

fDELTAfLsub(s)

*H1*H2N(2*H-

You may cancel a word by using the & (ampersand) as a deletion symbol. You may use as many of these deletions as you wish, if there is no word to the left of the & no deletion will occur. If you use more &'s than words on the line, the line will be eliminated and the additional &'s ignored. You may leave a space between the word you are eliminating and the &, but this is not mandatory. Also a space may or may not be left between consecutive &'s (i.e. &&& or & & & are equally effective). After the & we would like you to leave a space for easier proofreading; if you have not, however, the OCR will still give the right output (see example a, third line). Also, words may be deleted which have already been corrected with a @ (see example c below).

Examples:

a) A line typed as follows:

International Tricit & Triticale Symposium (or)

International Tricit& Triticale Symposium (or)

International Tricit&Triticale Symposium

will be output as follows:

International Triticale Symposium

b) A line typed as follows:

310!IDRC--042e & 310!IDRC--024e

will be output as follows:

310!IDRC--024e

c) A line typed as follows:

200!Effect of nam@ny& many pra@@article, many hole

will be output as follows:

200!Effect of many particle, many hole

d) A line typed as follows:

Induced *H24Na actovity && *H2*H4Na activity (or)

Induced *H24Na actovity & & *H2*H4Na activity

will be output as follows:

Induced ²⁴Na activity

III. Deleting an entire line

If for any reason you wish to delete a whole line in your typescript, you may do so by typing a space followed by a slash (/). Used in this way the slash causes the OCR reader to ignore all characters (letters and numbers) and spaces which occur on that line back to the left hand margin. You should then recommence typing on a new line.

Please note that if you want to use the slash as a character in the data you are typing you may do so, provided you use it without a space immediately preceding it. It is the combination "space-slash" which causes a line to be deleted, and the combination should therefore never be a part of the data which you are typing. Use of the combinations "character-slash-character" or "character-slash-space" are permitted and will be read without difficulty by the machine.

Examples:

a) Lines typed as follows:

```
009!M
100!MacIntyre, R; Campbell, M. ((ed /
100!MacIntyre, R.; Campbell, M. (eds.)
will be output by the OCR reader as follows:
009!M
100!MacIntyre, R.; Campbell, M. (eds.)
```

b) Lines typed as follows:

```
009M /
009!M
100!Mc@acIntyre, R.; Cam & Campbell /
100!MacIntyre, R.; Campbell, M. (eds.)
will be output by the OCR reader as follows:
009!M
100!MacIntyre, R.; Campbell, M. (eds.)
```

c) Lines typed as follows:

004!N

008!E30/N/AM/K

008!F30/N/AM/KV

will be output by the OCR reader as follows:

004!N

008!F30/N/AM/KV

IV. Deleting a space between two words

A space cannot be deleted by using the @ sign. There are however two other ways. By far the best way is to type two &'s this will delete both word particles, (see examples a and b). You must remember then to type the whole word again. A second method is to leave the space on the worksheet and after finish typing put a vertical stroke with a black felt pen in the space you are trying to delete, (example c). This is a bit tricky, however, for the stroke must extend a bit above the top of the letters and it cannot lean to the left or right. If you have already typed a word or two and cannot use the method of inserting &'s, it is probably wiser to delete the whole line, (example d).

Examples:

a) A line typed as follows:

200!Extraction of hafnium by&& by

will be output as follows:

200!Extraction of hafnium by

b) A line typed as follows:

200!The neutron-to- proton && neutron-to-proton ratio

will be output as follows:

200!The neutron-to-proton ratio

c) A line typed as follows and corrected by black felt pen

200!Extraction of 1-phenyl-3-methyl-4-ben|zoyl

will output as follows:

200!Extraction of 1-phenyl-3-methyl-4-benzoyl

d) Two lines typed as follows:

```
860!The author obtains a general cri terion for the
860!The author obtains a general' criterion for the
will be output as follows:
860!The author obtains a general criterion for the
```

V. Deleting the @ & and .

Of course an unwanted @ can be negated by simply retyping the character in front of it, (see example a). This will not work however, if you have also mistakenly left a space between the @ and the character. In this case back track and fill the space with an X, then delete the whole word, (example b). The same is true for the &; here however, the word may be long and complicated, and you do not wish to repeat it. If this is the case you must type the ampersand once again, back space twice, and overstroke both ampersands with a horizontal bar (i.e. ~~&&~~). The same method can be used for deleting an unwanted /; simply type another slash, and then overstrike them with verticle lines (i.e. ~~//~~). Care must be taken in using the horizontal bar and not the shorter hyphen (i.e. - and not -). Usually the horizontal bar is a shift of one of the numbers, whereas the hyphen is located near the period and comma.

Examples:

a) A line typed as follows:

```
850!Engl@lish
will be output as follows:
850!English
```

b) A line first typed as follows:

```
850!En @glis h
must be edited as follows:
850!EnX@glis h& 850!English
and will be output as follows:
850!English
```

- c) A line typed as follows:
 850!English& 850!English
 will be output as follows:
 850!English
- d) A line typed as follows:
 200!Extraction of 1-phenyl-3-methyl-4-benzoyl&& solution
 will be output as follows:
 200!Extraction of 1-phenyl-3-methyl-4-benzoyl solution
- e) A line typed as follows:
 200!Extraction of 1-phenyl-3-methyl-4-benzoyl solution #
 will be output as follows:
 200!Extraction of 1-phenyl-3-methyl-4-benzoyl solution

VI. Correcting the _ (underscore)

Four characters on the OCR-B golfball, the < (less than sign), > (greater than sign), ■ (black square) and _ (underscore) are forbidden. If you have accidentally typed one of the first three mentioned, simply eliminate it with the "a" or the whole word in which it appears with the "&". However, this will not work for eliminating the _ (underscore); the OCR machine creates an extra line for it. In this case you must backspace and type a letter which will touch the underscore, such as "g" (lower case only). Then you may delete the character with the "a" or the word with the "&". It does not matter whether a space or another character was above the underscore, before you backspaced and overtyped with the "g".

Examples:

- a) A line typed first as follows:
 500!v. 24(256) p. 1_2
 should be edited as follows:
 500!v. 24(256) p. 1g2a@-2
 and will be output as follows:
 500!v. 24(256) p. 1-2

b) A line typed first as follows:

200!Common scab in seed_potato

should be edited as follows:

200!Common scab in seedpotato & seed-potato

and will be output as follows:

200!Common scab in seed-potato

c) A line typed first as follows:

401!Katmandu (Nepal)

should be edited as follows:

401!Katmandu (~~gggg~~) & (Nepal)

and will be output as follows:

401!Katmandu (Nepal)

(overtyped with "g's")

PREPARATION OF AGRIS DATA FOR OCR PROCESSING

Now you are ready to begin preparing some data for OCR processing. The following instructions will tell you how to proceed. It may help you to understand the instructions if you also study the examples in Appendix 1 carefully.

The AGRIS worksheet

As you know, the AGRIS input sheet is designed to record information about documents, such as journal articles, chapters in books, conference papers, etc. In order to make it easier to record the information in a way that will permit it being processed by the computer later, the AGRIS input sheet is subdivided into a number of data fields. A data field is really just a space into which information is placed. Each piece of information relating to a document is recorded in its own field, and the fields are distinguished from each other by numbers or "tags".

Approximately 5 1/2 cm from the top of the AGRIS input sheet you see a double horizontal line. The data fields which occur above this double line are printed in the form of boxes, which are numbered 001 to 008. Below the double lines most of the data fields consist of rectangular spaces of varying height. Each space (or field) is separated from the next by a horizontal line and each field is labelled with an explanation of the data that should be entered in the field (e.g. "Personal Name", "Corporate Name", etc.) and the tag number that applies to that field (e.g. 100, 110, etc.). In the following explanation we will use the words "field" and "tag" over and over again, so it is important that you understand what we mean by them.

Copying the data from the AGRIS input sheet to the OCR form

Let us now see how the data should be copied for the AGRIS input sheet to the OCR form. First a number of general rules:

(1) When to start on a new OCR input sheet

Always start copying the information from a new AGRIS input sheet on a new OCR input sheet. There is one exception to this rule, namely when you have two AGRIS input sheets which together describe a single document. You will know that they belong together because the letter/number combination printed in the field numbered 001 (the TRN field) is identical on both AGRIS input sheets. In this case you should not go to a new OCR input sheet to copy the information from the second AGRIS input sheet. Instead simply continue typing on the first OCR input sheet until it is full. Only if you still have more information to copy which does not fit on the first OCR input sheets do you move to a second sheet. In this case please write the TRN with a red felt pen on the top of the second sheet (see example). In no case should you retype the tag 001! and TRN on the second sheet.

(2) Blank fields

In copying data from the AGRIS input sheets you should always ignore any fields that have not been filled out on the form. Simply pass to the next field in which there is data.

(3) Tag numbers

Each time you copy some information from a field on the AGRIS input sheet, you should first copy the tag number that belongs to that field. Each tag number that you copy should begin on a new line, as close as possible to the left hand margin of your OCR input sheet. After you have typed the tag number (do not forget the leading zeroes in 001 to 009!) you should type an exclamation mark (!) and then start typing the data itself. The exclamation mark must be used (e.g. 001!); encoding the exclamation mark should be done only when it appears in the text (e.g. Let us begin with OCR input *fexf*). There should be no spaces between the tag number and the exclamation mark or between the exclamation mark and the data.

(4) Continuation Lines

If you need more than one line to type the data for a given field, start the second and succeeding lines as close as possible to the left hand margin of your OCR input form. Do not indent your continuation lines.

(5) Copying the data exactly

Always be sure to copy the data exactly as it appears on the AGRIS input sheet. In particular be careful not to omit any spaces or full stops. At the same time do not put in any spaces or full stops which do not appear on the original AGRIS input sheet. Note that in general the data in each field is not terminated by a full stop.

Special instructions relating to certain tags

There are some special rules which you should observe when copying certain fields. The following are the tag numbers of the fields to which special rules apply:

Tag 001 (TRN field)

This is the most important piece of information and must always be typed. Please type this field exactly without any error corrections. Special care must be given in the usage of the letter "0" and number zero "0" (e.g. 001!M0750001).

Tag 002

Please be careful to copy all three characters which you will find in the field, i.e. the slash as well as the two numbers on either side of the slash. (see examples).

Tag 003

On the AGRIS input sheet two letters ("R" and "W") have been printed in this field. However, you should treat this field as if it were blank, (i.e. pass to the next item) unless a circle has been drawn around one of the letters. In that case you should treat the letter that has been circled as a piece of data that has been entered in that field.

Tag 004

As in the previous field, ignore the preprinted letters which have not been circled and simply copy the one that has a circle around it (but do not forget the tag number and the exclamation mark!). Note that tag 004 may never be blank.

Tag 005, 006, 007

The same as for tags 003 and 004.

Tag 008

This is a complicated field, but if you look at the examples in the Appendix, whilst you read the instructions you should not have too much trouble understanding the instructions.

The first part of this field consists of three pre-printed boxes, separated by semicolons. You will come across some AGRIS input sheets on which one or two of these boxes are empty. This is how you should proceed. First copy the contents of the first box (a letter followed by a 2-digit number). Now check if there is something in the second box. If there is, copy the semicolon and then copy the information in the second box. Then look at the third box. If there is information in that box, type another semicolon and copy the information from the third box. Now you can go to the other end of the dotted line which is printed on the AGRIS input sheet starting from the end of the three preprinted boxes and copy the slash.

However, the second and/or third boxes may be empty. Then, if you come to an empty box, do not copy the semicolon printed in front of that box but go directly along the dotted line to the point where you meet the slash and then type the slash.

Once you have typed the slash, look at the next set of boxes. Each has a preprinted letter in it; the letters being B,C,D,F,G,H, J,P,R, and T. A circle will have been drawn around one and only one of these letters. Copy the letter which has a circle around it and then copy the slash which you find on the AGRIS input sheet at the end of this set of boxes.

Now you find another set of 4 boxes with the preprinted letters A,M,S and C. Proceed as before, i.e. copy only the letter or letters which have circles around them. This time, however, before you copy the slash which is printed at the end of this set of boxes, look at the next set of boxes, to check if any letters in that set are circled. If no letters in the final set of boxes have a circle around them, you need not do anything more, simply pass to the next field (tag 009). However, if one or more letters in the first set of boxes have a circle around them then you must copy the slash which is printed on the AGRIS input sheet in front of that last set of boxes and must copy the letters which have circles around them.

Tag 009

This field consists of a single box with a letter written or printed in it. This field occurs twice on the AGRIS input sheet; the first time just below the horizontal double line; the second time approximately 7 cm above the bottom of the sheet.

Each time that tag 009 occurs on the AGRIS input sheet it is printed at the head of a block of other fields. The first block is identified by a big "1", printed to the left of tag 009. This block consists of 21 fields, numbered from 100 to 620. The second block is identified by a "2" and consists of 6 fields, numbered from 230 to 610. You will find that quite often on the AGRIS input sheets no information has been entered in the second block of fields. If that is the case, then you should not copy on to your OCR input sheet the tag 009 which is printed at the head of this second block of fields. In other words, in that case the last field which you should copy is tag 620. (See examples).

Tag 100 - Tag 610

Copy the information exactly as it appears on the AGRIS input sheet. Do not type a full stop at the end of any field unless it occurs on the AGRIS input sheet. Please note that the tags must not necessarily be typed in ascending order within one block of tags. (See example 3).

Tag 620

This field consists of five sets of boxes. In the last two sets of boxes a "G" has been printed in the first box. Again, not all the sets of boxes will have been filled in on all the AGRIS input sheets. When copying the information in Tag 620 you should follow a similar procedure to the one you followed when you copied the information in tag 008.

For example, if the first set of boxes has been filled in, copy the information onto the OCR input sheet. Then check: is the second set of boxes filled in? If so, then type the semicolon and copy the information from the second set of boxes. Then check the third set of boxes. If it has also been filled, type another semicolon and copy the information from the third set of boxes. As soon as you come to a set of boxes that has not been filled in, look to see if the "G" boxes contain information. If not, tag 620 is finished. If yes, then type the slash and then type the information which is entered in the first set of boxes following the slash (including the preprinted "G"). Also type the slash if none of the first three sets of boxes has been filled in, before typing the set with "G" (e.g. 620!/G514). If there is also information in the last set of boxes type the semicolon and the information entered in the last set of boxes. If there is no information in the final set of boxes, you have finished with tag 620 and you can go on to the next field. Again, a study of the examples should make the above explanation clear.

The Second AGRIS input sheet

As explained earlier, you will find on occasions that two AGRIS input sheets have been used to record all the information regarding a certain document. When you come to copy the information from the second AGRIS input sheet, continue typing on your OCR page with tag 002. Do not copy the information in tag 001 from the second AGRIS input sheet, as this is the same as the information in tag 001 on the first AGRIS input sheet and is therefore superfluous. (See example 1).

End of record

When you have copied all the information relating to a certain record, you should finish with a record terminator (end of document code). The end of document code is *RT (without intervening spaces). Note that the asterisk (*) must be used; encoding the asterisk is to be done only when it appears in the text (e.g. 610!fastfIDRC...) It should be typed on a new line following the last line of data and as close as possible to the left-hand margin of the OCR form (see example).

Special codings

Certain characters which are part of the AGRIS character set do not appear on the golfball, and have been used as control characters. The attached Table gives a comparison of the AGRIS character set versus the OCR-B (English) character set and shows the coding conventions to be adopted for those characters which are not available on the golfball.

Note: When using the OCR-B golfball you should be aware that in a number of cases a single character from the AGRIS character set is encoded by means of a combination of an asterisk (*) followed by another character or by means of the delimiter character f. Although the asterisk is available on the OCR-B golfball it should only be used for encoding special characters and for the end of document indication (*RT). If it occurs in the text as a character in its own right it should be encoded by means of the f delimiter as shown in the table.

Furthermore you should be aware that the exclamation mark is used as tag delimiter and, if encountered in normal text, special encoding must be used (fexf). Please note that four characters, the underscore "_", the less than sign "<", the greater than sign ">", and the square "■" are forbidden.

TRN

(fill in by hand with red felt pen only)

001!XL7506569
002!1/2
004!N
008!F26;E10/G/AM/V
009!A
110!Instituto Geografico Nacional, Guatemala City
200!*(Map of soil actual use *(Guatemala*(*)@@@)*)
230!Mapa: uso actual del suelo
403!1967
600!(Es)
610!Map. 32 x 82 cm. (scale 1:125,000)
620!/G356
002!2/2
009!M
100!Aguilera Vizcarra, H.E.
110!Universidad de San Carlos de Guatemala, Guatemala City.
Facultad de Agronomia
111!Tesis (Ing Agr)
200!*(Use and development of water resources of the Maria
Linda Basin for irrigation*)
230!Uso y aprovechameinto& aprovechamiento de Los recursos
hidraulicos de La cuenca del Rio Maria Linda para Riego&riego
401!Guatemala City (Guatemala)
403!Sep 1974
*RT

*Please note that tag 001 was not repeated after tag 620.
Note also the use of character and word correction symbols.*



008 **F 2 0 E 1 0** : [] [] []
 (Principal) (Secundaria)
 CATEGORIAS DE MATERIA

000 **AGROT/AGRIS**

C. I. Año No. consecutivo
1 F 7 5 0 6 5 6 9

001 **X L** NTR
 NR relacionado (NTR)

002 **1 / 2**
 No. de la hoja No. total de hojas

003 **R W**
 Revisión Eliminación
 Cambio

004 **N C D**
 Nuevo Cambiado Eliminado
 Status del registro

005 [] [] [] [] []
 NR afectado

005 **T /**
 Traducción Ganéctica

TIPO DE DOCUMENTO
 Monografía Ensayar Dibujo Folleto Mapa o Atlas Disco Artículo Patente Informe Prod. de Comput. / **B C D F G H J P R T**

NIVEL BIBLIOGR.
 Analítica Monográfico Púst. seriada Colectivo / **0 0 0 S C**

INDICADOR DEL TIPO DE LITERATURA
 Conferencia Diccionario Dicco. numéricos Teas e Diver. Legislación Bibliografía (Mapas) (Incluido) Sumario No-convenc. / **K L N U W Z Y E**

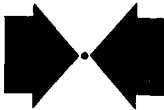
1 009 **A** (Utilice una hoja de entrada para cada nivel bibliográfico señalado y llene la casilla 009 con la letra pertinente. Para publicaciones seriadas usa la sección 2 de esta hoja de entrada)

		Campo	Datos
		100	(Usa siempre máquina de escribir)
Autor(es) personal(es) (Institución donde trabaje)			
Autor(es) Corporativo(s)		110	Instituto Geografico Nacional, Guatemala City
Grado académico		111	
Título en Inglés	Título principal	200	<u>Map of soil actual use [Guatemala]</u>
	Elementos secundarios	201	
Conferencia	Nombre	210	
	Lugar	211	
	Fecha	213	
Título Original	Título principal	230	Mapa: uso actual del suelo
	Elementos secundarios	231	
Edición		250	
No. Informa/Patente		300	
Números adicionales		310	
ISBN/IPC		320	
Pie de Imprenta	Lugar de publicación	401	
	Cese edit.	402	
	Fecha publ.	403	1967
Colación		500	
Idioma del texto		600	(Es)
Notas		610	Map. 32 x 82 cm. (scale 1:125,000)
Cod. de Objetos y Geogr.		620	[] [] [] [] ; [] [] [] [] ; [] [] [] [] / G 3 5 6 ; G [] [] [] []

2 009 **S**

Título de la publ. seriada	Título principal	230	
	Elementos secundarios	231	
ISSN		320	
Fecha publicación		403	
Colación		500	
Notas		610	

AGRINTER



HOJA DE ENTRADA

000 AGRINTER/AGRIIS

C. I. Año No. consecutivo

001 147506569

NTR

NR relacionado (NTR)

002 2/2 No. de la hoja No. total de hojas

003 R W Cambio Revisión Eliminación

004 N C D Nuevo Cambiado Eliminado

005 NR afectado

006 Traducción / Genérica T /

008 (Principal) (Secundaria) CATEGORIAS DE MATERIA

TIPO DE DOCUMENTO Monografía, Estándar, Dibujo, Película, Mapa o Atlas, Disco, Artículo, Pasante, Informe, Prod. de Congres.

NIVEL BIBLIOGR. Analítica, Monográfico, Publ. seriada, Colectivo

INDICADOR DEL TIPO DE LITERATURA Conferencia, Diccionario, Datos numéricos, Tests o Diar., Legislación, Bibliografía, Mapas (Incluidos), Sumario, No-conven.

1 009 M (Utilice una hoja de entrada para cada nivel bibliográfico señalado y llene la casilla 009 con la letra pertinente. Para publicaciones seriadas use la sección 2 de esta hoja de entrada)

Table with columns: Nivel, Campo, Datos. Rows include: Autor(es) personal(es) (100) Aguilera Vizcarra, H.E.; Autor(es) Corporativo(s) (110) Universidad de San Carlos de Guatemala; Grado académico (111) Tesis (Ing Agr); Título en Inglés (200) /Use and development of water resources of the Maria Linda Basin for irrigation/; Conferencia (210-213); Título Original (230) Uso y aprovechamiento de los recursos hidraulicos de la cuenca del Rio Maria Linda para riego; Edición (250); No. Informe/Patente (300); Números adicionales (310); ISBN/IPC (320); Pie de Imprenta (401-403) Guatemala City (Guatemala), Sep 1974; Colección (500); Idioma del texto (600); Notas (610); Cod. de Objetos y Geogr. (620)

2 009 S

Table with columns: Nivel, Título de la publ. seriada, Elementos secundarios, ISSN, Fecha publicación, Colección, Notas. Rows include: Título de la publ. seriada (230); Elementos secundarios (231); ISSN (320); Fecha publicación (403); Colección (500); Notas (610)

AGRINTER/Form. 1

TRN

(fill in by hand with red felt pen only)

001!XL7506483
002!1/1
004!N
008!F28/B/M/YV
009!AaM
100!Lohier, G.; Duvivier, L.; Dorville, R.; Dorismond, P.;
Denizard, J.R.
110!Departement de l'Agriculture desXR&, /
110!Departement de l'Agriculture, des Ressourd@ces Naturelles
et du Developpement Rural; Conseil National de Developpement
et de Planification; Institute@ de Developpement Agricole
et Industriel; IICA, Port-au-Prince (Haiti)
200!*(Project on erosion control of the Hospital Mountain
for the protection of the city of Port-au-Prince *(Haiti)*)
230!Projet controle de l'erosion au morne l'Hopital pour
la protection de la ville de Port-au-Prince
310!11LH/74
401!Port-auPrince& 401!Port-au-Prince (Haiti)
403!1974
500!122 p., suppl: 9 p.
600!(Fr)
610!18 tables; 4 maps, 8 ref. fastfBibliotheque de la Faculte
D'Agronomie et de Medecine Veterinaire, Port-au-R@Prince,
(Haiti)
620!/G324
*RT

Note use of line deletion and word deletion symbols.



HOJA DE ENTRADA

008 **F28** : : :
 (Principal) (Secundaria)
 CATEGORIAS DE MATERIA

000 **AGRINTER AGRIS**

C. J. Año No. consecutivo
 001 **FE 7500483**

NTR

NR relacionado INTR)

002 **1 / 1**
 No. de la hoja No. total de hojas

003 **P W**
 Reemplazo Eliminación

004 **N C O**
 Nuevo Cambiado Eliminado

005 : : : : :
 NR afectado

006 **T /**
 Traducción Gendécima

007 : : : : :
 NR relacionado INTR)

TIPO DE DOCUMENTO
G C O F H J P R T

NIVEL BIBLIOGR.
A M S C

INDICADOR DEL TIPO DE LITERATURA
K L N U W Z

1 009 **M**
 NIVEL

(Utilice una hoja de entrada para cada nivel bibliográfico señalado y llene la casilla 009 con la letra pertinente. Para publicaciones seriadas use la sección 2 de esta hoja de entrada)

Cam-po		Otros (Use siempre máquina de escribir)
Autor(es) personal(es) (Institución donde trabaja)	100	Lohier, G.; Duvivier, L.; Dorville, R.; Dorismond, P.; Denizard, J.R.
Autor(es) Corporativo(s)	110	Departement de l'Agriculture, des Ressources Naturelles et du Developpement Rural; Conseil National de Developpement et de Planification; Institut de Developpement Agricole et Industriel; IICA, Port-au-Prince (Haiti)
Grado académico	111	
Título en Inglés	Título principal	200 /Project on erosion control of the Hospital Mountain for the protection of the city of Port-au-Prince /Haiti/ /
	Elementos secundarios	201
Conferencia	Nombre	210
	Lugar	211
	Fecha	213
Título Original	Título principal	230 Projet controle de l'erosion au morne l'Hopital pour la protection de la ville de Port-au-Prince
	Elementos secundarios	231
Edición	250	
No. Informe/Patente	300	
Números adicionales	310	11LH/74
ISBN/IPC	320	
Pie de Imprenta	Lugar de publicación	401 Port-au-Prince (Haiti)
	Casa edit.	402
	Fecha publ.	403 1974
Colación	500	122 p., supl: 9 p.
Idioma del texto	600	(Fr)
Notas	610	18 tables; 4 maps, 8 ref. *Bibliotheque de la Faculte D'Agronomie et de Medecine Veterinaire, Port-au-Prince, (Haiti)
Cod. de Objetos y Geogr.	620	/ G 3 2 4 ; G

2 009 **S**
 NIVEL

Título de la publ. seriada	Título principal	230
	Elementos secundarios	231
ISSN	320	
Fecha publicación	403	
Colación	500	
Otros	610	

001!XL7506484
002!1/1
)a008!F00/B/MS/V
004!N
009!M
200!*(Summary report, 10a960-61 and 1973-74 *(cotton,
Venezuela*))
110!Centro Nacional de Investigaciones Agropecuarias,
Maracay (Venezuela). Instituto de Investigaciones Agronomi=
cas. Seccion Algodon
230!Informe resumen, campana 106a960-61 a 1973-74
401!Maracay (Venezuela)
403!1974
500!68 p.
600!!a(Es)
610!fastfServicios de Biblioteca y Documentacion. Centro
Nacional de Investigaciones Agropecuarias, Maracay
(Venezuela)
620!0430/G536
009!S
230!Informe Resumen - Centro Nacional de Investigaciones
Agropecuarias (Venezuela)
*RT

Please note that margins can be larger, but not smaller than 2.5 cm. Also note that tag 004 was typed after 008 and tag 110 after 200; these will be properly sorted by the computer.



008 **000** (Principal) (Secundaria) CATEGORIAS DE MATERIA

000 **AGRINTER/AGRI**
 C. A. No. consecutivo
 001 **7506484**

002 No. de la hoja **1/1**
 No. total de hojas

003 Cambio
 Revisión Eliminación
R W

004 Status del registro
 Número Cambiado Eliminado
N C D

005 NR afectado

006 Transcripción
 007 NTR
 NR relacionado (NTR)

TIPO DE DOCUMENTO
B C D T F G H J P R T

NIVEL BIBLIOCR. INDICADOR DEL TIPO DE LITERATURA
A M S C K L N U W Z Y E

1 009 **M** (Utilice una hoja de entrada para cada nivel bibliográfico señalado y llene la casilla 009 con la letra pertinente. Para publicaciones seriadas use la sección 2 de esta hoja de entrada)

Cam-p-o		Datos
(Use siempre máquina de escribir)		
Autor(es) persona(s) (Institución donde trabaja)	100	
Autor(es) Corporativo(s)	110	Centro Nacional de Investigaciones Agropecuarias, Maracay (Venezuela). Instituto de Investigaciones Agronomicas. Seccion Algodon
Grado académico	111	
Título en Inglés	Título principal 200	<u>Summary report, 1960-61 and 1973-74 /cotton, Venezuela/</u>
	Elementos secundarios 201	
Conferencia	Nombre 210	
	Lugar 211	
	Fecha 213	
Título Original	Título principal 230	Informe resumen, campana 1960-61 a 1973-74
	Elementos secundarios 231	
Edición	250	
No. Informa/Patente	300	
Números adicionales	310	
ISBN/IPC	320	
Pie de Imprenta	Lugar de publicación 401	Maracay (Venezuela)
	Casa edit. 402	
	Fecha publ. 403	1974
Colación	500	68 p.
Idioma del texto	600	(Es)
Notas	610	*Servicios de Biblioteca y Documentacion. Centro Nacional de Investigaciones Agropecuarias, Maracay (Venezuela)
Cod. de Objetos y Geogr.	620	0430 ; ; ; / G 536 ; G

2 009 **S** NIVEL

Título de la publ. seriada	Título principal 230	Informe Resumen - Centro Nacional de Investigaciones Agropecuarias (Venezuela)
	Elementos secundarios 231	
ISSN	320	
Fecha publicación	403	
Colación	500	
otas	610	

AGRINTER/Form. 1

CHARACTER SETS

	<u>1. normal</u>	<u>AGRIS</u>	<u>OCR-B</u>	<u>Remarks</u>
USED IN TYPING AGRIS DATA 		A - Z	A - Z	
		a - z	a - z	
		0 - 9	0 - 9	care must be taken not to use letter "0" for number zero "0"
		.	.	
		,	,	
		:	:	
		;	;	
		/	/	a space cannot precede this character
		=	=	used also for word continuation at end of line
		+	+	
		- (hyphen)	-	do not confuse hyphen "-" with long dash "--"
		%	%	
		((
))	
		?	?	
	' (apostrophe)	'		
	"	"		
	<u>2. encoded</u>	[*(
]	*)	
		\$	<i>fdollarf</i>	
		*	<i>fastf</i>	
		<	<i>fltf</i>	this must be encoded; use of < is forbidden
		>	<i>fgtf</i>	this must be encoded; use of > is forbidden
		!	<i>fexf</i>	if in text, this must be encoded; use of ! is for tag delimiter only
	<u>3. control</u>		@	character delete
			&	word delete
			W/	line delete (space must precede /)
			*	used for [,], and record terminator (*RT)
			!	tag delimiter
			- (long dash)	used for deletion of & and / (see correction procedures, part V)
	<u>4. forbidden</u>		_ (underscore)	should never be used; for deletion of this character (see correction procedures, part VI)
			<	should never be used, to delete use @
			>	should never be used, to delete use @
			■	should never be used, to delete use @

Appendix VI

Specifications for an OCR Machine Suitable

for Bibliographic Data Processing

(INIS and AGRIS)

INPUT: Stack-fed typewritten sheets prepared on a standard IBM Selectric typewriter.
Paper size: DIN-A-4, scanned area should not be less than 15 cm x 23 cm.
Paper colour: White, with the possibility of having nonscan colour background printing.
Paper quality: The equipment should be tolerant to some variations in weight, whiteness, and finish of the paper.

TYPE FONTS:

- (i) Standard ECMA-II OCR-B font in the full 121 character set;
- (ii) Recognition of full Cyrillic alphabet in upper and lower case should be delivered to the IAEA within 4 months of acceptance of our order;
- (iii) The successful tenderer should be prepared to enter at a later stage into a contract with the IAEA for the development of specifications for a typewriter element and for the preparation of a recognition program to read the full 120-character INIS character set (Appendix IV). Simultaneously with the submission of their offer for the supply of an OCR device, tenderers should submit an offer to cover the development of such specifications and software.

PROCESSING: *Read rate:* not less than 100 characters/second.
Error rate: not more than 1 recognition error for each 10,000 characters read.

Editing features:

- (i) *On input sheets:*
 - Cancellation of 1 single character by typewriter or manually (editing pen)
 - Cancellation of 1 single word by typewriter or manually (editing pen)
 - Cancellation of 1 single line by typewriter or

manually (editing pen)
Insertion editing (changes and additions) by
typing of words and lines.

(ii) *During processing:*

Facility for viewing characters that have been
read and for correcting, altering, deleting,
or adding characters, words, or lines during
processing through console panel or keyboard.

- OUTPUT: (i) 8-level paper tape according to ISO standards and
recommendations. Punch rate not less than 75
characters/second.
- (ii) 9-track, IBM compatible $\frac{1}{2}$ inch magnetic tape, 800
b.p.i., NRZl.
- (iii) On-line interface with IBM 370 series computers
should be possible at additional cost.

SOFTWARE: Full documentation with all application and
recognition software should be provided. If required
the tenderer should be prepared to provide
assistance to IAEA programing staff in altering
and amending software.

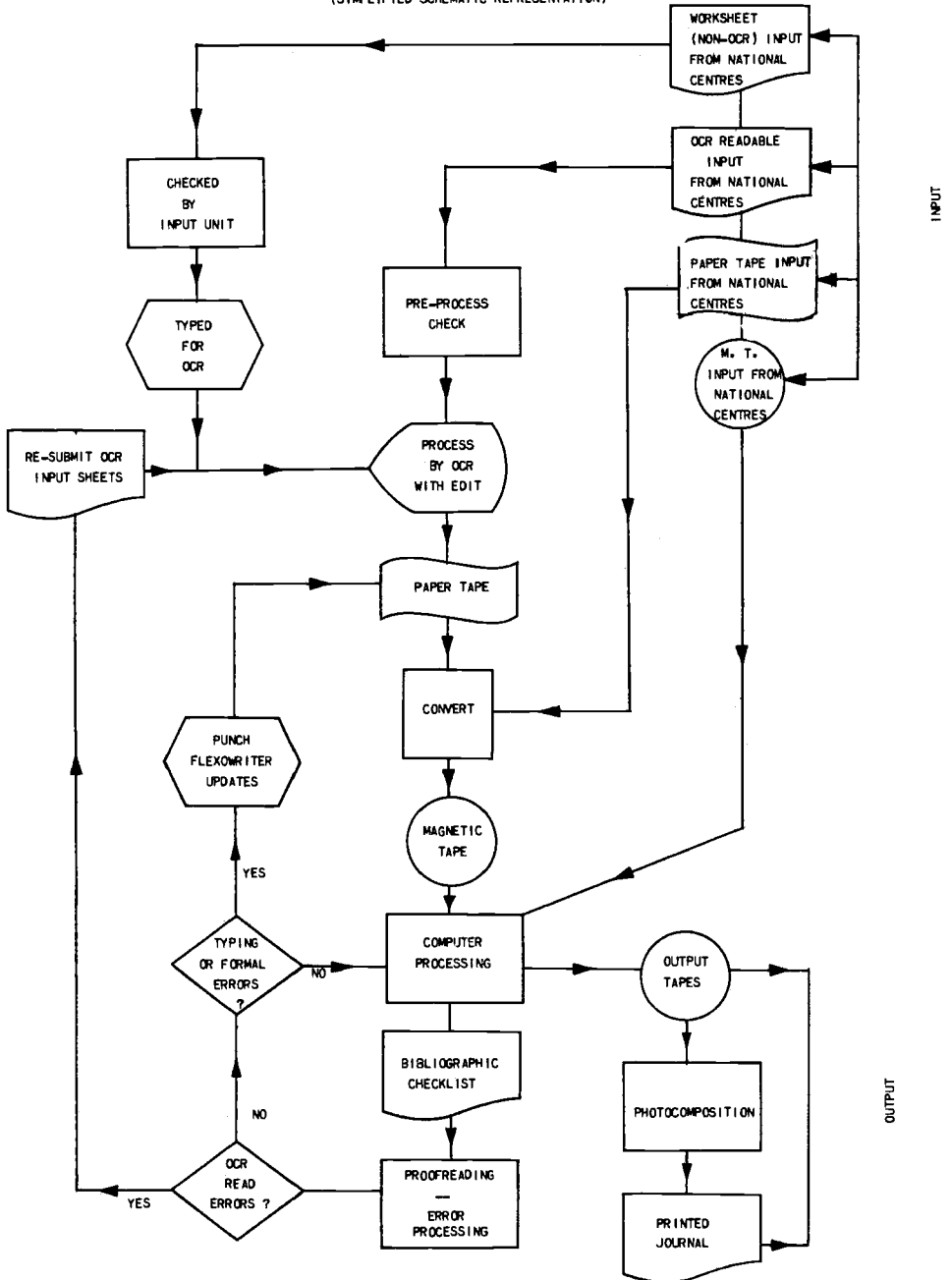
TRAINING: Facilities for training operating staff should be
made available at the IAEA premises.

POWER SUPPLY: 220 volts, 50 cycles.

Appendix VII

Integration of OCR Techniques into INIS and AGRIS Processing Cycle

PROCESSING CYCLE FOR INIS & AGRIS INTEGRATING OCR TECHNIQUES
(SIMPLIFIED SCHEMATIC REPRESENTATION)



N.B.1: For AGRIS, input from one centre would normally not be mixed, i.e., the centre would submit, either on worksheets, or on paper tape, or on magnetic tape.

For INIS, input may be mixed, e.g., bibliographic data could be on worksheets and abstracts in a form for OCR processing, etc. When this happens *Computer Processing* includes a matching operation.

N.B.2: The Autoreader 5300 purchased by the IAEA produces output directly on magnetic tape thus eliminating the paper tape to magnetic tape conversion step.

N.B.3: Some worksheet input is still converted to machine-readable form by methods other than OCR, e.g., Flexowriter, magnetic tape encoder. For the sake of simplicity these alternative methods are not shown in the diagram.

N.B.4: Only serious OCR reading errors are corrected by resubmission of the OCR worksheets. Minor errors are corrected by the punching of Flexowriter updates.

N.B.5: Output tapes are prepared at the end of each processing cycle. A number of updates can take place prior to each cycle.

N.B.6: The diagram does not show quality checking routines performed by INIS Subject Specialists.

