

The FACIT Project
LIB FACIT / 1-1044



Technical Report no 5

**Retroconversion of Older Card
Catalogues using OCR and
Automatic Formatting.**

Project Overview and Final Report

Niels Erik Wille

November 1996

General Preface to the FACIT Reports

The present report is one of five presenting the results of the FACIT project.

FACIT is short for Fast Automated Conversion with Integrated Tools. The project has been supported by the EU under the Libraries section of the Telematics Programme. It started in January 1993 and finished in February 1996.

The project has been concerned with two main questions:

- 1 To determine the feasibility of converting older card catalogues into modern OPACs using scanning, Optical Character Recognition, and automatic formatting into a bibliographic format (such as UNIMARC).
- 2 To develop a prototype system capable of handling automatic formatting, and automatic or semiautomatic detection and correction of the errors produced by scanning and OCR.

The first objective has been achieved in the sense that the project has shown that such retroconversion can only be expected to be feasible under certain conditions.

The Achilles' heel of fully automatic procedures in retroconversion is still the speed and quality of OCR. And this depends to a large extent on the state of the source material. The project was based on the assumption that commercially available equipment and OCR programs would be able to handle older typewritten and printed catalogue cards in a satisfactory way, so that the main effort could be aimed at formatting the cards, but very much more time has been spent on problems of OCR than was originally envisaged.

The results concerning OCR are based mainly on the use of commercially available scanners and OCR packages in the lower or middle price range (such as seem attractive to most libraries). The results may have been different if more sophisticated (and more costly) equipment had been used, or even custom built

equipment. But in general the conclusion has to be that many older card catalogues are not suitable for this type of methodology because of the state of the source material: Yellowed by age, worn, smudged, with handwritten additions, sometimes swollen or made uneven by dampness, written with a series of typewriters with varying typefaces and with ribbons that are more or less worn out, copied by stencilling etc.

The conclusion is not that the methodology is not feasible at all, but that its application is limited to fairly "well behaved" catalogues. A library wondering whether to apply scanning and OCR to retroconversion should carry out extensive tests in order to assess the suitability of this.

Formatting the cards after scanning and OCR does not, on the other hand, seem to present serious problems, if the output from OCR have a low level of errors. Based on a thorough formal analysis of the catalogue and the rules used in producing it, it will in most cases be possible to write a series of programs specific to that catalogue to do the job. This is confirmed by other projects.

The main focus of the project was to investigate the possibility of producing one application, able to handle a wide range of card catalogues as found in European libraries, avoiding the necessity of writing the formatting programs from scratch every time. This is done by feeding the application a formal description of the catalogue at hand, using a relatively simple formal language. At the same time the application should provide a set of integrated tools for the range of different procedures that go into retroconversion work. The project has demonstrated that this is in fact feasible.

But the work of formal analysis is quite demanding, both in terms of time and the necessary skills and knowledge. And it will have to be done again with each new catalogue, since no two catalogues are exactly alike. This process is needed in order to produce the necessary formal specifications for the formatting programs, both with a system like the FACIT Prototype and with custom built formatting programs. This is definitely a specialist job.

With automatic conversion a fair amount of the costs go into setting up and testing the system with each new library and each new catalogue. This means that this methodology is not suitable for a small or medium size library to handle alone without expert assistance - from a commercial service or a large library that has already done some work in this area.

An important problem that has not been solved in a satisfactory way in this project, is the need for detecting and correcting errors produced by scanning and OCR. The project has investigated various possible solutions, and it seems worth while to pursue this further. Meanwhile corrections will have to be done by the human operator with some support from the computer.

The second objective of the project, as stated above, has only been partly reached. A software package has been developed that is able to demonstrate the principles involved in automatic formatting of library catalogues and in customizing the procedures for use in libraries with widely different cataloguing practices as well as catalogues produced over time to different specifications. But the package does not include more advanced facilities for error detection and correction, and it still lacks a series of features that are necessary for use in large scale conversion of catalogues.

Nevertheless the results of the project are promising for further development work, and constitute a solid basis for future work by the partners and the subcontractors of the project as well as others. The aim of the published reports is therefore to make available the information generated by the project, in order to help making realistic judgements about the prospects of using the methodology described in a particular library for the conversion of a particular catalogue, and in order to make the information useful for other research and development projects.

The published reports from the FACIT project consists of the following:

Optical Character Recognition for Retroconversion of Catalogue Cards: Hardware, Software and Character Representation. By Niels Erik Wille. (FACIT Technical Report no 1). Statens Bibliotekstjeneste, Copenhagen. October 1996.

The report summarizes the experiences with scanners and OCR programs. Special treatment is given to the question of character sets and representation of characters, since this is normally of great importance in converting multilingual catalogues.

A Framework for the Analysis of Catalogue Cards. By Niels Erik Wille and Vera Valitutto. (FACIT Technical Report no 2). Statens Bibliotekstjeneste, Copenhagen. Revised version, October 1996.

The report describes the problems involved in analysing a catalogue in order to evaluate the feasibility of converting it by automatic means, as well as the formal language to be used in setting up the FACIT Prototype. This information should also be useful for someone aiming at developing similar tools for retroconversion.

Error Analysis and Correction in Retroconversion. By Hans Erik Jensen (FACIT Technical Report no 3). Statsbiblioteket, Aarhus. October 1996.

The report summarizes the issues involved in automatic or semiautomatic error detection and correction, and outlines plans for further development of the Prototype in order to incorporate more sophisticated handling of OCR errors.

The FACIT Prototype. Manual and Documentation. By SYNERGI (FACIT Technical Report no 4). Statens Bibliotekstjeneste, Copenhagen. October 1996.

The report describes the Prototype in detail and the procedures to use when setting up the demonstration version. The level of information is highly technical. Due to a series of limitations the demonstration Prototype is not suitable for large scale conversion work, but using it with a smaller sample will provide a good grasp of the problems and procedures involved in automatic formatting etc.

Retroconversion of Older Card Catalogues using OCR and Automatic Formatting. Project Overview and Final Report. By Niels Erik Wille (FACIT Technical Report no 5). Statens Bibliotekstjeneste, Copenhagen. November 1996.

This report presents the project as a whole and the main results reached. It includes a summary of the information included in the previous reports.

These reports are available free of charge.

A workable demonstration version of the FACIT Prototype is available. This is a combination of a suite of DOS programs and an interface produced as an application for Microsoft Access. The Prototype will run on a PC with Windows 3.11 or Windows 95 and Microsoft Access 2.0 or later versions. The Demon-

stration Prototype is available free of charge for use in European libraries.

All correspondence concerning the reports and the Prototype should be sent to:

Niels Erik Wille
Senior lecturer
Dept. of Computer Science, Communication and
Education
Building P4
Roskilde University
P.O.Box 260
DK-4000 Roskilde

or posted by e-mail to: new@snow.ruc.dk (Internet)

Information about the project and copies of the reports and the demo-version of the FACIT Prototype are also available on the World Wide Web at the address: <http://www.komm.ruc.dk/FACIT/>

The reports are in PDF format requiring an Acrobat Reader for reading and printing. The demo-version is in ZIP format.

Contents

Project Summary and Results	8
A. Hardware and Software for Scanning of Catalogue Cards	10
B. Optical Character Recognition	11
C. Representation of Character Sets	14
D. Analysis of Catalogue Cards	16
E. Formatting of Input	16
F. Error Detection and Correction	17
G. The FACIT prototype	20
H. General conclusions	21
Background of the Project	23
A. General background	23
B. Rationale of the project	27
C. Wider framework of the project	29
Workplan of the Project	31
Hardware and Software Platform	36
A. General Specification of the Platform	36
B. The Input System used	37
C. Alternatives to the proposed Input System	38
D. Equipment for the Main Processing system	38
The FACIT Prototype	40
A. Introduction	40
B. Scanning and OCR	41
C. Internal Record Format	43
D. Internal 16-bit Character Set	44
E. Formatting Specifications	45
F. Dictionaries	47
G. Specification of Output Format	48
H. Outline of the Overall Process of Using the Facit Prototype.	49
Calculation of Costs	52
A. Introduction	52
B. Cost of acquiring the necessary equipment	52
C. Cost of setting up the formatting system, including formal analysis of the catalogue(s) to be converted	53
D. Cost of producing the plain text input for the formatting process	54

E. Cost of proof reading and correction of results	55
F. Cost of formatting and production of output files in the desired format	56
G. Cost of introducing retroconverted records into the electronic catalogue system of the library	57
H. Comparisons of Total Costs	57
Concluding Remarks	58
A. Compliance with international standards	58
B. Benefits and Results	59
C. Exploitation plans	61
References	62
Appendix 1: Names and Adresses of Participants	68
Appendix 2: List of reports and other products of the project	71

Project Summary and Results

The FACIT-Project (Fast Automated Conversion with Integrated Tools) was initiated in 1991 with an application for support from the Telematics Programme. The starting time was January 1993 with a scheduled finishing time in March 1995. After two prolongations the actual finishing time was the end of February 1996.

The main objective of the project was to contribute to the development of tools for fast and relatively cheap large scale conversion of catalogue cards from different libraries and different time periods, with special regard to pre-ISBD catalogues. As part of this objective the project should produce a working prototype for automatic formatting and automatic or semiautomatic error detection and correction of scanned catalogue cards.

The prototype was to be based on an existing hardware and software platform for conversion of the cards into a plain text-format, using image scanning and optical character recognition (OCR). This input had then to be formatted according to a cataloguing format specified by the user, and errors stemming from the OCR process detected and corrected. The output had to conform to the UNIMARC format with optional other output formats.

The following elements were included in the project:

- 1 Analysis of formal features of catalogue cards from different libraries, with different cataloguing traditions and different rules for the representation of bibliographic and other information on the cards, in order to produce specifications for automatic formatting of machinereadable copies of the cards.
- 2 Analysis of typical errors produced by the scanning and OCR process, as well as errors inherent in the source, in order to produce specifications for automatic or semi-automatic error detection and correction. The analysis was to include analysis of possible links between the formatting process and the process of error detection.

- 3 Specification of an application to convert scanned catalogue cards in plain text-format into a bibliographic format specified by the user (primarily UNIMARC), suitable for exchange of bibliographic records as well as for import of records into a computer based library catalogue. The specification was to include a specification of a suitable user interface for an integrated package.
- 4 Production of a working prototype implementing the specifications, including user interface and draft user's manual, both in English. The application should be adaptable by the user to specific formatting needs and error conditions using integrated facilities. It should be easy to translate the user interface of the prototype into any EU language, also using integrated tools.

Some of the assumptions of the original project has turned out not to be valid. Of major importance the assumption that hardware and software for scanning and OCR would not be a concern of the project, has turned out to be entirely wrong. The developments in this area as well as developments in the general architecture and operative systems of personal computers has made it necessary to investigate the market again and to experiment with new systems that have become available during the run of the project.

Special problems were encountered with the Greek characters, because the range and complexity of the characters in the catalogue of the National Library of Greece were larger than expected. The use of a special character set, the "polytonic" character set, in some parts of the catalogue has made it necessary to invest in an extension of an existing Greek OCR package to handle this. The "polytonic" character set uses special diacritical marks not used in modern Greek writing, with up to three diacritical marks to one basic character. The need to solve this problem prior to converting the cards has made the path followed by the Greek partner somewhat different from the rest of the project.

The following is a summary of the main results of the project as described in detail in the underlying reports, especially Technical Reports No 1 to 4:

A. Hardware and Software for Scanning of Catalogue Cards

The main problems with handling catalogue cards in bulk are related to the automatic feeding of the cards, the existence of cards printed on both sides, and the variable quality of well used machinewritten cards.

Most feeders for standard scanners produced for office automation are not able to handle cardboard catalogue cards, mainly because of the thickness of cards, but also because of the small size (typically 7.5 x 12.5 cm). The cards normally have to be passed round a sharp bend in an automatic document feeder (ADF) in order to be placed on the glass plate of the scanner and then back again. This works fine with paper but not with cardboard.

The project has identified four scanners able to handle catalogue cards at acceptable speed: The Fujitsu 3096/3097 series, the Fujitsu M3099, the Kodak Image-link 900 and the Hybrid 4512.

The last two are by far the fastest but are more difficult to integrate in the overall process of scanning, OCR and formatting, as well as working with a lower resolution (200 dpi). The cards pass at high speed in front of a reading head using feeder mechanisms without any need for bending the cards.

All the Fujitsu scanners use a roller feeder that passes the cards straight between two rollers and then in front of one or two reading heads. (The 3096/97 models also have a flat bed part, which is not used when the ADF is active.) All models have some problems with skewness (not pulling the cards straight through the feeder), but are otherwise very reliable. They work at resolutions up to 400 DPI - which is acceptable - and may be controlled directly by the OCR software making for a smoother work flow in production. The M3099 is by far the fastest, it is also able to handle a larger amount of cards in the feeder, and can scan both sides of the cards at the same time (duplex scanning).

The speed of the scanning process does not depend entirely on the scanner itself, but also on the controller card (as well as the speed of the controlling PC). The Fujitsu scanners are able to work with two different video interfaces provided by a Kofax card or a XIONICS card. The choice of controller card has consequences for the choice of OCR (see below). The XIONICS card can be made TWAIN compatible, using optional drivers.

The variable quality of the printing on the cards creates problems with bulk processing, since it is often not possible to set brightness and contrast in a way that will not leave a significant number of cards difficult to process correctly in optical character recognition. The use of some sort of automatic brightness control is able to reduce, but not eliminate this problem.

In the project only the Fujitsu 3096 or 3097 scanners have been used extensively for experimental work, both with Kofax and XIONICS controllers. The 3097 is the fastest of the two. Tests have shown that it is possible to process about 400 cards an hour, including OCR (or about 2.500 cards per working day), which is deemed acceptable. The main part of this time is spent in the actual feeding and reading of the cards, so that expenditure on a faster scanner, like the M3099, will pay off in production time.

Statsbiblioteket has been using the KODAK Imagelink 900 on a production basis for an independent project on retroconversion. The Hybrid 4512 and the Fujitsu M3099 has only been demonstrated to the project.

A custom-build card feeder for the Fujitsu 3096 that was intended for use in the project, turned out not to be adaptable to the OCR packages actually used in the project. The faster speed provided by newer models of PC's, and the possibility of handling larger number of cards in models like the Fujitsu 3099, the Kodak 900 and the Hybrid 4512 makes it unreasonable to pursue this line of development further.

B. Optical Character Recognition

The market offers a variety of affordable OCR packages with many nice features that have been optimized for office automation. Most packages rely on sophisticated recognition algorithms enhanced with post-processing modules using dictionaries and language specific rules to reduce the inevitable errors in the character recognition. Many of these features are not very helpful in retroconversion of catalogues because of the special nature of older catalogue cards from multilanguage collections.

The main features to look for in an OCR package for catalogue conversion is speed and accuracy in recognition of individual characters, and the character set supported. The ability to work with the model of scanner and controller card used is of course also important. The OCR packages support different ranges of scanners, but there is a general tendency towards

support for TWAIN-compatible scanners, making this less of a problem.

The main problems with printed and machine-written catalogue cards stem from the number of different typefaces (fonts) used - most of them non-standard from the point of view of the modern office. All the OCR-packages investigated provided high performance with modern standard fonts, but coped with varying success with older types. A mixture of different fonts in the same batch of cards also makes it difficult to maintain the same standard of recognition with all cards, even with packages that may be trained to recognise new fonts.

The variations in printing quality also creates problems as noted above, since this makes it almost impossible to have settings at which all or most cards are recognized at optimal accuracy and consistency.

All the packages investigated support only an 8-bit character set, meaning no more than 256 different characters, of which only 100 - 114 represents letters. Some packages are limited to the characters of the 8-bit ASCII-character set, with the variations created by the Code Page system of DOS (e.g. 87 Latin letters + 15 Greek letters in Code Page 865 and 114 Latin letters in Code Page 850), or the ANSI-character set supported by Windows 3.1 (114 Latin letters).

One, Recognita Plus, is able to handle a very wide range of characters, Latin as well as Greek, in all 319 (200 Latin letters, including combinations with diacritics, 65 Greek characters, 10 digits and 44 other characters (punctuation marks etc.)). The characters have to be mapped to the 8-bit character set by suitable assignment of codes. The characters are shown correctly on screen within the package itself, and may be exported to other applications according to code values defined by the user.

Other packages, such as Perceive and CharacterEyes, may be trained to recognise any character, which will then have to be assigned more or less arbitrarily to a position in the 8-bit character set, making it very difficult to proofread the result. Perceive have the option, though, to use an OEM font supplied by the user for representation of the characters (code 128 - 256) on screen.

The project has encountered specific problems with non-Latin characters, especially the Greek character set, since most OCR packages cater for Latin characters only. The Greek national library needed not only a full set of Latin and Greek characters at the same time, but also a very large range of Greek characters

with up to three diacritical signs to one basic character in order to reflect the letters used in older parts of the catalogue, the so called "polytonic" character set. It has been necessary to invest in the extension of an existing Greek OCR package, Anagnostis by IdeaTech, to handle this character set.

Most OCR packages rely on postprocessing (using spell checker techniques with dictionaries and language specific information about the character set used and the character sequences normally allowed) in order to find and correct errors in the "raw" OCR. This is very efficient if all the cards are in the same language and the vocabulary is the same as in the dictionary used. But with catalogues of multilingual collections this will never be the case. And in monolingual catalogues the vocabulary will certainly be special. So these features of the OCR package will normally have to be disregarded.

With retroconversion of older catalogues in national, academic and special libraries one has to evaluate the OCR packages on the basis of the "raw" recognition power, character by character, and not on the accuracy reached when using also the supplied post-processing modules.

The project has investigated the following commercially available OCR packages: iBS! Gigaread, Calera Wordscan, Caere Omnipage, Xerox Textbridge, Ocron Perceive, Recognita Plus and Ligature CharacterEyes.

iBS! Gigaread was the original candidate, but it had to be discarded, since it was not possible to get a Windows-version. Wordscan and Omnipage are both well-recommended but not suitable when a large range of "international" characters is needed. The ability to be trained to recognize non-standard typefaces is limited, and its use lowers the overall accuracy significantly. Xerox TextBridge is limited in the range of characters supported and provides no training facilities.

Recognita Plus was selected for use in the project, because of the large range of "international" characters supported, including Greek characters, and for the possibility to export to a 16-bit UNICODE text format. The package may be trained to recognize new variations of the characters supported, but only within the built-in range. As mentioned above Recognita Plus does not support the full range of characters needed by the Greek National Library.

Ocron Perceive and Ligature CharacterEyes may be trained "from scratch" to recognize non-standard

typefaces, like the ones found in older typewriters, black-letters, Greek letters etc. But the packages are then limited to one typeface at a time, not a mixture of typefaces, and the characters may only be "mapped" against the built-in character set (the ANSI set for Windows). As mentioned above Perceive has the ability to display the trained characters using an OEM font provided by the user, which is a very useful feature for retroconversion.

The accuracy of the packages when used without post-processing modules seems to be at the same level for the investigated OCR programs, that is about 98% correct (about 2% of the characters of the source misrecognized or not recognized), when used with trained typefaces. With catalogue cards this means 4 to 6 errors per card on the average. With much worn cards or very uneven printing as well as multi-font cards the accuracy is obviously much less.

The conclusion is that none of the tested packages fullfil all the demands of retronversion of older multi-font, multilingual card catalogues. The usefulness of each will have to be evaluated on the basis of the demands of each library concerning the range of characters and the nature of the typefaces used.

Some older card catalogue are not at all suitable for conversion using OCR, especially cards with handwritten or stamped additions, cards worn by much use and cards printed with very worn down and uneven typewriter ribbons. At the present state of the art conversion of handwritten cards is not feasible.

C. Representation of Character Sets

The handling of the character sets present in the card catalogue is a crucial point in retroconversion. The range of Latin characters used in European languages is larger than the 256 positions supported by an 8-bit character set, when all variations including diacritical marks, such as accents, are taken into account. The presence of Greek or Cyrillic characters extends the demands further.

It is difficult to estimate exactly the number of different characters that a retroconversion system will have to support. One list includes 144 different small Latin letters (including combinations with diacritics) and 120 different capital Latin letters, in all 264 different Latin letters used in languages with a Latin alphabet. Appendix 2 - 4 of the FACIT Technical Report no 1 lists 210 Latin letters that seem to be needed for European languages with Latin

alphabets, 84 Greek letters and 66 other characters (punctuation marks, typographical signs etc.). The range actually found in the catalogue will depend on such things as the accuracy in rendering the spelling of the original, translitteration practices and the range of characters that the typewriter was able to produce.

A main point in retroconversion of card catalogues should be not to lose any crucial information in the cards, and the use of specific characters, such as the combination of a basic character with one or more diacritical marks, represents such crucial information. This can only be represented accurately in a 16-bit character set, such as UNICODE (a subset of ISO 10646-1).

The present generation of PC's, using DOS or Windows 3.1, support an 8-bit character set only, and Windows 95 has only limited support for 16-bit characters. The problems of the larger range of characters used in European languages is "solved" by the Code Page technique in which the character codes 128 - 256 represent different selections of characters (so called "international" characters) according to the Code Page used. Code Page 850 (International Character Set) provides 114 Latin letters, while Code Page 865 (Nordic) provides 87 Latin letters and 15 Greek. The ANSI character set used in Windows 3.1 has 114 Latin letters.

Even when the number of different letters fit into the 8-bit coding scheme, the selection of characters may not fit nicely into the selections provided by the different Code Pages.

In order to provide a unique identification of characters across Code Pages and across different hardware and software platforms a 16-bit character set, such as UNICODE, is needed. But UNICODE is at present only supported by Windows NT, and only in a limited way. Very few fonts in UNICODE are provided with Windows NT and none of the ready made applications useful for retroconversion, such as scanning software, OCR, text-editors etc. are fully Windows NT compatible at the moment.

The version of Windows planned to supersede Windows 95 is expected to provide full support for UNICODE and fully UNICODE-based applications can be expected within the next few years, making it uneconomic at the present time to develop special applications, e.g. for on-screen editing of texts with 16-bit characters, showing the right graphical representation of each character.

The FACIT-prototype represents a compromise in this respect. It provides an internal subset of a 16-bit character set, fully customizable and with facilities for converting both input and output to and from this character set. This means that the identity of a specific character in the input set can be established and preserved in the processing of the records. The character may then be mapped to the character set of the exchange format or the OPAC. But it is not possible to display the characters correctly on screen (using the correct "glyph"), nor has it been possible to provide editing facilities with access to the records while in the internal format.

As said above the solution to this problem ought to become available soon as a ready made editing tool able to work with UNICODE files.

D. Analysis of Catalogue Cards

The project has developed a methodology for describing catalogue cards for the purpose of retroconversion using OCR and automatic formatting. This description is a central - and normally very time consuming - part of the overall process.

The information contained in the cards and the way it is represented using lay out, punctuation, specific words (= strings of characters) etc. has to be described in such a way that a computer may use it in a formal analysis of the card.

This is actually a formal reconstruction of the rules applied when the card was produced. But this may and may not be the same as the cataloguing rules set down by the library. The analysis has to reflect the actual practice of the cataloguers, not the rules they were supposed to apply.

The analysis is used partly to decide whether it is possible to convert the catalogue using a tool like the FACIT prototype, partly to prepare for such conversion.

The formal description aimed at is the same as the one used in the control files for formatting the input (cf. next section).

E. Formatting of Input

The project has shown that the formal specifications needed for formatting ("parsing") the input may be written using a very simple language providing great flexibility. The technique used is a simplified ver-

sion of the one used for syntactical analysis of natural languages or for the writing of compilers for computer programs.

This is possible because most catalogue cards have a very flat structure with a fixed sequence of fields. Repeating fields are found but not nested fields. With ISBD cards the fields are marked by special punctuation making it very easy to write a formal parser, but with older cards the effort to identify useful delimiters is much bigger.

Dictionaries may be used to identify words or other character sequences that are useful for the identification of fields, but it is not possible to work with a closed vocabulary for syntactical analysis as is the case with natural languages.

The actual formatting is performed by two programs written specifically for the FACIT application. They will interpret the formal specifications provided by the user and transfer the information found in fields on the cards to fields (in the internal card format) named and defined by the user.

The core programs make very few assumptions about the input, making it possible to handle a large variety of cataloguing conventions. But the cost of this flexibility is that the effort needed to set up and test the necessary control files is big and the task can only be carried out by specially qualified staff. Still the effort is - in most cases - economic compared to writing a suitable formatting program from scratch.

The variability found in real life catalogues makes it difficult to reuse control files when working with a new catalogue, but an experienced systems analyst will of course be able to apply principles and solutions derived from other catalogues, thus cutting down the time needed to set up a new system. Dictionaries associated with control files may be reused to a certain extent.

Writing the control files and debugging them is a specialist job, combining knowledge of cataloguing and of computing. Using the application after the system has been set up only requires skills in routine work with computers.

F. Error Detection and Correction

The project has applied a general methodology for analysing errors in OCR and application of the analysis in error detection and correction.

The methodology is based on the statistical analysis of misrecognitions and nonrecognition of characters in the source, identifying special "error prone" characters.

Some errors produced in OCR are highly predictable and they may be used when checking the result against a dictionary, as in speller checker routines, while others are more diffuse (stochastic).

The main sources of errors are:

- * The shape of characters in general or in a specific typeface (font) may make it difficult for the OCR package to distinguish the characters. Often encountered examples are "1"(one), "l" (el) and "I" (capital i); "c", "C" and "e"; "5" and "S".
- * Broken characters due to bad printing or wrong setting of brightness control or contrast in scanning. Often encountered examples are "rn" or "in" for "m"; "lc" for "k"; "I." for "L".
- * "Ligatures" or characters merging due to bad printing or wrong setting of brightness control or contrast in scanning. The result will normally be an unrecognized character marked according to the set up of the OCR package used.
- * Inserting or deleting spaces. Especially with proportional printing the OCR may have problems with correctly identifying spaces, but an OCR package set to handle proportional printing may also have problems with fixed distance printing, such as older typewriting.

The accuracy of the recognition is strongly influenced by the following:

- * The number of different typefaces (fonts) in the source. Some OCR package may only work with one font at a time, others can handle a range of modern standard typefaces, but may be baffled by older printing or typewriting.
- * The number of different characters to recognize combined with the number of different characters in the character set of the OCR. Accuracy may be heightened by making the two fit closely, so that the OCR package does not have to take characters into account not found the source. (This is only possible with some OCR, such as Recognita Plus, when more than one language or a nonstandard mixture of characters is used.)

- * The evenness of the printing, making it difficult to handle cards written on typewriters with worn down ribbons, and with varying printing methods.
- * The state of the cards: yellowing, smudged, handwritten additions, stamps etc. lowers the accuracy significantly or may prevent the use of OCR entirely.

Routines for error detection and correction will have to be adapted to the error conditions of the actual set up of source (catalogue cards), scanner, OCR and character representation in the plain text input file, in order to be as effective as possible. This means that the application will have to be presented with the result of the error analysis in a suitable way.

The postprocessing facilities provided with most OCR packages today have very limited usefulness in retro-conversion of multilingual card catalogues, because they are relying on monolingual dictionaries and other types of linguistically based analysis.

Monolingual and special purpose dictionaries may be used in error detection and correction, but only after formatting and after the language used e.g. in the title field has been identified. The project has sketched some possible procedures for identifying the language of the title field, but it has not been possible to test them fully within the scope of the project.

The number of languages found in many catalogues prohibits the use of large standard dictionaries (with 200. - 300.000 wordforms). It is normal to find 10 - 15 different languages in the catalogue of an academic library or a national library. The possibility of applying very compressed special purpose dictionaries has been investigated during the project, but it has not been possible to implement this in the prototype.

The prototype includes facilities for systematic substitution of characters and strings of characters occurring in certain surroundings. It also includes routines for dictionary look up, using small dictionaries provided by the user as needed for formatting.

Facilities for using statistical information about errors (= character substitutions) produced during OCR to identify possible matches in dictionary look up are planned but have not been implemented yet.

G. The FACIT prototype

The project has produced a working prototype demonstrating some of the principles arrived at in the project.

The prototype consists of a shell, providing the user interface, based on Microsoft Access 2.0 for Windows, and a series of underlying, integrated DOS-programs handling the actual conversions. It is possible to use the DOS-programs independently of the shell, allowing the application to work at higher speed.

The main input to the prototype is a plain 8-bit text file with copies of the catalogue cards separated by one or more characters assigned the role of card separators. This input file will typically be the result of a scanning and OCR process, but may also be the result of manual keying in the text of the cards.

The user will also have to provide a series of control files that govern the conversion process: Conversion from 8-bit to 16-bit characters, conversion into the internal format, general substitution of strings, e.g. systematic errors, rules for formatting the input files into bibliographic fields, dictionaries for recognition of field specific expressions, conversion from the internal format to target format, such as UNIMARC, and conversion of the internal 16-bit character set into an external 8-bit or 16-bit character set.

The working of the FACIT Prototype is described in more detail later in the report as well as in the Technical Reports No 3 and 4.

The main output of the application is a file containing the formatted catalogue records in a tagged text format or a tape format conforming to ISO 2709 and the UNIMARC specifications.

The application will run on a standard PC with Windows 3.1. A PC with at least 486 DX2, 66MHZ, 16 MB RAM and a large hard disk (400 - 600 MB) is recommended, as well as a high resolution colour monitor (SVGA 1024 x 768, 16 colours, 16").

The application is able to process about 200 cards per hour or about 1200 cards per working day using the equipment described above. The actual speed will of course depend on the equipment used for future projects. The Prototype was able to recognize about 95% of the cards correctly with input files that had been carefully checked and corrected for OCR errors.

The prototype does not include facilities for editing records in the internal format nor the ability to show the internal 16-bit characters on screen with the right graphical representation (cf. section C above). These facilities are planned to be provided by standard (off the shelf) tools, when such tools supporting UNICODE files becomes available.

The prototype does not include the planned modules for semiautomatic error detection and correction nor the ability to edit errors in the source with direct reference to an image of the original card. These facilities will have to be included in further developments of the prototype outside the scope of the present project.

The production of UNIMARC output includes information not found in the original source, such as type of material, status of cataloguing and language of publication. This has to be provided manually after inspection of the records or supplied as default information. Facilities are provided for this.

A Danish project has investigated the possibility of automatically deducting the language of the publication from the string of characters and words in the title field using Neural Network techniques. It seems possible to reach a very high degree of accuracy, about 95% correct guesses, but further work has to be done before this technique can be applied in actual production work (Hørning, 1995).

H. General conclusions

The project has demonstrated that it is possible to create an application that may be adapted to the variety of different cataloguing formats found in older card catalogues. This implies detailed formal analysis of the catalogue and the writing of formal specifications ("control files") which is a specialist job requiring the combination of knowledge about the catalogue and about computing.

As has also been shown in other projects the work load is shifted from routine keying in of the data from the cards to preparation of the conversion process and setting up the system. The economics of this depends partly on the complexity of the catalogue, partly on the number of cards to be processed.

The process investigated does not seem suitable for in-house work in smaller libraries (probably less than 20.000 cards), but should be feasible for larger libraries and for companies providing retroconversion services for all types of libraries. The actual

economics of such a project will have to be established empirically as described in the section on the calculation of costs later in this report.

The weakest part of the overall process has turned out to be the accuracy provided by the OCR process, and the need for careful proofreading of the result. Some support may be expected from the computer in error detection and correction, but the nature of the material to be converted in a multi-font, multi-language catalogue does not make the prospects of fully automated error detection and correction very promising.

The direction of further development should rather be towards easing the proofreading process, using on screen representation of the original card and the converted text (in UNICODE characters), and using the computer to identify records with possible errors. In some case a speller checker program, based on user provided, special dictionaries, and using information about typical errors, also provided by the user, may be used to enhance the proof reading proces.

Further work is needed on the prototype in order to provide an application suitable for large scale production work.

In spite of this, the project has provided the partners with a basic tool for retroconversion, as well as valuable insights into the nature of retroconversion using OCR. And it has provided a greater awareness in other libraries in the countries involved about the problems involved and the tools available for retroconversion.

Background of the Project

A. General background

The success of library automation, creating user-friendly on-line catalogues integrated with circulation-systems and other administrative facilities, has created an urgent need for retroconversion of the older parts of the catalogues.

As the users get used to the new catalogue medium, the holdings not registered in machinereadable form become so to speak "invisible" to the users. As a typical case of this Statsbiblioteket in Aarhus, Denmark, found 33 requests in the OPAC (a queue almost three years long) for a standard work on the fashionable subject of fractals, while an older, and just as useful edition of the same work, which could only be found in the card catalogue, was not requested once.

Apart from such evident waste of the resources invested in the collections of the library, the difference in the cost of handling loans and other routine administrative matters using the automated system, and the cost of the equivalent manual routines, makes it imperative to get machinereadable records for all works in the library likely to be searched for and requested by the users, in order to make efficient use of the investment in the new technology.

This has resulted in a search for cost-effective means for converting the old catalogues into machine-readable form, a search that has been going on ever since the libraries became convinced that the new catalogue medium was the medium of the foreseeable future.

Several methods have been developed, representing four families of methods that differ in principle:

1. Cataloguing the collections from scratch, using modern standards of cataloguing and the input tools of the automated system, thus ensuring the same format and the same level of information as the existing OPAC of the library. (Retrocataloguing)

2. Copying the information from the card catalogue and other registers using the input tools of the automated system, but without adding information not found in the old catalogues, and with only minimal changes of the information in order to comply with modern standards. (Retroconversion by direct keying)
3. Copying the information from the card catalogue and other registers using optical scanners and OCR (Optical Character Reading), and automatically tagging the resulting machinereadable records before the final conversion into the target format. (Retroconversion by OCR and automatic formatting)
4. Substituting the old records with preexisting machinereadable records of the same works, usually provided by a conversion agency or the producers of national bibliographies. (Retroconversion by substitution)

The method to be used in each individual case depends on several factors, among others available funds, the state of the catalogue to be converted, the availability of machinereadable records matching the collection, and the quality of cataloguing aimed at.

If the aim is to create records complying with the standards of today and fitting seamlessly with the rest of the records in the OPAC, retrocataloguing is the only viable solution, specially with older collections consisting predominantly of works in languages other than English. But the cost and the time scale of such a "manual" process may be more than most libraries can consider seriously.

The recommendation of the Council of Europe Guidelines for retroconversion (Council of Europe, 1989) is for the libraries of Europe to cooperate in creating a pool of machinereadable records from national bibliographies and catalogues of important collections, so that the retroconversion of the large of majority libraries can be made using the substitution method. In practical terms this means that some libraries or national agencies will have to do the necessary retrocataloguing of their national collections, while everybody else will have to wait for the arrival of a significant number of records relevant to their needs, in order to attain a cost-effective conversion of their foreign collections.

As the ideal solution may be either prohibitively expensive or too late, most libraries will have to search for an optimal solution within the limitations imposed. The FACIT project (Fast Automatic Conversion

with Integrated Tools) explores the possibilities of fast and relatively cheap mass conversion of typed or printed card catalogues using OCR.

The use of OCR in retroconversion has been investigated earlier, but with disappointing results. Typically the handbook on retroconversion by Beaumont & Cox (1989) dismissed the possibility of using OCR with the following remark: "There has been some efforts made to use the information directly from shelf-list cards using optical character recognition (OCR) systems. On the whole, such efforts have not been very successful because the quality of shelf-list cards has varied so much over the years and the information may have to be coded for fields in the record. It is not a viable option for small or medium-sized libraries [...]."

The remarks on OCR are made in the context of using OCR to get the key information necessary for retroconversion by substitution, using an outside database as the source of substitution. Using OCR for copying all the information from the main catalogue as an alternative to direct keying of the information is not even considered.

The general dismissal of the OCR technique was based on practical tests (Avram, 1972; Smith & Merali, 1985). The disappointment with the results of these were partly due to unrealistic expectations based on an incomplete understanding of the nature of the technology, partly due to limitations in the equipment available at the time, and in the computer based routines used to handle the conversion. The arrival of cheaper, faster and more reliable hardware and software for OCR as a result of developments in office automation has made it reasonable again to investigate the possibilities for retroconversion. Within the last few years several small and a few large projects to that end has emerged, trying out different approaches.

In most cases the tests carried out by libraries have consisted in using a cheap commercial OCR package to make a copy of the card or other source in machinereadable form, then using a text-editor or wordprocessor to correct any errors and to provide tags and other alterations in order to produce an input acceptable to the target database-system.

After some initial exhilaration over the fact that the OCR makes almost as few errors as a typist but is very much faster, the realization surfaces that the time spent in finding and correcting the inevitable errors and editing the result to conform to the proper input-format might just as usefully be spend

keying the card directly into the desired format. Without the appropriate post-processing software OCR is not cost-effective for retroconversion, and the post-processing package (error detection/correction and formatting) is not available as a standard, low-cost commercial package, since it is neither necessary nor useful in the office environment.

Some of the most successful commercial OCR packages used for office automation nowadays come with some form of postprocessing programs for enhancing the character recognition and for error correction. They are usually based of some sort of linguistic information and use dictionaries in the same way as a spell checker for a wordprocessor. But since they function best with a source containing only text in one language, typically English, a lot of the strength gained by the post-processing turns into weaknesses when confronted with the catalogue of a typical foreign language collection. And even with monolingual collections the vocabulary of the cards is very much different from the one in the supplied dictionary.

A few projects have concentrated on the processes that follow the OCR part. Work done by Martin Harrison in the early eighties investigated the possibility of automatic formatting into the MARC format, starting with the machinereadable copy of the catalogue card produced by an OCR-system. The formatting program used the parsing methods of computational linguistics to analyse the information in the cards, including the clues provided by the typographical lay-out (Harrison, 1984). The results of the project were never implemented in a production system for reasons not known. It seems they had to do with problems in the OCR part of the project, rather than the formatting programs.

Later projects too have relied on syntactical analysis of one form or another for the formatting part. A small feasibility study carried out in Canada in 1989 used parsing tools developed for programming languages and input-systems (LEX and YACC) for automatic formatting of copies of records in a printed bibliography (Crawford & Lee, 1990). A project commissioned by Deutsches Bibliotheksinstitut to Compulex Biblioscan in Zürich (Shah, 1992; Deutsches Bibliotheksinstitut, 1993) used the techniques of computational linguistics for both error correction/detection and formatting. And a project carried out at the Universidad Complutense on Madrid in 1991-92 also used techniques of computational linguistics in the form of logic-programming (PROLOG). This has resulted in the the program LAURA produced by the Verba Logica group at the Logic Department of the Universidad Com-

plutense, and used for retroconversion at the university library and the library of the Scientific Research Council (Repiso & Ríos, 1994).

MORE, a project supported by the CEC under the Telematics programme, has demonstrated that these techniques can be used to convert a printed bibliography, the national bibliography of Belgium, into formatted machinereadable records. The project also explores the use of SGML (Standard Generalised Markup Language) to produce intermediary structures in the formatting proces.

The forerunner in Denmark of the FACIT project used the same approach for the formatting part. And a working system for formatting the output of an OCR-system was made by the Royal Library of Copenhagen in 1991. This system is now used for production and more than 400.000 records has been converted with the system, which was presented to the library world at a conference in Copenhagen in december 1992. The system uses custom build software generated with the help of generally available tools like LEX and YACC (Boserup & Holtse, 1992). The library has by now converted a significant part of its newer card catalogue of the foreign collection using this system.

Alternative approaches have also been tried. A small scale feasibility study was carried out in Copenhagen by a Danish company in collaboration with a university departemental library, using the techniques of neural networks for formating the records, but this was not very successful. Still the neural network approach may be useful for certain aspects of the conversion process as shown by Hørning (1995).

B. Rationale of the project

In connection with large scale conversion of catalogues it is essential to get a very high throughput with a minimum of labour. The time scale of such a process is in itself important in order to make the older parts of the collections accessible through library automation, including the OPAC (Online Public Access Catalogue) and the circulation system, within a reasonable number of years. The overall cost of the conversion proces is also a critical factor, since the usually considerable cost of retroconversion has to compete with other claims for funding within the individual libraries and within society at large.

More expensive scanners are able to maintain a much bigger throughput just producing a digital image of the cards, but then the OCR process slows down the

overall throughput rate for this first part of the conversion.

If scanning and OCR becomes fast and fairly reliable the bottlenecks in the conversion process move from the input stage (scanning and OCR) to the stages of detection and correction of the residual of errors (OCR errors and source errors) and tagging the elements of information to go into the fields and sub-fields of the target bibliographic format.

These processes should be handled by a set of tools integrated with the OCR package and be automatic as far as possible in order to match the speed of the OCR part of the conversion.

The overall effort of setting up the system to convert a given catalogue and the actual conversion of the catalogue will have to compare favourably with direct keying from the source into the target bibliographic format by a skilled operator with the bibliographic knowledge to make the necessary decisions about the bibliographic content of the source.

The preliminary investigations by the partners and others seemed to show that such an automatic process was in principle possible with a series of known techniques, but that no integrated package existed that could be set up easily by the user to convert any new catalogue.

The best results so far had been obtained with monolingual catalogues using the ISBD conventions. This is so, because if all the works in the catalogue are in one language - which is the same as the language of the cataloguing agency - the techniques of for instance speller checkers with dictionaries supplemented by linguistic analysis of character sequences can be used easily for error detection/correction, and this process may be integrated with the OCR process, as is often the case with so-called Intelligent Character Recognition (ICR). And the ISBD punctuation makes it easy to program a machine to identify the bibliographic elements.

The present project - on the other hand - was aimed at handling the multilingual catalogues that are typically found in large academic libraries including the foreign language collections of national libraries. Since a large part of the catalogues to be converted were made before the advent of ISBD, it is not possible to rely on the ISBD punctuation etc. for identification of bibliographic elements. This means that the methodological approach has to be more general both in the field of error detection/correction and in the field of formatting, than in existing

techniques used for ISBD-based monolingual catalogues.

C. Wider framework of the project

The FACIT project is focusing on one aspect only of the total process of conversion by OCR. In order to optimize the whole process other aspects will have to be investigated and suitable methods found or developed.

The application produced by the FACIT project takes as input a plain textfile containing records that are simple copies of the source, retaining the typographical lay-out. This input may be produced by OCR software and hardware, where suitable, but may also be produced by having a skilled typist copy typewritten cards of poor quality or with many handwritten additions as well as hand-written cards and registers.

Whether it is cost-effective to use the application in connection with direct keying compared to having staff with bibliographic skills key the information directly into the target database, is a matter for investigation. The answer will depend very much on the speed and reliability of the application, and the cost of using it.

The project relies on existing systems for scanning and OCR, and to undertake developments in this area was not meant to be a part of the project. This does not mean that there is no room for more work aimed at optimizing scanning and OCR for retroconversion - on the contrary. At present the process of handling catalogue cards in large volumes, the level of accuracy attainable in the OCR process itself with older materials, and the specific problems with character sets and vocabulary all leave ample room for more optimal solutions.

It will also be necessary to focus on the process of merging converted records with the already existing machinereadable catalogues, in order to avoid duplication of records. This should also be done automatically, with as little intervention by humans as possible, in order to make the overall process cost-effective.

A good merging package that makes it possible to identify duplicate records of different quality, and to substitute high quality records for low quality records without losing information like shelf marks, subject indexing, class marks, location etc. would make it possible to combine conversion with OCR and

conversion with substitution (possibly at a later time) to get the best of the two methodologies.

Workplan of the project

The project was divided into four phases:

1. The Analytical Phase, to result in two reports: Analysis of card features, and analysis of errors in the OCR process. The methods for these analyses had to be established up as part of the project, but some solid ground work has been done by amongst others Statsbiblioteket, and this together with the experiences of Det Kongelige Bibliotek and SYNERGI has been the main basis of the methodology used.
2. The Specification Phase, to result in specifications of error detection/correction and formatting routines and of the user interface. The core routines were specified using prototyping techniques, as well as formal specifications of the features of the cards etc. using Backus-Naur notation and notation of so-called regular expressions for lexical analysis. This phase also included detailed plans for Phase 3, including software development strategy and software architecture design.
3. The Production Phase, resulting in the working prototype and plans for evaluation of the prototype as well as the overall process of conversion.
4. The Evaluation Phase, resulting in an evaluation of the prototype based on the conversion of a sample of cards from each library, revision of the prototype and the draft manual and production of the final report.

Overview of the Workplan

In order to reach the objectives the project was divided into 15 Workpackages (WP1 - WP15) with 3 Milestones (M1 - M3) as shown in the following outline:

- WP1: Project Management
- WP2: Information about alternatives to the OCR-scanning platform.

- WP3: Analysis of catalogue cards: Identification of data elements and formal features
- WP4: Installment of equipment and training of technical staff (including preliminary tests)
- WP5: Conversion of 1st sample. Selection of 2500 cards from each library. Conversion into simple ASCII file. Advanced trimming of OCR equipment.
- WP6: Analysis of errors in OCR-conversion of cards: Identification of possible methods for detection and correction.
- M1: Reports on the results of the analytical phase (WP3 and WP6).
- WP7: Specifications of Prototype, including User Interface
- WP8: Programming of User Interface and General Facilities of Prototype
- WP9: Programming of Automatic Formatting Procedures
- WP10: Programming of Error Detection and Correction Procedures
- M2: Prototype, ver 1.0 (WP8 - WP10).
- WP11: Establishing methods of assesment and evaluation of economy and effectiveness of prototype
- WP12: Test of prototype
- WP13: Evaluation and finalization of prototype
- WP14: Dissemination of results.
- WP15: Final Report
- M3: Release of Final Report and Prototype ver. 1.1 (WP13 and WP15)

Some adjustments of the working plan has been necessary during the run of the project, but the main outline has remained the one shown above.

Organization of the work

The organization of work reflected the dual objectives of producing a working prototype adaptable to a broad spectrum of libraries and with a solid empirical foundation, and giving the participating libraries a thorough insight into the combined bibliographic-technological problems of retroconversion of card catalogues and the functions of the prototype produced.

A Joint Committee with one representative from each library and the coordinating partner, has had - apart from its role in Project Management - a function as the main forum for discussion of the methods and procedures of the project. A shared framework for the execution of the tasks to be carried out by each library in the individual workpackages was to be agreed in this forum. A series of meetings was scheduled for these purposes in the relevant workpackages.

The analysis of cards and of errors was carried out in each library as part of a close collaboration to produce generalized descriptions and methods of analysis.

The work on selection of samples, conversion of cards and testing of the prototype was carried out by each library, and both the librarians and the technical experts from the libraries were intended to acquire a good understanding of the processes and problems.

Methods of assessment of costs and evaluation of efficiency of retroconversion was to be developed under the leadership of one partner experienced in the use of such methods, but all participants had to acquire some knowledge of the background and implications of the methods proposed.

The work on specification and development of the prototype involved mainly the project manager and the staff at SYNERGI (and to some extent subcontractors in Italy and Greece), but again the librarians and technical experts of the libraries had to be involved, partly in order to participate in decisions relevant to the catalogue of each library, partly to get insight into the working of the prototype and the different software modules constituting the application.

The sort of insight outlined above was also seen as an important prerequisite for the librarians when involved in disseminating the results of the project.

Deviations from the original planning

The project was delayed from the very beginning because funding for the necessary equipment was not available in Greece and Italy right away. This was circumvented by having scanning and OCR made in Denmark on existing equipment in stead.

This proces revealed problems with the Greek character set that had to be dealt with in a special way, forcing the Greek partner to follow a special course, ending in the production of a special OCR application for the extended Greek character set (the "polytonic" character set).

In the mean time the other processes progressed as scheduled but slightly late. The card analysis was carried out in the Italian libraries and in the Danish library and the specifications for the prototype produced according to the original plan.

The analysis of errors carried out at the library in Firenze showed that the level of errors to be expected from the older parts of the catalogues was much larger than originally foreseen. And later it turned out that the flood in Firenze had made it almost impossible to feed the cards automatically in large numbers.

These problems with scanning and OCR delayed the testing of the prototype, while certain problems in the programming phase contributed to further delays.

Formal specifications were produced for the Danish and Italian libraries, but because of the problems with scanning and OCR they could not be tested on a large sample as originally foreseen.

Firenze produced a sample of images of cards using mainly manual feeding. This sample was used also in a special test of an online catalogues based on images with limited indexing on authors names and titles.

Napoli concentrated on deleveloping strategies for incorporating dictionaries for special purposes as well as providing the vocabularies for this.

Statsbiblioteket was all the time active in developing algorithms for formatting etc. as well as working on the problems of errors detection and economics.

The problems with scanning and OCR and the effort used in trying to solve these problems account for the main part of the deviations and som of the delays, while a certain division of labour, especially

between the two Italian libraries, account for some deviations from the original parallel work.

Apart from this all the partners, including the coordinating partner, have had problems with the actual allocation of key personnel to the project, since these persons were also involved in other processes vital to the institution. This has lead to an overall tendency to delay the project and the reporting.

Hardware and Software Platform

A. General Specification of the Platform

The hardware and software used in the project can logically be divided into two parts: 1) the Input System (scanner and OCR) and 2) the Main Processing System (formatting and error detection/correction).

The Input System converts a series of typewritten or printed catalogue cards into a plain text file ("ASCII file") containing separated records as character-by-character copies of the original cards. Each character in the cards is represented by one or more characters in the 8-bit character set, according to a documented standard.

The Input System has to be fast, measured as the average number of cards processed per hour over a period of six hours. The quality of the conversion should be high, measured as the number of not recognized or misread characters per 100 characters on a sample of a 100 cards.

The Main Processing System converts the plain text file produced by the Input System into an intermediate file. In the FACIT prototype a database allowing 16-bit characters is used. During the conversion or through further processing of the intermediate file bibliographical formatting, error detection and error correction takes place. From the intermediate file an output file in the desired format (such as UNIMARC) is made by a separate conversion module.

The formatting etc. carried out by the Main Processing System ought as far as possible to match the speed of the Input System, again measured as the average number of records converted per hour over a period of six hours.

The interconnection between the two systems was to be constructed in such a way as to minimise the overall time used in converting a sample of catalogue cards into a UNIMARC file.

The Input System and the Main Processing System was required to have a clear interface so that the specific hardware and software used in the Input System could be substituted with others without jeopardizing

the whole proces. This was done by only requiring the Input System to produce a plain text file as specified.

B. The Input System used

In the project the following configuration has been used for the Input System:

- 1 standard 386 or 486 PC, with a 14" or 16" SVGA colour screen, DOS 3.30 or higher (preferably DOS 6.2), a mouse, hard disk(s) with a capacity of 100 to 600 MB, and at least 8 MB of extended RAM.
- 1 Fujitsu M3096 or M3097 flat bed scanner (A3 format, 400 dpi), with an automatic document feeder (ADF) capable of holding 50 sheets of paper or 20 catalogue cards. The choice of the scanner version depends on the interface with the computer, and that again on the interface cards supported by the OCR package.

The Fujitsu scanner is the only flat bed scanner that is able to feed catalogue cards automatically, because of the extra roller feeder and reading head at one end of the scanner.

- 1 OCR package with good performance in "raw" character recognition - without enhancement with linguistic rules and dictionaries, and able to handle the range of different characters (including diacritics) found in the source. The package has to be Windows based in order to work well with the Prototype Application, but most commercial OCR packages are today.

Several OCR packages have been investigated and they have various strengths and weaknesses from the point of view of retroconversion of catalogues. At the moment the Hungarian Recongnita Plus (Windows version) seems to offer a reasonable solution, specially with regard to the range of character sets and the range of scanners supported.

The equipment used in the project for scanning and OCR was chosen for the combination of speed and cost that it offers. But the project aims at independence of any one specific input system so that new and better hardware and software for OCR can be substituted when they become available.

C. Alternatives to the proposed Input System

The Input System was originally selected because earlier tests had shown it to be fast and reliable as well as within the price bracket reasonable for a project of this type. But other combinations of hardware and software may be used to produce the same type of ASCII-copy of the source.

As part of the project information about alternatives has been collected and checked to provide a wider picture of the options available to the libraries. A special study of the Kurzweil systems is reported in a separate report: Using Kurzweil scanning systems as platform for cards scanning (FACIT Technical Report No 1.1). Other systems are mentioned in Optical Character Recognition for Retroconversion of Catalogue Cards: Hardware and Software. (FACIT Technical Report No 1).

The digital images of the cards may themselves be used as a microform copy of the card catalogue, either for archival purposes, or with minimal indexing as a substitute for the card catalogue. This is also be an interesting alternative for catalogues with hand-written entries or additions, since these are not suitable for conversion using OCR.

This may be produced using the same hardware and software for scanning and production of digital images. A special study of this was carried out by BNCF in connection with the test of the FACIT Prototype and the resulting prototype was made available on the World Wide Web, showing the feasibility of such an approach.

D. Equipment for the Main Processing system

After the initial scanning and OCR, which demands special equipment using a powerful PC, the next steps of the process may be executed using standard 386 or 486 PC's with a good SVGA colour screen.

In principle the same PC may be used as for the Input System, making it easy to integrate the use of data-files. This will also be suitable for a workprocess in which the cards are scanned, processed by OCR and formatted etc in smaller batches (20 cards per batch or 100 cards per batch).

Alternatively one fast Input System may feed several Main Processing Systems working in parallel to match the speed of the Input System. This creates the need for suitable transport of data from the Input System

to the other workstations, but this has not been investigated in this project.

The FACIT Prototype

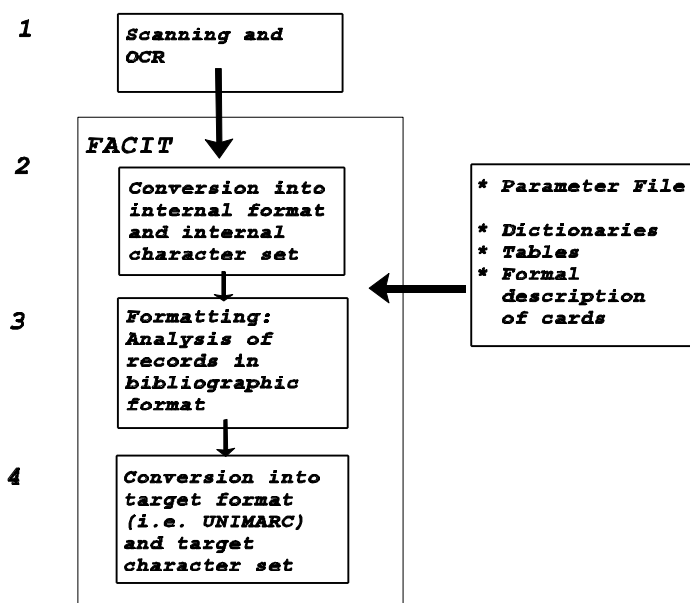
A. Introduction

The FACIT prototype consists of a series of programs running under DOS with a shell running under Windows on an IBM-compatible Personal Computer (PC).

The programs handle post-processing of the output from an OCR program in order to format bibliographic records and to detect and correct spelling errors in the records. The programs also handle storage and retrieval of intermediary files, editing of files and updating of dictionaries, tables etc. for use in the formatting and error detection processes.

Windows is selected as the operating system mainly for two reasons: 1) Most commercially available OCR packages for the IBM-compatible PCs nowadays run under Windows, making it easier to establish a good work-flow if the FACIT application runs under Windows too. 2) The graphic interface of Windows may be utilized to handle a user defined character set for the retroconversion, and to handle the images of the original cards in a fairly simple way.

The overall process is divided into four phases:



B. Scanning and OCR

Scanning and Optical Character Recognition (OCR) is not part of the FACIT application as such, but provided by third party hardware and software. Apart from the ability to handle basic needs such as scanning cardboard cards in large numbers and recognizing all characters occurring in the source with a low error rate, the OCR part will have to be able to provide input to the FACIT programs in the following form:

- 1 Bibliographic records in one or more text files ("ASCII text files"), consisting of printable characters and a few control characters: NL (New Line), FF (Form Feed) and EOF (End of File).

The FACIT application is able to accept 7-bit, 8-bit and 16-bit characters ("UNICODE") as input. Most OCR packages will be able to produce 8-bit characters, making it possible to represent up to about 250 printable characters simultaneously. The actual number may be less, since the OCR package may only allow letters, punctuation marks and other typographic characters from the ANSI character set or the Code Page system.

But to represent the Latin characters (including diacritics) occurring in the European languages, as well as Greek characters and perhaps Cyrillic, one needs more than 250 characters. Recognita Plus 2.0A is able to differentiate between 200 Latin characters, 65 Greek characters, 10 digits and 44 other characters (punctuation marks etc.), in all 319 characters (but of course not simultaneously).

Until the OCR packages are able to handle 16-bit characters, the characters in the source have to be "mapped" to the 8-bit character set of the input file, perhaps using character combinations to represent one character in the source (if the OCR package allows this). This will be translated into a true 16-bit character set used internally by the FACIT programs.

One character, like "@" will have to be used to represent an unrecognized character. Form Feed as a separator between cards (pages) may be substituted by a unique sequence of characters like "\$\$", but this is not necessary.

- 2 A series of digitized images in uncompressed TIFF-format, one image for each card (page) in the text-file. The images will have to be named and numbered in such a way that it is easy to establish the necessary links between the individual images and the pages. This information will have

to be provided in the general parameter file (see below section H).

The user will also have to provide three tables with information about the character set in the source file(s), and how to interpret this during the FACIT processing.

1. A Conversion Table (in the form of a text file) showing the correspondance between the 7-, 8- or 16-bit character codes of the input file and the character codes of the internal 16-bit character set. The codes are given in hexadecimal numbers. The input character may also be designated as literals in quotes if unambiguous. The Conversion Table may be used to document the character representation of the input file in a manner independent of the character sets of printers and video displays by providing standardized descriptions (ISO 10646) after a "//" (double slash):

```
// Sample Conversion Table
65, 0065 // Latin small letter a with acute
E1, 03B1 // Greek small letter alpha
"â", 00E2 // Latin small letter a with circumflex
"^a", 00E2 // Latin small letter a with circumflex
.
.
.
```

This table is used when converting an input text file into records in the internal card format and in the internal 16-bit character set. Other information about these characters are provided by specific font files (see section D below).

2. A Transition Table (in the form of a text file) showing the relationship between the characters of the original source (the cards) and the characters of the input file, based on a statistical analysis of a representative sample, using the same scanner and OCR package with the same settings on the same type of cards. The characters are represented by the codes of the internal 16-bit character set, in hexadecimal form. Comments to provide interpretation for a human reader may be added after a "//" (double slash).

```
// Sample Transition Table
0061,0061,530 // Latin a -> Latin a, 530 times
0061,0065,56 // Latin a -> Latin e, 56 times
0061,0073,85 // Latin a -> Latin s, 85 times
.
.
.
```

This table is used in the verification and error detection/correction processes (see sections F and H).

3. A Substitution Table (in the form of a text file) giving general substitutions to be carried out while converting the source file into the internal format. The table consists of two columns separated by commas:

- Strings in the source file (may be specified as so-called regular expressions).
- Resulting string.

Comments for documentary purposes may be added after a "//" (double slash).

```
// Sample Substitution Table
"Edltor", "Editor"
"ae", "æ"
"I."[a-z]+, "L"[a-z]
    // "I." is changed to "L" when preceding a
    // string of lower case letters
.
.
.
```

Additionally the user will have to provide a formal description of the bibliographic structure and data elements of the cards, but this will be taken up below in section E (Formal Specifications).

C. Internal Record Format

During processing in the FACIT application all records are held in a custom-built object oriented database. The basic elements of the records are the record identifier, the text from the input file, one record per card, and a link to the image file of the card. This link makes it possible to display the image on the screen when the human operator needs this in order to visually verify or correct the output from the OCR. As the formatting proceeds the record is expanded to incorporate the various data elements found. The fields and sub-fields are also handled as objects that are established as needed, making the structure (field lengths, number of fields, repeating fields) very flexible.

The field and subfields to be used in the formatting process are determined by the formal description of the cards provided by the user. The internal record format does not put any general restrictions on the representation of the data elements. This means that

the application is easy to customize, but the user will have to provide quite a lot of very specific information in order to do this.

The built-in database management system is intended to handle creation, editing and deletion of records, as well as maintaining the link between individual records and the associated image file.

D. Internal 16-bit Character Set

In order to handle the range of characters to be expected in a typical multilingual European catalogue, an internal 16-bit character set is established, with the necessary string-manipulation routines, such as string-matching, string-sorting, display of strings etc. These routines are not as yet provided as standard in the programming languages.

The character set to be provided with the prototype is intended to conform to the first plane of ISO 10646-1, UCS-2, sometimes called "UNICODE", but in principle it is completely customizable by the user. The set should include the correct graphical forms for display on screen for the purpose of monitoring the progress or editing the records. A font editor was planned for customization of the character set and the display of the characters, making it possible for the user to determine exactly how the characters will appear on screen.

The font editor is to handle input and display of other information about the characters, like classification as capital or small letters, links between characters (e.g. telling that "A" and "a" are the same letter in capital and small versions), constituent elements of composite characters (like "^" and "a" for "â"), values for alphabetical sorting etc.

The idea is to have the characters on screen look as much as possible like the characters in the source, rather than having some of them represented as codes in two or more letters. This should make it easier for the human operator to visually verify and correct the results, both of the OCR and the formatting process.

The planned 16-bit editor and font editor has been postponed because true UNICODE based editors can be expected to be offered as ready made tools in the not so far future, as part of Windows NT or the next version of Windows for personal computers.

E. Formatting Specifications

The formatting of the cards are carried out according to formal specifications provided by the user in a special formal language, based on the so-called Backus-Naur notation. The input takes the form of one or more text files (ASCII text files), giving information about dictionaries and the so-called production rules (or just "rules") to be used in formatting.

This information is the result of the analysis that the library will have to carry out on the catalogue to be converted, since no two libraries or even catalogues require the same rules. One catalogue may require two or more formal specifications in order to cover all possibilities in the simplest way. The FACIT application can be told to make several runs through the records, each time using a different set of rules. In each run only the records rejected by earlier runs have to be processed.

The formal specifications are illustrated by a very simple case below. The space does not allow a more detailed presentation of this, but Technical Report No 3 describes the principles and presents the specification language in details. In order to produce such a specification a thorough knowledge of the catalogue is needed, as well as more than average understanding of the workings of a formal grammar. This will in many cases mean close collaboration between a librarian and a person with some training in computer science or computational linguistics.

The sample Formal Specification is to be taken as an illustration of principles, not as a fully worked out, realistic sample.

```
// Sample Formal Specification File

[Dictionaries]
PlaceNames = "PLACES.TBL" ;
FirstNames = "FSTNAMES.TBL" ;
LastNames = "LSTNAMES.TBL" ;

[Rules]
//*****
//* Author Rules      *
//*****
Author = AULastName Comma [Space] AUFstName
        {[Space] AUFstName} ;
AULastName = InitialWord [ "-" InitialWord ] |
            DictLookUp(LSTNAME) ;
AUFstName = InitialWord | Initial |
            DictLookUp(FSTNAME) ;
```

```

/*****
/* Title Rules          *
/*****
Title = "This is a title" : "Another title" ;

/*****
/* Imprint Rules       *
/*****
Imprint = ImpPlace [Comma] Blanks ImpYear [0
ImpPlace = (ImpCityName [ Comma Blanks
             ImpStateName ] ) | DictLookUp(PLACES) ;
ImpCityName = InitialWord | ("["InitialWord"]") ;
ImpStateName = InitialWord [FullStop] | Initial
              Initial ;
ImpYear = [ "cop." Blank ] ImpActYear |
           (" ImpActYear ")" | "[" ImpActYear "]" ;
ImpActYear = (("14" | "15" | "16" | "17" | "18" |
              "19" ) Digit Digit ) | (("14" | "15" | "16" |
              "17" | "18" | "19" ) Digit Digit ) | "S.A." |
           "s.a." ;
Pages = [ RomanNumber ("+" | ",") ] (( Number [
           "+" Number ] ("p." | "pages.") Comma " " (Ro-
           manNumber | "[" Number "]" ) "tab.")) ;

/*****
/* Series Rules        *
/*****
Series = "(" SeriesSpec ")" [FullStop] ;
SeriesSpec = SeriesTitle SeriesDelim SeriesNumber;
Series Title = IntitalWord { Blank Word } ;
SeriesDelim = FullStop Blank | Blank SemiColon
             Blank ;
SeriesNumber = Number | "Vol." Number | ("Vol." |
           "Bd." ) Blanks Number Comma Blanks "H. " Num-
           ber | "Hft." Blank Number | "H. " Number ;

/*****
/* Medium Level Rules  *
/*****
RomanNumber = RomanDigit { RomanDigit } ;
RomanDigit = "I" | "i" | "V" | "v" | "X" | "x" |
             "L" | "l" | "C" | "c" | "D" | "d" | "M" | "n";
InitialWord = UpperCase { LowerCars } ;
Initial = UpperCase FullStop ;
Word = Letter { Letter } ;
Number = Digit { Digit } ;
Blanks = Blank { Blank } ;

/*****
/* Low Level Rules    *
/*****
Digit = "$[Digit]" ;
Letter = "$[Letter]" ;
AlphaNum = "$[AlphaNum]" ;

```

```

UpperCase = "$[UpperCase]" ;
LowerCase = "$[LowerCase]" ;
NewLine = "$[NewLine]" ;
CardSepar = "$[ESCAPE]" ;
UnMatched = "$[Any]" ;
Blank = " " ;
Comma = "," ;
FullStop = "." ;
SemiColon = ";" ;
Colon = ":" ;
Hyphen = "-" ;
Slash = "/" ;
LeftParen = "(" ;
RightParen = ")" ;
LeftBrack = "[" ;
RightBrack = "]" ;

```

If needed a separate Formal Specification file may be provided by the user specifying areas of the card containing data elements defined by position in the card, like Top Left Corner, Top Right Corner, Left Margin, Bottom Left Corner and Bottom Right Corner. Then the computer will first identify text or other characters in these areas and place them at a designated location in the sequence of the text. Further handling of this information is then carried out by sections in the Formal Specifications file.

F. Dictionaries

For verification of strings etc. a set of dictionaries may be provided by the user. The dictionaries to be used are listed in the Parameter File (See section H) and at the head of the Formal Specification (see sample above). The formal language includes orders to call dictionaries for specific operations such as verification, substitution of strings or look-up of coded information.

Typical dictionaries could include: Lists of accepted Location Marks or Class Marks. List of publishing places (with country code). List of typical First Names. List of Typical Last Names or Family Names. List of Special Names, like names of kings and other princes, popes etc. List of publisher's names. List of typical words and phrases used in cataloguing: "Edited by", "Herausgegeben von", "S.L.", "S.A." etc. List of word typically used in titles (with indication of language(s) where that word occur).

The FACIT application is planned to include tools to interactively update the dictionaries with new names, words, location marks etc. identified in the process of verification.

At the moment a dictionary is a simple text file with comments after a "//" (Double Slash). The entries do not have to be sorted, but alphabetical sorting is useful from the point of view of the human user. But more sophisticated dictionaries and dictionary look up routines have to be developed at a later stage. This has been investigated especially by the Italian partners.

```
//Sample First Name Dictionary
"Alexander"
"Alexandra"
"Algernon"
"Anders"
"Anne"
"Arthur"
.
.
.
"Ben"
"Benjamin"
"Birthelme"
.
.
.
```

When checking the dictionaries for verification of strings during formatting or other processes the information provided by the Transition Table should be used to determine possible matches, taking into account possible misreadings of characters. If a match is not found further searching is to be made on strings constructed by substituting error prone characters with characters that could be the correct ones. The most error prone characters will be substituted first and the process carried on until a match is found or the possibilities exhausted.

If a match has been found after such substitution this may be taken as an indicator that the original string contained an error. The error may be corrected automatically or by user intervention depending on the circumstances and information provided in the Parameter File (see section H).

G. Specification of Output Format

After the formatting and weeding out of "spelling errors" the resulting formatted records have to be converted from the internal record format into an output file in the target bibliographic format and a character set acceptable to the cataloguing system where the records are going to be used.

This - like the input - makes it necessary for the user to provide two tables:

1. A Specification file telling the computer how to tag the bibliographic data elements (kept in the internal record format as separate "fields" or "objects") to conform with the target format. This could be any bibliographic format suitable for exchange of bibliographic records, but UNIMARC is the target format aimed at by the project.

2. A Conversion Table telling the computer how the characters in the internal 16-bit character set are to be represented in the output files. UNIMARC is expected to support UNICODE characters in the foreseeable future, in which case no conversion should be needed.

H. Outline of the Overall Process of Using the Facit Prototype.

As indicated above the FACIT application includes a suite of programs to handle the diverse processes of input and output of record files, internal record management, string substitution, dictionary look up, and of course the overall management of the formatting and error detection process.

Dialogue with the user, both the system administrator maintaining the system, and the operator doing the actual conversions, takes place through a Windows based graphical interface. In the Prototype the language used will be English, but all texts in the interface can be translated into other languages (using both Latin and Greek characters).

The basic setup of the system for a specific conversion task is made by providing a general Parameter File (also an ordinary text file, that may be produced with a standard text editor).

The Parameter File tells the system which Dictionaries, Tables and Specification files to use for the session in progress. The Parameter File will also include information about values for certain parameters to be used, like threshold values for error statistics calculated on the basis of the Transition Table.

A formatting session will typically start by reading in the general Parameter file, then the first of the input files with accompanying image files, using the Conversion Table and the Substitution Table. Then the records will be checked for bibliographic records running over more than one card, using information

provided in the Parameter file. The text will be concentrated on one record and the images of the cards linked to that record.

When the internal records have been established the formatting process will start. This may be a two step process, first decomposing the cards by extracting information from specific areas like corners and margins, then the actual formatting using the formal specifications provided. Each record is analysed according to the specifications to see if it complies with the rules specified. In the course of this possible errors may be detected and presented to the operator for verification and possible correction, or errors may be corrected automatically by the program. The image of the card may be displayed at any time alongside the text as represented in the internal record, so that the operator may check the original text without having to find the original card in the catalogue. The internal record may be edited directly on the basis of this if needed. It will be possible to browse in the images in sequence in case an image was misplaced in the input process.

If the card is accepted by the formal analysis it is marked as O.K. It will have been updated so that it now includes fields and subfields with the data elements identified during the analysis. If the card is not accepted it will be marked as failed, and it will not be updated - apart from corrections made by the operator.

If the Parameter File specifies a series of Formal Specifications to be used in a certain sequence, the next in sequence will be activated and used to analyse the failed cards. The process is repeated until the sequence is exhausted. Any cards still remaining as failed will then have to be corrected by the Human operator, normally using the image of the card as a reference point. This will be the case with cards that were badly recognized by the OCR for various reasons, and with cards that differ in format and bibliographic contents from the majority of the cards in the catalogue in question. Then the sequence will have to be activated again. If cards are still remaining they will have to be output to an Error File in order to be handled "manually".

"Debugging" of the all files used to control the process may be carried out interactively using an editor.

After the formatting, which will include a substantial part of the error detection and correction, certain parts of the records like the title may then be submitted to further error checking using a speller

checker or an analysis of character sequences based on n-grams (sequences of n-characters, n being 2, 3, 4 ...). This process may be combined with determining the language of the publication. The details of this part of the proces has not been worked out yet, but the feasability was investigated in Hørning (1995).

The prototype demonstrates the interface aimed at, but the actual implementation, using Microsoft Access 2.0, is too cumbersome to be used for production purposes. In fact using the DOS programs without the interface will speed up the conversion process significantly.

Due to certain memory restrictions in the DOS programs the prototype is not - at the moment - able to process input files with more than 20 -25 records at a time.

Further development is needed to optimize the workings of the interface and the underlying programs, as well as providing the missing editing tools, sophisticated dictionary look up, and computer assisted error detection and correction.

Calculation of costs

A. Introduction

Due to a series of unforeseen and special circumstances encountered in the project - as reported in the Summary and the previous sections - it has not been possible to establish valid empirical data for the costs of using the FACIT Prototype for a large scale production. However a costing model has been worked out to show the crucial factors to take into account when estimating costs.

The costs may conveniently be divided into the following groups:

1. Cost of acquiring the necessary equipment
 2. Cost of setting up the formatting system, including formal analysis of the catalogue(s) to be converted.
 3. Cost of producing the plain text input for the formatting process
 4. Cost of proof reading and correction of results
 5. Cost of formatting and production of output files in the desired format
 6. Cost of introducing retroconverted records into the electronic catalogue system of the library.
1. - 6. add up to the total cost of the retroconversion exercise.

B. Cost of acquiring the necessary equipment

The cost of equipment, which includes computer(s), scanner, OCR software, text-editing software, other applications needed (such as Microsoft Access for the FACIT Prototype) etc. is a fixed cost.

The cost of equipment differs from country to country and is changing all the time. It also depends on the hardware and software systems available at the time of carrying out the retroconversion project.

When estimating the cost of the equipment to be used it is worthwhile to consider duplicating certain parts in order to work with parallel production lines so as to reduce the time span of the project. This will of course also include a duplication the labour effort at certain parts of the overall process.

The main point in calculating the cost of equipment - apart from taking into account all the components needed for the project - is to make provision for the depreciation of the equipment in accordance with normal accounting principles and the time span of the project. If the equipment is used for other purposes during the time span of the project this should also be taken into account.

C. Cost of setting up the formatting system, including formal analysis of the catalogue(s) to be converted.

Analysing the catalogue, working out the correct formal specifications and setting up and debugging the system is a unavoidable feature of using OCR and automatic formatting, whether using the FACIT Prototype or any similar system, even creating a customised formatting program from scratch.

This effort will involve specialists in cataloguing as well as systems analysis, preferable persons with knowledge of computational linguistics or compiler construction.

The effort depends on the complexity of the catalogue, not on the size, and something like 2 - 3 man months will have to be allowed for this.

Since all new catalogues or catalogue sequences will have to be analysed and described in equal depths no real saving of effort is to be expected. For persons doing this for the first time allowance will have to be made for learning a new way of working, while cumulated experience will of course tell in the time needed to do this job.

When calculating the cost one should include 1) the actual wages of the persons involved, 2) indirect costs, such as social costs, pension benefits etc, 3) overheads in the form of general costs of administration, office support, telephone, fax, electricity, heating, building maintenance etc. It is important to include the full costs of in-house staff when comparing with the cost of having the job done by an outside agency.

If the purpose of the calculation exercise is to compare the cost of retroconversion using OCR and automatic formatting with using an external agency using other methods, one should realize that a thorough analysis of the existing card catalogue and specification of the data to transfer from the card catalogue to the electronic catalogue will be needed in most cases so that the cost of this analysis will have to be included in all the cases to be compared.

The conclusion is that the cost of analysis and setting up of the system is best treated as a fixed cost for each catalogue/catalogue sequence to be converted, dependent on the complexity of the catalogue and the skill of the persons carrying out the job.

D. Cost of producing the plain text input for the formatting process

Transferring text from the cards to a plain text digital file is a variable cost that is to all practical purposes directly proportional to the number of cards to be processed. Apart from this it depends on the cost of labour used (with all direct and indirect costs included as in C. above) and the speed of processing.

The speed of transfer can be ascertained by carrying out a small series of tests on a sample of the actual catalogue. Using this an average transfer rate per hour is calculated and from this an average transfer rate per working day is interpolated. For the transfer rate per day remember to take normal breaks etc into account as well as the working hours normal for the group of employees involved.

The formula will be

$$\text{Cost of Transfer} = \frac{A}{B} * C$$

where A is the number of cards to be retroconverted, B is the average number of cards processed per working day and C is the cost per working day of the staff doing the work (including social benefits, overheads etc.).

The actual speed mainly depends on the equipment being used but also on the average amount of text on the cards. It will have to be ascertained empirically as noted above. With the fast development of computers, scanners and OCR software it is not possible

the give a number that will stay valid for a very long time. With the equipment used for the FACIT Project is was possible to transfer 400 cards per hour or something like 2.500 cards per working day, including scanning and OCR, but not proof reading and correction of errors.

With cards that cannot be processed with scanner and OCR because of the state of the cards, it might still be cost effective (when the whole process, including formatting, is taken into account) to have a skilled typist copy the cards into digital form by reproducing the text letter by letter.

The procedure given above for estimating the cost of transfer can also be used in this case, since it does not depend on the method used. But the speed will of course be much slower.

When estimating the cost of alternative ways of doing retroconversion it is important to be aware of costs have to be incurred by the library in all cases. If the retroconversion is to be carried out by an outside agency the cards may have to photocopied before shipping to the agency. Photocopying cards will in most cases be more expensive than scanning the cards with a suitable document feeder, since very few photocopiers are able to handle cards in bulk. Usually they will have to laid out on the glass plate by hand, which is a slow and unstable process.

E. Cost of proof reading and correction of results

This cost will also be a variable cost and to all practical purposes directly proportional to the number of cards to be processed. And the method of estimating the cost is the same.

$$\text{Cost-of-correction} = \frac{A}{B} * C$$

where A is the number of cards to be retroconverted, B is the average number of cards processed per working day and C is the cost per working day of the staff doing the work (including social benefits, overheads atc.).

In this case the speed depends partly on the number of errors to be corrected, partly on the method used to check the cards against the original. There are many possible ways of reducing the time spend checking the cards, but they have not been measured as

part of the FACIT Project. Empirical tests will have to be carried out in each case.

Proof reading and correction of errors may be carried out as part of the OCR processing in which case it will be more convenient to measure the cost of transfer and the cost of correction as one.

Proof reading and error correction may also be carried out as part of formatting, in which case it will be more convenient to measure the cost of correction and the cost of formatting as one.

Separate calculation of the cost of correction is suitable when the process is carried out independently of the other processes in the retroconversion chain.

F. Cost of formatting and production of output files in the desired format

Again is this cost is a variable cost and to all practical purposes directly proportional to the number of cards to be processed. And the method of estimating the cost is the same as for cost of transfer.

$$\text{Cost-of-Formatting} = \frac{A}{B} * C$$

where A is the number of cards to be retroconverted, B is the average number of cards processed per working day and C is the cost per working day of the staff doing the work (including social benefits, overheads atc.).

The speed depends partly on the equipment used, partly on the complexity of the cards and the average amount of text on the cards.

In the tests carried out with the FACIT Prototype on a 386DX, 60 Mhz PC with 16 MB RAM the speed was about 200 cards per hour or something like 1200 cards per working day. But these numbers can only be taken as pointers, since newer and faster equipment is already available.

G. Cost of introducing retroconverted records into the electronic catalogue system of the library.

It is important to realize - with all types of retroconversion - that the records produced will have to be introduced into the electronic catalogue of the library and that this will incur costs as well.

If the output of the retroconversion process is a file in a format that may be read and processed directly by the input routines of the system this cost may be fairly small. Otherwise a conversion program may have to be written specifically for this purpose.

Eliminating duplicates and preventing the new records from overwriting existing information may also be a concern in certain cases.

These processes are outside the scope of the FACIT project, but they have to be included when estimating the cost of a specific retroconversion strategy.

H. Comparisons of Total Costs

When comparing different ways of doing retroconversion it is important to take all costs into account, not only direct costs like cost of equipment or payments to external agencies.

It is a normal error to underestimate the cost of using the library's own staff when deciding whether to use an external agency or not, so that inhouse processing appears cheaper than it actually is.

On the other hand libraries tend to underestimate the effort required of its own staff when looking at offers from commercial agencies, like forgetting the cost of photocopying cards before shipping to the outside agency and the cost of incorporating the results into the catalogue system.

Useful examples of cost comparisons is found in: Deutsches Bibliotheksinstitut: Retrokonversion. Konverzion von Zettelkatalogen in Deutschen Hochschulbibliotheken - Methoden, Verfahren, Kosten - Redaktion: Kirsten Weber. Deutsches Bibliotheksinstitut. Berlin 1993.

Concluding Remarks

A. Compliance with international standards

The primary target format of the formatting programs is UNIMARC, as described in the UNIMARC manual and supplementary publications (IFLA, 1987), but other formats are also possible.

Specific formats used by the library systems of the involved libraries, and other (national) standards, i.e. danMARC, may have to be handled by special conversion programs, having the internal records of the FACIT prototype or an intermediate file produced from this as a source. Writing specific conversion programs is not part of the project.

The bibliographic data elements of the source (the existing card catalogue) should be identified according to the ISBD standard (IFLA, 1991 and IFLA, 1987). Only the data elements present in the card will be present in the resulting formatted file (unless they can be automatically inferred from the information present or the context).

Information about holdings should comply with: ISO/DIS 10324:1991 Information and documentation - Holdings statements - Summary level.

The level of bibliographic description in the converted catalogue and the differentiation possible in the bibliographic format is entirely dependent on the information in the source. No bibliographic information not present or implied in the source will be added in the conversion process. This will in some cases mean that the level of cataloguing will fall below the minimum standards used today. Enhancing the records produced by the conversion process - either by external sources or by reference to the original publications - is outside the scope of the present project.

Several different standards exist for the representation of character sets in machinereadable form, and the project includes a discussion of the problems of character sets in the context of retroconversion projects. The two main objectives of the representation used are 1) to conserve the differentiation of characters in the source file, so that no information is lost in the conversion, and 2) to make the characters

easily recognizable to a human operator inspecting the records on the screen or on paper for control and correction purposes. These two objectives may to a certain extent be in conflict with each other because of limitations in available software and a practical compromise has to be reached.

The project relies on official standards (ISO standards and national standards) as well as de facto standards like the "Code Page" system of IBM compatible personal computers.

At the beginning of the project the ISO 6937/2:1983 and the ISO 8859 series was used as the reference standard for Latin Characters. For the Greek characters a standard used by the National Library of Greece for representation of Greek characters with all variants (accents, diacritics etc) was used, since the ISO 8859 for Greek characters is too restricted for library purposes.

After the introduction of ISO 10646-1 and the subset of this called UNICODE, these are used as the point of reference for both Latin and Greek characters.

Since very few applications for the PC and Windows environment exist that has implemented UNICODE, a solution allowing internal representation of 16-bit characters has been developed for the project as part of the FACIT Prototype. The character codes may be assigned arbitrarily by the user, but it is recommended to follow ISO 10646-1.

B. Benefits and Results

The project has led to a better understanding of the issues involved in the retroconversion process and is a step towards producing better tools and methods for the application of OCR on suitable parts of library catalogues.

The prototype produced is in itself an important result, but the analyses carried out of the catalogues and the errors, and the methods developed for this should provide a broader basis for further work in the field to the advancement of library automation.

A preliminary analysis seem to show that the methods developed to convert an ASCII file into a formatted bibliographic file could also be used to advantage with sources not suitable for OCR, i.e. bad or worn typewritten/printed cards, cards with handwritten additions and all handwritten cards and registers. In

this case the cards have to be copied by direct keying to the text format.

The use of techniques for converting existing catalogues rather than relying on external sources is of special interest with unique collections, among others collections of national literature and special collections of foreign literature, as well as collections of non-book material not represented in national bibliographies, where corresponding machinereadable records of high quality are not normally available in significant numbers and never will be unless produced by this or a similar method.

With the techniques investigated in the project suitable parts of important collections may be made widely accessible nationally and internationally within a reasonable time and with a reasonable effort in terms of money and labour. The day-to-day administration of the collections will also gain by having catalogue records with the signatures, class marks and indexing terms of the library as well as basic bibliographic descriptions.

But it has to be recognized that the FACIT Prototype is not yet suitable for proper production work, and that scanning and OCR will only be possible with certain older catalogues.

Since the records produced by the process investigated will not be enhanced, the bibliographic quality of the source will determine whether the resulting records will be suitable for exchange with other libraries as a source of retroconversion. In the case of national collections and other unique collections they will be the best available. The records may be used as the basis of further enhancement by local effort (typically for the purpose of providing a machinereadable national bibliography) referring to the original publications.

In other libraries and other collections the records may be enhanced later by automatic means using external sources, while retaining the information specific to the library in question.

As part of an European co-operation in the field of libraries the project has produced valuable insights into similarities and differences in earlier cataloguing rules and practices, thus contributing to our understanding of the problems involved in establishing common standards and shared resources.

C. Exploitation plans

The prototype developed in the project will be distributed on a Public Domain basis for non-commercial use in libraries, together with the draft user's manual (English only). Commercial use will need a license from the partnership.

The main subcontractor in the project will be granted a license for further development of the prototype, provided that only the cost of further development and the cost of marketing and sale is included in the price calculation of the resulting application, but not the cost of developing the prototype. The main subcontractor will further have the opportunity to market training and support of the prototype as well as any further developments.

The other companies involved as subcontractors will also be granted a license to provide a service based on the methods and tools developed or to use the work as a starting point for further development based on the prototype or directly on the specifications produced by the project.

The companies in Italy and Denmark have all expressed a wish to carry on the work on a commercial basis if it is possible to attract the necessary customers.

The prototype will be used by the libraries involved in the project for further retroconversion work, and where necessary the libraries will undertake to convert files in UNIMARC format produced by the prototype to the internal format of the catalogue system of the library.

The partners are committed to implementing the methods established in the project and to further work on retroconversion of catalogues in their own country.

References

Allen, 1987

Allen, James: Natural Language Understanding. The Benjamin/Cummings Publishing Company Inc. Menlo Park, Calif. 1987. 574 p.

Avram, 1972

Avram, Henriette D.: RECON Pilot Project. Library of Congress, Washington D.C. 1972. 49 p.

Beaumont & Cox, 1989

Beaumont, Jane & Joseph P. Cox: Retrospective Conversion. A Practical Guide for Libraries. Meckler, Westport/London. 1989. 198 p.

Bokos, 1993

Bokos, George: "UNIMARC, CDS/ISIS and conversion of records in the National Library of Greece." In: Program vol.27 (2), April 1993. Pp. 135 - 148.

Boserup & Holtse, 1992

Boserup, Ivan & Lisbet Holtse: "Automatic Conversion at The Royal Librar, Copenhagen. A Progress Report." Paper given at The Second International Conference on Retrospective Cataloguing, München, 28.-29.January 1992.

Cain, 1993

Cain, Chris: "WordScan Plus 2.0 for Windows." In: Personal Computer World, November 1993. P. 382 - 384.

CEC, DG XIII B, 1990

Report of the Workshop on Retrospective Conversion of Catalogues. Problems, Priorities and Projects under the Library Plan. Commission of the European Community, Directorate General XIII B, Luxembourg. 1990. [var.pag.] Printed as draft.

Council of Europe, 1989

Guidelines for Retroconversion Projects prepared by the LIBER Library Automation Group. Council of Europe, Council for Cultural Co-operation, Working Party on Retrospective Cataloguing. 14 July 1989 (revised). 13 p.

- Crawford & Lee, 1990
 Crawford, R.G. & Susan Lee: "A prototype for fully automated entry of structured documents." In: The Canadian Journal of Information Science/Revue canadienne des sciences de l'information. Vol. 15, No. 4. December 1990. Pp. 39 - 50.
- Diehl & Eglowstein, 1991
 Diehl, Stanford & Howard Eglowstein: "Tame the Paper Tiger." In: Byte, April 1991. Page 220 - 238.
- Deutsches Bibliotheksinstitut, 1993
Retrokonversion. Konverzion von Zettelkatalogen in Deutschen Hochschulbibliotheken - Methoden, Verfahren, Kosten - Redaktion: Kirsten Weber. Deutsches Bibliotheksinstitut. Berlin 1993. 411 p.
- Fresko & Brindley, 1992
 Fresko, Marc & Lynne Brindley: Optical Disc Technology and European Libraries. A Study of User and Technical Requirements. A report for the Commission of the European Communities. Bowker-Saur. London ... 1992. 280 p. (Specially p. 41 - 78 (Scanner Technology) and p. 109 - 116 (Optical Character Recognition))
- Georghiades & Jacobs, 1993
 Georghiades, Panicos & Gabriel Jacobs: "Let your PC do the typing." In: PC Today, Vol 6, No 12, April 1993. p. 119 - 121. + "A question of character." In: PC Today, Vol 7, No 1, May 1993. P. 146 - 147. (Short evaluations of: SKZI Recognita Plus, Textpert and Calera WordScan Plus).
- Gough, 1992
 Gough, John K.: Syntax Analysis and Software Tools. Addison Wesley Publishing Company, Reading, Mass./ Wokingham (U.K.), 1992. 400 p.
- Govindan & Shivaprasad, 1990
 Govindan, V.K. & A.P. Shivaprasad: "Character Recognition - A Review." In: Pattern Recognition, Vol. 23, No. 7, 1990. Pp 671 - 683.
- Grunin, 1991
 Grunin, Lori: "Calera WordScan Plus" In: PC Magazine 10: 174-77, January 15, 1991.

Harrison, 1985

Harrison, Martin: "Retrospective Conversion of Card Catalogues into Full MARC Format Using Sophisticated Computer-Controlled Visual Imaging Techniques." In: Program 19 (July 1989). p. 213-30.

Hein, 1986

Hein, Morten: "Optical Scanning for Retrospective Conversion of Information." In: The Electronic Library, December 1986. Vol.4, No.6. P. 328 - 331.

Hendley & Pritchard, 1993

Hendley, Tony & John Pritchard: "Recognition technology for data entry, document indexing and re-publishing." In: Information Management & Technology. Vol 26, No 2, 1993. P. 70 - 78.

Hørning, 1995

Hørning, Annette: Language Identification using an Artificial Neural Net. Report, Department of Lexicography and Computational Linguistics The Aarhus School of Business, 1995. (Available from Statens Bibliotekstjeneste, Copenhagen or the FACIT Project Manager)

IFLA, 1987

ISBD (M). Revised Edition. IFLA Committee on Cataloguing. International Federation of Library Associations, London. 62 p.

IFLA, 1987

UNIMARC Manual. Edited by Brian P. Holt with the assistance of Sally H. MacCallum & A.B.Long. IFLA Universal Bibliographic Control and International MARC Programme/British Library Bibliographic Service, London. 1987. 482 p.

IFLA, 1990

IFLA Journal 16 (1). Special issue devoted to an overview of projects and approaches to retrospective conversion, providing an international perspective.

IFLA, 1991

ISBD (S). Revised Edition. Joint Working Group of the IFLA Committees on Cataloguing and Serial Publications. London 1991. 62 p.

ISO 5426

Extension of the Latin alphabet coded character set for bibliographic information interchange. Second Edition. International Standards Organisation. 1983.

- ISO 5427
Extension of the Cyrillic alphabet coded character set for bibliographic information interchange.
International Standards Organisation. 1984.
- ISO 5428
Greek alphabet coded character set for bibliographic information interchange. Second Edition.
International Standards Organisation. 1984.
- ISO 6937
Information Technology - Coded graphic character sets for text communication - Latin alphabet.
Second edition. International Standards Organization. 1993
- ISO 8859-1
Information processing - 8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1. International Standards Organisation. 1987.
- ISO 8859-2
Information processing - 8-bit single-byte coded graphic character sets - Part 2: Latin alphabet No. 2. International Standards Organisation. 1987.
- ISO 8859-3
Information processing - 8-bit single-byte coded graphic character sets - Part 3: Latin alphabet No. 3. International Standards Organisation. 1988.
- ISO 8859-4
Information processing - 8-bit single-byte coded graphic character sets - Part 4: Latin alphabet No. 4. International Standards Organisation. 1988.
- ISO 8859-7
Information processing - 8-bit single-byte coded graphic character sets - Part 7: Latin/Greek alphabet. International Standards Organisation. 1987.
- ISO/DIS 10324
Information and Documentation - Holdings Statements - Summary Level. ISO/DIS 10324:1991. International Standards Organisation. 1991.
- ISO/IEC 10646-1
Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane. ISO/IEC 10646-1:1993 (E).

- Jain, 1991
 Jain, Raj: The Art of Computer Systems Performance Analysis. Techniques for Experimental Design, Measurement, Simulation and Modeling. John Wiley and Sons, Inc., New York. 1991. 685 p.
- Jennings, Newman & Wilkinson, 1982
 Jennings, Newman & Wilkinson: "Data capture by optical scanning of published material for database enhancement." In: Program 16 (1). January 1982. Pp. 17 - 27.
- Jensen, 1986
 Jensen, Hans Erik: Problemer i forbindelse med retrospektiv inddatering af kortkataloger og optisk læsning. (Problems of retrospective conversion of card catalogues and optical character recognition.) Statsbiblioteket, Århus. 1986. 52 p.
- Nagy & Seth, 1996
 Nagy, George & Sharad Seth: "Modern Optical Character Recognition" In: Froehlich, Kent & Hall (eds.): The Froehlich/Kent Encyclopedia of Telecommunications. Volume 11. Marcel Dekker, Inc. New York / Basel / Hong Kong. 1996. P. 473 - 531.
- Ogg & Ogg, 1992
 Ogg, Harold C. & Marlene H. Ogg: Optical Character Recognition: A Librarians Guide. Meckler, London. 1992. 171. ISBN 0-88736-778-X.
- Ortiz-Repiso, Virginia & Yolanda Ríos: "Automated Cataloguing and Retrospective Conversion in the University Libraries of Spain." In: Online & CD-ROM Review, Vol. 18, No. 3. 1994. p. 157 - 167.
- Peruginelli, Bergamin & Ammendola, 1992
 Peruginelli, Susanna, Giovanni Bergamin & Pino Ammendola: "Character sets: towards a standard solution?" In: Program, vol 26, no. 3, July 1992. Pp 215 - 223.
- Reigem, Ore & Hofland, 1990
 Reigem, Øystein, Espen S. Ore & Knut Hofland: OCR - teknologi for innlesing av humanistisk kildemateriale. Status for optisk tegngjenkennung i dag. Rapportserie no 49. NAVFs Edb-senter for Humanistisk Forskning. Bergen, 1990. 27 s. ISBN 82-7283-058. ISSN 0800-5796.
- Seymour, 1990
 Seymour, Jim: "HP ScanJet Plus". In: PC Magazine 9:166, January 16, 1990.

Shah, 1992

Shah, Raimond: "Automatisches Erfassung von gescannten Bibliothekskarten: Wissensbasiertes Verfahren." In: OUTPUT, 9403 Goldach, Jubiläumsausgabe 1992. p. 89 - 90.

Simon, 1991

Simon, Barry: "Recognita Plus: OCR with Strength in Hardware." In: PC Magazine 10: 50, April 1991.

Smith & Merali, 1985

Smith, John W.T. & Zinat Merali: Optical Character Recognition: The Technology and its Application in Information Units and Libraries. Library and Information Research Report 33. British Library, London. 1985. 125 p.

Steve, Dickman & Parascandolo, 1990

Steve, Roth, Chris Dickman & Salvatore Parascandolo: ScanJet Unlimited. Peachpit Press, Berkeley, CA. 1990. 226 p.

Sun et.al. 1992

Sun, Wei, Lon-Mu Liu, Weining Zhang & John Craig Comfort: "Intelligent OCR Processing." In: Journal of the American Society for Information Science 43(6):422-431, 1992.

Süle, 1990

Süle, Gisela: "Bibliographic Standards for Retrospective Conversion" In: IFLA Journal 16 (1). 1990. P. 58 - 63.

Syré, 1987

Syré, Ludger: Retrospektive Konversion. Theoretische und praktische Ansätze zur Überführung konventioneller Kataloge in Maschinenlesbare Form in den USA, Grossbritannien und der Bundesrepublik Deutschland. Deutsches Bibliotheksinstitut, Berlin. 1987. 231 p.

Wayner, 1993

Wayner, Peter: "Optimal Character Recognition." In: Byte, December 1993. P. 203 - 210. (Mentions: ExperVision, Caere OmniPage and XIS (Xerox Imaging System) Lexifier)

Wheelright, 1992

Wheelright, Geoff: "OmniPage Professional 2.0." In: Personal Computer World September 1992. P. 190 - 192.

Names and addresses of participants

Partners of the FACIT Consortium

1. Statens Bibliotekstjeneste (SBT)
(National Library Authority)
Nyhavn 31 E
DK-1051 København K
Danmark

Tele: +45 33 93 46 33

Fax: +45 33 93 60 33

Contact: Ulla Dresler, Head of Administration

Other Contacts: Niels Erik Wille (Project Manager), now Senior Lecturer at Roskilde University (E-mail: new@snow.ruc.dk)

2. Statsbiblioteket (SB)
(State and University Library)

Universitetsparken
DK-8000 Århus
Danmark

Tele: +45 89 46 20 22

Fax: +45 86 13 27 04

Contact: Hans Erik Jensen, research librarian

2. Biblioteca Nazionale Centrale (BNCF)
Piazza Cavalleggeri 1
I-50122 Firenze
Italia

Tele: +39 55 24 44 41

Fax: +39 55 23 42 482

Contact: dr. Claudia Miconi, librarian

Other contacts: Gian Luca Corradi, librarian, BNCF
Guiseppe Vitiello, senior librarian, BNCF

3. Biblioteca Nazionale "V.E.III" (BNN)
Palazzo Reale
Piazza Plebiscito
Napoli
Italia

Tele: +39 81 40 79 21
Fax: +39 81 40 38 20

Contact: dr. Mena Savarese
Other contacts: dr. Vera Valitutto , now Director
of the University Library of Napoli

4. Ethnike Bibliotheke tes Hellados (EBH)
32 Panepistimio Str.
GR-106 79 Athena
Hellas

Tele: +30 1 36 08 141
Fax: +30 1 36 08 495

Contact: dr. George Bokos, head of cataloguing
dept.
Other contacts: Joanna Demopoulos, acting head of
cataloguing dept.

Associated contractor

6. Det Kongelige Bibliotek
Christians Brygge 8
P.O.box 2149
DK-1016 København K
Danmark

Tele: +45 33 93 01 11
Fax: +45 33 32 68 30

Contact: Ivan Boserup, senior librarian
Other contacts: Lisbeth Holtse: research librarian

Participating Companies

SYNERGI
Bakkevej 13
DK-2950 Vedbæk
Danmark

Tele: +45 45 66 00 56
Fax: +45 42 89 44 56

Contact: Kim Mikkelsen, managing director

StudioErre s.n.c.
Sede largo Duca della Ferrantina 1
CAP 80121
Napoli
Italia

Tele + fax: +39 81 41 19 73

Contact: Gianluigi Visco, director

Stefano Tulini
Consultant in Computer Science
Via di Goletta 50
I-56121 Pisa
Italia

Tele + Fax: +39 50 40 361

Appendix 2

List of reports and other products of the project

- 1 Optical Character Recognition for Retroconversion of Catalogue Cards: Hardware, Software and Character Representation. By Niels Erik Wille. (FACIT Technical Report No. 1). Copenhagen, October 1996.
- 2 Using Kurzweil scanning systems as platform for card scanning. By Ernst Pedersen, ADP. (FACIT Technical Report No 1.1). Copenhagen, 1994.
- 3 A Framework for the Analysis of Catalogue Cards. By Niels Erik Wille & Vera Valitutto. (FACIT Technical Report No 2). Copenhagen, Revised version, October 1996.
- 4 Description of the Card Catalogue at Statsbiblioteket. By Hans Erik Jensen & Dorete Larsen. (FACIT Technical Report No. 2.1). Århus, 1994.
- 5 Analysis of Three Card Catalogues of Biblioteca Nazionale Centrale di Firenze. By Claudia Miconi & Gian Luca Corradi. (FACIT Technical Report No. 2.2.1). Firenze, 1993.
- 6 Descrizione formale delle schede del Catalogo Palatino, Biblioteca Nazionale Centrale di Firenze. By Database Informatica. (FACIT Technical Report No. 2.2.2). Firenze, 1994
- 7 Description of Italian Cataloguing Rules: 1886 - 1979. By Rosella Ruoppolo & Vera Valitutto. (FACIT Technical Report No 2.3.1). Napoli, 1993.
- 8 Lex Rules in FACIT Format for BNN Catalogue - a first approach. By StudioErre di Gianluigi Visco. (FACIT Technical Report No 2.3.2). Napoli, 1994.
- 9 Descrizione formale delle schede del Catalogo BNN. By Stefano Tulini. (FACIT Technical Report No. 2.3.3). Napoli, 1994.
- 10 Analysis of the Card Catalogue of the National Library of Greece. Prepared by Joanna Demopoulos. (FACIT Technical Report No. 2.4). Athens, 1994.

- 11 Error Analysis and Correction in Retroconversion.
By Hans Erik Jensen. (FACIT Technical Report No 3). Århus, October 1996.
- 12 Error Analysis of converted cards from the catalogue, the State and University Library. By Hans Erik Jensen (FACIT Technical Report No 3.1). Århus, 1996.
- 13 Cataloguing Files through Retrospective Conversion. Error Analysis. By Database Informatica. (FACIT Technical Report No 3.2). Firenze, 1994.
- 14 First sample of Catalogue Cards: Error Analysis. By StudioErre di Gianluigi Visco. (FACIT Technical Report No 3.3). Napoli, 1994.
- 15 First Sample of Catalogue Cards: Error Analysis. By Joanna Demopoulos. (FACIT Technical Report No 3.4). Athens, 1995.
- 16 Final Technical Report including Error Analysis of the Cards of the National Library of Greece. FACIT Technical Reptot No 3.4.2). Athens, January 1996.
- 17 The FACIT Prototype. Manual and Documentation. By SYNERGI. (FACIT Technical Report No 4). Copenhagen, October 1996. (= System Administrators Manual)
- 18 Noder. Analyse af kort og kartoteker i forhold til retrokonvertering. By Hans Erik jensen (Preliminary work for FACIT Technical Report No 4.1) Århus, 1994.
- 19 The Retroconversion of the Palatino Catalogue. Final Report. By Stefano Tulini. (FACIT Technical Report No 4.2.1) Firenze, 1995.
- 20 Examples of retroconversion of the Palatino Catalogue. By Stefano Tulini. (FACIT Technical Report No 4.2.2). Firenze, 1995.
- 21 Text treatment and realization of specialized dictionaries. By Stefano Tulini. (FACIT Technical Report No 4.2.3). Firenze, 1995.
- 22 Report on scanning the Palatino Catalogue. By SOFTTEAMware. (FACIT technical Report No 4.2.4). Firenze, 1996.
- 23 Report on the End of Project by BNN, Naples. By StudioErre di Gianluigi Visco. (FACIT Technical Report No 4.3.1) Napoli, 1996.

- 24 Text treatment and realization of specialized dictionaries. By StudioErre di Gianluigi Visco. (FACIT Technical Report No 4.3.1). Napoli, 1996.
- 25 Retroconversion of Older Card Catalogues using OCR and Automatic Formatting. Project Overview and Final Report. By Niels Erik Wille. (FACIT Technical Report No 5). Copenhagen, October 1996.
- 26 FACIT Prototype: Suite of DOS programs with a Windows Shell (Microsoft Access application). By SYNERGI. (Working Prototype, version 1.1). Copenhagen, 1995.
- 27 Report on Dissimination Activities of BNN. By Vera Valitutto. Napoli. 1993.

Items 1, 3, 11, 17, 25 and 26 are generally available and can be ordered free of charge from:

the Coordinating Partner

Statens Bibliotekstjeneste
Nyhavn 31 E
DK-1051 Copenhagen K
Denmark

Fax: +45 33 93 60 33
E-mail: SBT@sbt.bib.dk

from the Project Manager

Senior Lecturer Niels Erik Wille
Dpt. of Computer Science, Communication and Education
Roskilde University
P.O.Box 260
DK-4000 Roskilde
Denmark

Fax: +45 46 75 34 15
E-mail: new@snow.ruc.dk

or directly from the Web site of the FACIT Project:

<http://www.komm.ruc.dk/FACIT/>

The rest of the reports are distributed at the discretion of the responsible partner.

Conference Presentations

Meeting of Italian research libraries on Retroconversion of Catalogues. Biblioteca Nazionale Centrale, Firenze, 22. October 1993.

Papers by Niels Erik Wille and Claudia Miconi.

Conference on Retroconversion of Catalogues, Napoli, 19-20 May 1994. Organised by the FACIT and MORE projects in collaboration with the Napoli section of the Italian Library association.

Papers by Niels Erik Wille, Claudia Miconi and Vera Valitutto.

Conference on Library Networking in Europe. 12.-14. October 1994 in Bruxelles. Organized by EFLC in collaboration with CEC DG XIII.

Papers by Giuseppe Vitiello and Niels Erik Wille on Retroconversion of Library Catalogues and the FACIT project.

Digital Imaging. Concertation Meeting under the Libraries Programme. DG XIII. Luxembourg, 7. November 1994.

Paper on the problems of character representation in Retroconversion by Niels Erik Wille.

New Methods in Retroconversion of Library Catalogues. Conference organized by the Greek Library Association in collaboration with the MORE project. Athen, 18. nov. 1994.

Papers by Niels Erik Wille, on the FACIT prototype, and Joanna Demopolous on Optical Character Recognition of Greek characters.

Articles etc.

Niels Erik Wille: Retroconversion of typewritten or printed catalogues. An introduction to the FACIT project. Statens Bibliotekstjeneste. Kbh. 1993. 28 pp. 2. ed. June 1994. 24 pp.

Niels Erik Wille: "Retroconvertire con lo scanner. Un'introduzione al progetto FACIT." In: Bollettino AIB. Rivista italiana di biblioteconomia e scienze dell'informazione. Vol. 33 no. 4. Dec. 1993. p. 467 - 474. [Translated by Giuseppe Vitiello.]

Claudia Miconi & Gian Luca Corradi: "Politiche di retroconversione e ricerche sperimentale nel campo della scannerizzazione alla Biblioteca nazionale centrale di Firenze " In: Bollettino AIB. Rivista italiana di biblioteconomia e scienze dell'informazione. Vol. 33 no. 4. Dec. 1993. p. 475 - 478.

Guiseppe Vitiello & Niels Erik Wille: "Using scanning for Retrospective Conversion of Catalogues" In: H.P. Geh & M. Walkiers (eds.): Library Networking in Europe. The Proceedings of a European Conference organised by EFLC in cooperation with EBLIDA and LIBER with the support of the European Commission. Bruxelles 12.-14. Oct. 1994. TFPL Publishing. London. 1995. P. 319 - 326.

Niels Erik Wille: "OCR/ICR in Retroconversion of Older Card Catalogues" In: Digital Imaging Proceedings of the Consertation Meeting for Image Processing Projects supported under the Telematics Programme, 7 Nov. 1994. CEC DG XIII E. Luxembourg. 1995. 5 pp.

Niels Erik Wille: "OCR/ICR in Retroconversion of Older Card Catalogues : The FACIT Prototype." In: Vivliothikes kai pilroforisisi 1995; 12,13; 74-80.

J. Tsoutsou-Demopoulos: "FACIT : Stohl kai prooptikes : H Hellenki empiria." In: Vivliothikes kai pilroforisisi 1995; 12,13; 81-83.