The Trinity College Dublin 1872 Online Catalogue

John G. Byrne

Department of Computer Science
O'Reilly Institute
Trinity College
Dublin 2
John.Byrne@cs.tcd.ie

Abstract. The development of an online version of the Trinity College Dublin Printed Catalogue, which list books from the 14th C to 1872, is described. The principal benefit of the system is the ability to search on words and word stems in the title field. As the entries are in at least fourteen languages the language of each Roman script entry was determined, with a success rate of over 90%. The image of the entry from the catalogue is displayed. This hides the OCR errors.

1 Introduction

Most of the books and periodicals in the Library of Trinity College Dublin received up to 1872 are listed in the so called Printed Catalogue [1]. In the author's experience it is not always easy to find the book sought. Recently I was told by a historian interested in railways that the Library did not have Herapath's Railway Magazine. It does have this work which is indexed under MAGAZINE which extends over six pages.

It was the difficulty of looking up the catalogue and the fortuitous finding of some rare books when preparing an exhibition on "Computing before the Electronic Computer" that the idea for this project was conceived. It was carried out by undergraduate [2, 3] and postgraduate students [4, 5] commencing in 1990.

In the following the history of the Library [6] is described briefly. The development and general structure of the catalogue is then described. The major part of the project consisted of five principal steps: scanning and OCR, OCR output correction, natural language recognition, indexing, and the development of the user interface. A demonstration will be given at the Workshop.

2 The Library of Trinity College Dublin

Trinity College Dublin was founded in 1592. The Library surely was started later in that decade. Luke Challoner, who made a major contribution to the development of the College, and James Ussher, one of the first Scholars and Fellows of the College, were sent on book buying expeditions to London, Oxford and Cambridge and by 1610 there were about 4000 books in the Library. Over the years there were a number of major donations. The first major donation was that of Ussher's Library in 1661 by the British House of Commons. He had resigned from the College to accept an appointment as a bishop and he eventually became Archbishop of Armagh. He was a renowned scholar and his library consisted of about 10,000 books and manuscripts. It is

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 17–27, 2004.

stored today in the Long Room on the shelves on which it was originally placed in 1732 when the Library building was eventually completed, construction having started in 1712. In 1743 the library of 13,000 books collected by Claudius Gilbert, a former Vice-Provost and Librarian, was given to the College. The next major acquisition was purchased by the College from Hendrik Fagel, Chief Minister of Holland, in 1802. This consisted of 20,000 books and added considerably to the number of Dutch and other continental books. In 1801, following the Act of Union between Great Britain and Ireland, the Library became a legal deposit library and was able to claim a free copy of every book and periodical published in Great Britain and Ireland, a privilege it retains to this day. Consequently the number of acquisitions increased dramatically and by the 1830s the existing manuscript catalogues were in poor condition. In 1831 James Henthorn Todd, a Hebrew and Irish scholar, was appointed assistant librarian at no salary and he set about improving the cataloguing with a view to preparing a printed catalogue.

3 The Printed Catalogue

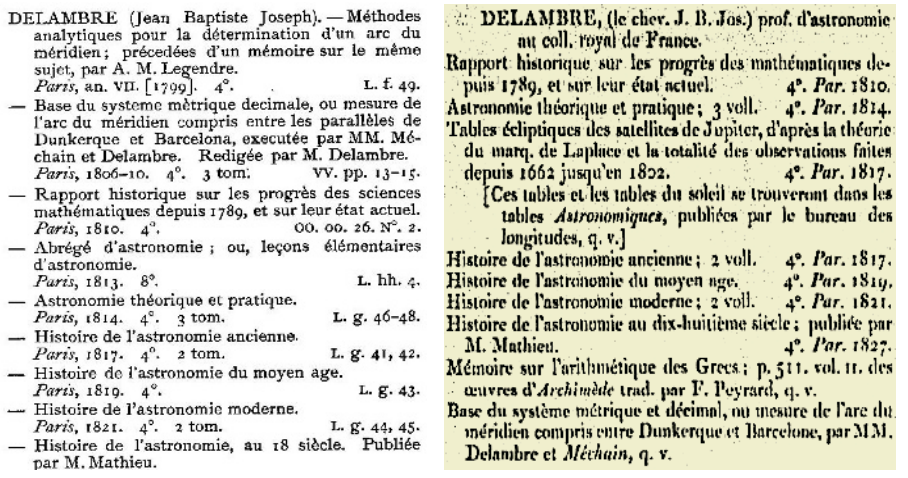
Finding the existing catalogues to be almost useless Todd persuaded the Board of the College to appoint permanent Library Clerks for the first time and set about cataloguing the books on slips of paper in 1835. This work was completed in 1846 and Todd was now in a position to commence printing. He used Bandinel's Catalogue [7] of the Bodleian Library in Oxford as a model but made significant improvements. Unlike Bandinel he decided not to set a date at which he would stop adding books as he knew that printing would take a long time and many books would be unreasonably excluded. As the catalogue was to contain both primary and secondary entries and was to be printed over many years it is possible that secondary entries will be found without primary entries and vice-versa. A *primary* entry which is listed under author (or other keyword such as Ireland) and contains the title, place of publication, date size, edition, no. of volumes etc. and finally as many shelf marks as there are copies of the book. A secondary entry is listed under a different heading and the entry is an abridgement of the title with the primary entry heading printed in capitals. An example of both is shown in Figure 1. A knowledge of Latin was presumed!

CHRISTIUS (Johannes Fridericus).—De Nicolao Machiavello libri tres, in quibus de vita et scriptis, item de secta eius viri atque in universum de politica nostrorum post instauratas litteras temporum . . . ratio habetur.
Lipsiæ et Hal. Magd. 1731. 4°. OO. f. 1.
Fag. G. 1. 43.

— Joh. Frider. CHRISTII de N. Machiavello libri tres.
OO. f. 1.

Fig. 1. A primary and secondary entry.

The imprint in the primary entry starts at the beginning of a new line unlike the older catalogues. Fig. 2 shows how effective this is compared with the Bodleian Catalogue of 1843. Note that an elongated hyphen — is placed at the start of each separate entry in the TCD Catalogue unlike the Bodleian Catalogue where it is only used for



DELAMBRE (Jean Baptiste Joseph). — Méthodes analytiques pour la détermination d'un arc du méridien; précedées d'un mémoire sur le même sujet, par A. M. Legendre.
Paris, an. VII. [1799]. 4°. L. f. 49.
— Base du système mètrique decimale, ou mesure de l'arc du méridien compris entre les parallèles de Dunkerque et Barcelona, executée par MM. Méchain et Delambre. Redigée par M. Delambre.
Paris, 1806–10. 4°. 3 tom. VV. pp. 13–15.
— Rapport historique sur les progrès des sciences mathématiques depuis 1789, et sur leur état actuel.
Paris, 1810. 4°. 00. 00. 26. N°. 2.
— Abrégé d'astronomie; ou, leçons élémentaires d'astronomie.
Paris, 1813. 8°. L. hh. 4.
— Astronomie théorique et pratique.
Paris, 1814. 4°. 3 tom. L. g. 46–48.
— Histoire de l'astronomie ancienne.
Paris, 1817. 4°. 2 tom. L. g. 41, 42.
— Histoire de l'astronomie du moyen age.
Paris, 1819. 4°. L. g. 43.
— Histoire de l'astronomie moderne.
Paris, 1821. 4°. 2 tom. L. g. 44, 45.
— Histoire de l'astronomie, au 18 siècle. Publiée par M. Mathieu.

DELAMBRE, (le chev. J. B. Jos.) prof. d'astronomie au coll. royal de France.
Rapport historique sur les progrès des mathématiques depuis 1789, et sur leur état actuel. 4°. *Par.* 1810.
Astronomie théorique et pratique; 3 voll. 4°. *Par.* 1814.
Tables écliptiques des satellites de Jupiter, d'après la théorie du marq. de Laplace et la totalité des observations faites depuis 1662 jusqu'en 1822. 4°. *Par.* 1817.
[Ces tables et les tables du soleil se trouveront dans les tables *Astronomiques*, publiées par le bureau des longitudes, q. v.]
Histoire de l'astronomie ancienne; 2 voll. 4°. *Par.* 1817.
Histoire de l'astronomie du moyen age. 4°. *Par.* 1819.
Histoire de l'astronomie moderne; 2 voll. 4°. *Par.* 1821.
Histoire de l'astronomie au dix-huitième siècle; publiée par M. Mathieu. 4°. *Par.* 1827.
Mémoire sur l'arithmétique des Grecs; p. 511. vol. II. des œuvres d'*Archimède* trad. par F. Peyrard, q. v.
Base du système métrique et décimal, ou mesure de l'arc du méridien compris entre Dunkerque et Barcelone, par MM. Delambre et *Méchain*, q. v.

Fig. 2. (a) 1872 Printed Catalogue on the left (b) Bodleian Catalogue on the right.

different editions of the same work. Todd claimed, with justification, that his arrangement was advantageous and made it easier for the eye to run down the page when trying to find a book. The Printed Catalogue [1] is a finding-catalogue and Todd did not aim to produce a full descriptive catalogue. A shelf number is included in both the primary and in the principal type of secondary entry. This latter is very convenient, especially when a secondary appeared before a primary. Note that Bandinel's Bodleian Catalogue did not include a shelf mark. These were written in by hand. Todd realised that the position of some books may change and indeed they have but the vast majority are in the same place. Examples of the five types of shelf mark are shown in Figure 3.

Long Room: L. f. 8,9
Long Room Gallery: Gall. MM. 6. 32
East and West End Gallery: Gall. 3. f. 34.
Fagel: Fag. L. 2. 15.
Quin N° 123

Fig. 3. Examples of shelf marks.

The A-B volume, together with a Supplement, was in the printer's hands from 1849 to 1862 and was published in 1864. Todd, who had been appointed Librarian in 1852, died in 1869 before the other volumes were printed. Henry Dix Hutton was given the task of editing the remaining seven volumes and a supplement which also contained an addenda and corrigenda. This has been made available on the Workshop website. The T-Z volume was printed in 1885 and the Supplement in 1887. The whole project took 52 years. The 5121 pages of one set of the eight volumes were separated in 1987 in order to make a microfiche copy and these pages were used in the project described in this paper. There are about 250,000 entries in the catalogue.

The catalogue contains entries in at least fourteen languages. English and Latin occur most frequently and other languages in the Roman alphabet include French, Italian, Spanish, Portuguese, German, Dutch, Danish, Norwegian, Swedish and Irish.

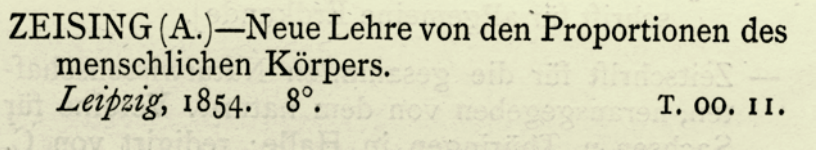
There are also entries in non-Roman alphabets: Greek, Hebrew, Cyrillic, Arabic and Syriac. If an author's works are mainly in Latin the Latin version of the name was used. For example 'KEPLER' is listed under 'KEPPLERUS'. It is similar for other languages. 'PETRARCH' is listed under 'PETRARCA'.

4 Optical Character Recognition

Several OCR packages were tested at the time (1990) but none were very satisfactory. As the catalogue pages had a clear structure as shown in Fig. 2(a) it was decided to make use of this and write our own software which is described by Anderson [4] in his M.Sc. thesis. Each entry has a clear starting point, either a string of capitals followed by an elongated hyphen or just an elongated hyphen. Consequently each entry can be easily extracted. Each page consists of about 4800 characters, excluding spaces, and each line has 48 character positions. There are 84 lines in each column, including the blank lines before each new author.

The objective of the OCR and associated processing is to extract each entry from the image and to transform it into the form shown in Fig. 4(b). Template matching was used. The templates were selected by a human operator and there are more than 300 of them. However using a state machine shown in Fig. 5 during the recognition process it was possible to split the templates into 8 sets corresponding to the A, F, E, T, C, D, L fields in Fig. 4(b) and the Series field. This speeded up the process considerably and probably improved the recognition rate. The number of templates matched can be reduced, by about 70%, by filtering based on the width and height of a character. Of course the segmentation problem was difficult and some progress was made. The OCR program was written in C and initially run on the now extinct Inmos T800 transputer. The success rate was 96% on a small sample of 9 pages. It will be appreciated that counting errors is very tedious. As there are about 4800 characters on each page this indicates that there are about 240 errors per page.

(a)


ZEISING (A.)—Neue Lehre von den Proportionen des menschlichen Körpers.
Leipzig, 1854. 8°. T. oo. 11.

(b)

P: 1130R
H: ZEISING(A.)
A: ZEISING
F: A.
E:
T: —Neue Lehre von den Proportionen des menslichen Korpers
C: Leipszig
PR:
D: 1854
X:
L: GERMAN
S: T. oo. 11.

Fig. 4. (a) 1872 Printed Catalogue entry (b) Entry after processing.

Trying to manually edit 5000 pages with 240 errors is out of the question and consequently an OCR output corrector was written and is described in the next section. The scanning of the 5121 pages was done in the summer of 1991 using a Microtek 3600 legal size scanner. Improved OCR software was transferred back to a PC. As the OCR was slow this was carried out simultaneously on multiple PCs.

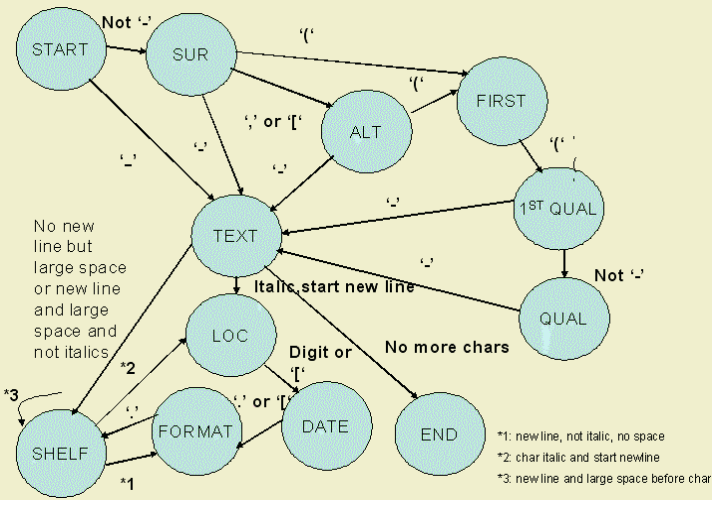


Fig. 5. State machine.

Table 1. Percentage OCR errors from a 50 page set.

Substitution	57.1
Insertion	4.64
Deletion	1.25
Transposition	0.06
Run-on words	6.8
Split words	1.8
Noise	1.0
'ir1' to 'ri'	4.28
'1 o' errors	6.32
Dot within word	2.26
Number substitution	2.74
Wrong Guesses	3.64
Indecipherable	7.64
Miscellaneous	0.47

5 OCROC – An OCR Output Corrector

Because of the numerous errors on each page it was decided to write an OCR Output Corrector (OCROC) which would be as automatic as possible and would facilitate interactive correction. The OCROC program [2] was written in the Spitbol [8] implementation of Snobol 4. A further error analysis was carried out on a set of 50 pages. This yielded 1,364 errors or 273 errors per page which is somewhat more than on the

earlier sample of nine pages. The recognition rate is about 94%. Table 1 lists the percentage occurrence of the OCR errors. About 16% of the errors were language independent and could be corrected automatically. Examples are: 'ir1' -> 'ri' which occurred in 'prr1ests' corrected to 'priests' and 'chir1stophorum' corrected to 'christophorum'. To attempt to correct the remaining errors it was desirable to know the language of the title. It was not feasible to use a conventional dictionary as many of the words would not be found in it. It was found that the number of word types increased approximately linearly with the number of word tokens. Zipf's Law [9] does not hold for words in the catalogue.

5.1 Natural Language Recognition

Suitable dictionaries for the languages used in the catalogue were not available a priori. The entries contain proper nouns, archaisms and variations in dialect and some of the titles are quite short. Initially it was decided to use unique function words for each of the languages. English, Latin, French, Spanish, Italian, German, and Dutch were initially chosen. It quickly became clear that function words alone would not be effective. For example, consider:

Pratique de la geometrie

'de' and 'la' both occur in more than one language so this would be unrecognised. Inflectional morphemes such as '-ique' were then included in the language tables. This method of recognition only yielded a success rate of 70% when tested on 100 pages. The recognition rate improved as more words and morphemes were added but saturation was reached at 70 entries. Improved results were obtained by allowing duplicate function words and morphemes and by assigning a score of 1 for each language found and a success rate of 88% was achieved on the same 100 pages. Analysing the effect of function words alone and suffixes alone it was found that the latter were less effective so a weighting scheme which favoured function words was introduced. It was found that assigning a weight of 2 to a unique function word in the title and 1 to suffixes and words which occurred in more than language was best. A success rate of 90.35% was achieved on the 100 pages [5]. 6.06% of the entries were marked as *unrecognised.* Table 2 shows the result for the title:

trad. en Franc. par Iaques de Miggrede

Table 2. Recognition of Language of Title.

Word or suffix	Language(s)	Weight
En	Dutch, French	1
Par	French	2
De	Spanish, Dutch, Italian, French, Portuguese	1
-ede	Dutch	1

The score for French is 4 and for Dutch 3. The unweighted algorithm would have been unable to differentiate between French and Dutch. Now having the language of 90.35% of the entries correctly recognised it was possible to build up dictionaries from the titles themselves and a success rate 99.3% was obtained.

Knowing the language of the entries enables a search to be made based on the language. It also enabled errors in those languages to be discovered. An approach is described in Nic Gerailt and Byrne [10]. This used trigrams generated for English, French and Latin from corpora obtained from the Internet, a dictionary built up from the catalogue from correct words and heuristic rules. Ignoring split words which would have to be processed in other ways over 80% of incorrect words were detected.

6 Indexing

The entries are indexed on eight fields: Author, forenames, author description, title, place of publication, date, series and shelf number. Two and three letter words and the first four characters of longer words are indexed. These are used to compute the hash key. The complete word is stored in the file. This allows for both stem searches based on four or more characters including complete words. An '*' is used as a wildcard and sometimes this can overcome language problems as well as OCR errors. A search using 'math*' will retrieve entries in English, French, Italian, German and Latin. But looking for all books on houses presents a difficulty. 'House' is 'maison' in French, 'haus' in German, 'casa' in Italian and Spanish.

There is a separate index file for each of the fields which are quite different in nature. Hash coding is used to create the index on disk and the formula is based on that used by Smith and Devine [11] in the Queen's University Belfast Microbird System. It is a radix transformation function based on the radix 7. The formula is:

$$\text{Key} = \text{Integer}(\Sigma(\text{letter}(i)-20)*7\text{^}(3-i),\ i=0 \text{ to } 3/\text{blocksize})+1$$

Radix transformation aims to create a random distribution of keys from a clustered and non-uniform set of keys. This cannot be completely achieved and there is provision for overflow. The separate index file for each field allowed for more efficient use of disk space and it enables convenient searching in described section 7.

7 The User Interface

The User Interface was built using the Delphi GUI builder and is shown in Fig. 6 which shows the result of a search for the works of Pertrarch which will be discussed below. The principal objective of the interface was that it must be easy to learn, easy to use and easy to remember how to use otherwise it would not be used by non-computer literate reader. The entries in the Catalogue are very diverse and readers consequently would come from a wide variety of backgrounds. Pull-down menus are used which cover all options. The options which are likely to be used most frequently are indicated by the icons on the bottom of the screen. The function of some of these icons is described in the following sections.

7.1 Search on Author

The form is shown in Fig. 7. In searching for an author it is possible to use just a surname but in the case of a common name such as 'SMITH' it is not of much use. A first name may also be used. Many author headings include a description of the author such as Duke, Bishop, S.J. etc. As usual the descriptions in the Catalogue are in the

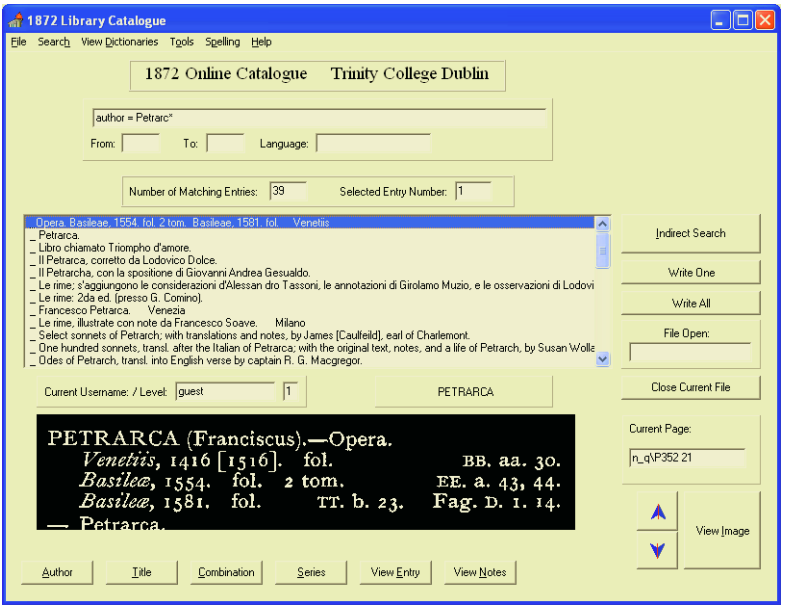


Fig. 6. The User Interface.

language of the entry. Many of the initials are expanded by OCROC to make them easier to search and all are in English. For example 'S.J.' is expanded to Jesuit but it is also possible to search on the abbreviations. A search on 'S.J.' yields 1947 entries. If a full author name yields no results it is possible to use a wild card. For example a user might be unsure whether 'PETRARCH' is in English, Latin or Italian. It is actually in Italian (Petrarca) but it is wise to use 'PETRARC*'. This yields the screen shown in Fig. 6 above. The search produced 39 entries. The selected title is listed in the box in the middle of the screen. The top entry has been selected and the actual entry from the Catalogue is displayed in the box below. This is intended to give confidence to the user and it has the major benefit of hiding OCR errors, if any! The actual post-processed OCR entry is shown in Fig. 8. This is not correctly laid out but it does contain all the information.

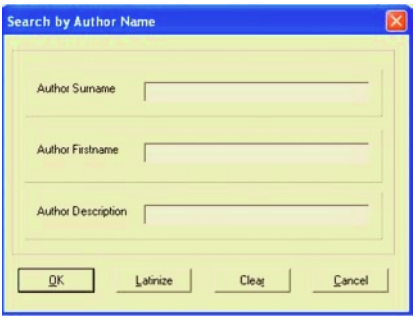


Fig. 7. Search by author name.

H: PETRARCA (Franciscus)
A: PETRARCA
F: Franciscus
E: .
T: _Opera.
Basileae, 1554. fol. 2 tom.
Basileae, 1581. fol.
C: Venetiis
PR:
D: 1416
X:
L: LATIN
S: Fag. D. 1. 14.BB. aa. 30.EE. a. 43, 44.TT. b. 23.

Fig. 8. The postprocessed entry for the first Petrarca entry.

As many of the author names are in Latin it is possible to latinize a name after an unsuccessful search. A search on 'CARDANO' is unsuccessful but the latinized version 'CARDANUS' yields 23 entries. There are sometimes secondary entries for author names. A search on 'JOHN NAPIER' gives the response:

Vid. Joannes NEPERUS

The 'Indirect Search' will place NEPERUS in the author name field. The resulting search yields 9 entries including his books on logarithms.

7.2 Search on Title

By far the most useful field to search on is the title field. The form is shown in Figure 9. As is clear from the examples shown words are implicitly combined with a logical 'and'. The use of the wildcard is very important as it allows four letter stems to be used. This is particularly important for entries under A and B as the font used in this volume is different from that used in the other volumes and the OCR is poor. The A-B volume was printed 19 years before the first of the others.

Apart from overcoming the deficiencies of the OCR searching on words from the title is the principal advantage of the 1872 Online Catalogue. The Herapath magazine mentioned in the opening paragraph was found in this way and it is clearly the only realistic way to look for items listed under terms such as 'IRELAND' (169 occurrences), 'PARLIAMENT' (843 occurrences), 'MAGAZINE' (246 occurrences), 'JOURNAL' (1066 occurrences) and 'BIBLIA' (803 occurrences).

7.3 Combination Search

A Boolean search query may be formulated using the Combination form shown in Fig. 10.

Recently a small collection of Syriac manuscripts have been fully catalogued. To find books in Latin with 'TITLE= Syriac' yields 116 responses. To reduce this to books about Codices the search shown in Fig. 10 was made. It yielded six books.

The date range can be specified. A search on 'TITLE=AGRICU*' from 1700 to 1799 yielded 7 books one of which was wrong due to an error in the date recognition. A search on 'DATE =1470' returns 8 books. There is one older book in the Library (1469) but there is no date in the catalogue entry for it.

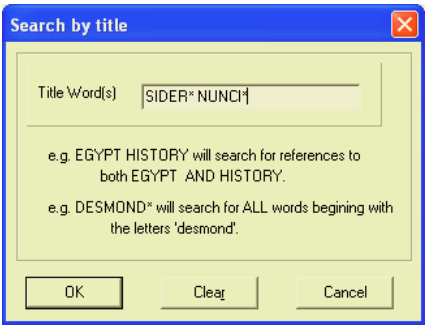


Fig. 9. Title searching form.

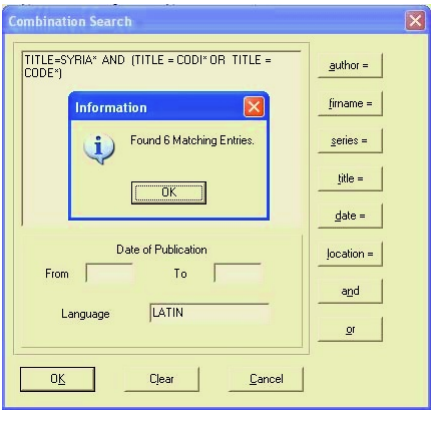


Fig. 10. Combination search showing a search for books about Syriac codices.

7.4 Other Features of the User Interface

If the incorrect image appears it is possible to bring up the OCR entry as shown, for example, in Fig. 4(b) by clicking on ‘View Entry’. Sometimes all of the relevant image may not appear. Clicking on ‘View Image’ displays a larger surrounding image. The arrows to the right of ‘View Image’ allow one to display preceding and following images.

8 Conclusion

The system described in this paper provides access to the Catalogue of books in the Old Library in a way not hitherto available e.g. one can search on words in the title. The OCR is not perfect but this has been overcome to a significant extent by automatic correction of errors (OCROC), by being able to search on four letter stems and by providing a variety of methods of searching and viewing images. The display of the entry image helps to give confidence to the user and gives the right shelf number, especially when this has been overwritten in manuscript. Even the A-B entries, for which the OCR is particularly bad, can often be discovered by approaching the search

in several ways. The system is used on a daily basis in the Department of Early Printed Books. It requires about 2 GB of disk space and retrieval performance is very satisfactory even on an Intel 386 PC. A zip file of images of the pages of the Supplement can be found at http://www.cs.tcd.ie/John.Byrne/

Acknowledgements

The project described in this paper was undertaken by Glynn Anderson, Brendan Culligan, Ruth Clarke, Donnla Nic Gerailt and Mark Fynes under the supervision of the author. Equipment support from IBM (Ireland) is also acknowledged. The catalogue pages were provided by the Keeper of Early Printed Books, Dr. Charles Benson.

References

1. Catalogus Librorum Impressorum qui in Bibliotheca Collegii Sacrosanctae et Individuae Trinitatis, Reginae Elizabethae, juxta Dublin. 9 vols. 1864-1886
2. Clarke, R. M. OCROC Optical Character Recognition Output Corrector. Final year project Trinity College Dublin, May 1993
3. Culligan, B. T. Design of an On-Line Database Query System for the 1872 Printed Catalogue. Final year project Trinity College Dublin, May 1993
4. Anderson, Glynn. Computerising a Library Catalogue using Optical Character Recognition. M.Sc. thesis, University of Dublin, 1992
5. Clarke, R. M. User-Oriented Access to a Multilingual Database. M.Sc. thesis, University of Dublin, 1995
6. Kinane, Vincent and Walsh, Anne (Eds.). Essays on the History of the Trinity College Library Dublin. Dublin: Four Courts Press, 2000
7. Bandinel, B. Catalogus Librorum Impressorum in Bibliotheca Bodleiana, Oxford 1843
8. Emmer, Mark B., Quillen, Edward K., Dewar, Robert B. K. MACRO SPITBOL The High-Performance SNOBOL Language. Catspaw Inc., 1991
9. Zipf, G. K. HumanBehaviour and the Principle of Least Effort. Addison Wesley, 1949
10. Nic Gerailt, D. Byrne, J. G. Error Detection in Several Languages for an OCR-Generated Multilingual Database. Proc. Third International Workshop on Applications of Natural Language to Information Systems. June 26-27 Simon Fraser University, Canada, 1997
11. Smith, F.J. and Devine, K. BIRD, QUILL and MicroBIRD - A successful family of text retrieval systems. Literary and Linguistic Computing, Vol 4, No 2, pp 115-120, 1989.