ED 215 680	IR 010 143
AUTHOR	O'Neill, Edward T.; Aluri, Rao
TITLE	A Method for Correcting Typographical Errors in Subject Headings in OCLC Records. Research Report.
INSTITUTION	OCLC Online Computer Library Center, Inc., Dublin, Ohio.
REPORT NO	OCLC/OPR/RR-80/3
PUB DATE	15 Oct 80
NOTE	31p.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Algorithms; *Cataloging; Databases; Information Retrieval; *Online Systems; *Subject Index Terms
IDENTIFIERS	*Authority Files; *Error Detection; Library of Congress Subject Headings; OCLC

ABSTRACT

The error-correcting algorithm described was constructed to examine subject headings in online catalog records for common errors such as omission, addition, substitution, and transposition errors, and to make needed changes. Essentially, the algorithm searches the authority file for a record whose primary key exactly matches the test key. If an exact match is not found, the algorithm identifies records in the authority file, first with the same initial characters, or if that is unsuccessful, with similar endings. The heading is then examined to see if by making simple changes, it can be modified to match a valid record in the authority file. If no match can be found, even after modification, it is then assumed that the heading is one of questionable validity--being either a valid heading with no corresponding record in the author file or an invalid heading containing extensive errors. The algorithm separates the subject headings into groups of valid headings, corrected headings, and questionable headings that require manual examination. Provided are one table, five figures, and 21 references. (Author/RBF)

Report Number: OCLC/OPR/RR-80/3
Date: 1980 October 15

Research Report

on

A Method for Correcting Typographical Errors in Subject Headings in OCLC Records

by

Edward T. O'Neill
Rao Aluri

OCLC, Inc.
Office of Planning & Research
Research Department
1125 Kinnear Road
Columbus, Ohio 43212

The Research Report Series is OCLC's formal dissemination vehicle through which OCLC research project results can be made public.

The Research Department's Research Reports are published and distributed through OCLC, Inc., until the report is available from ERIC (approximately six to nine months after publication). A maximum of three copies of any single Research Report is provided at no charge to interested persons and institutions. Requests for Research Reports should be directed to OCLC, Inc., User Services Division, Installation Services Section, 1125 Kinnear Road, Columbus, Ohio 43212.

Research Report Series

Subject Heading Patterns in OCLC Monographic Records
by E. O'Neill and R. Aluri
Report Number: OCLC/RDD/RR-79/1; ERIC ED 183 167

An Overview of a Proposed Monitoring Facility for the Large-scale, Network-based OCLC On-line System
by W. Dominick, D. Penniman, and J. Rush
Report Number: OCLC/OPR/RR-80/1; ERIC ED 186 042

Analytical Review of Catalog Use Studies
by K. Markey
Report Number: OCLC/OPR/RR-80/2; ERIC ED 186 041

A Method for Correcting Typographical Errors in Subject Headings in OCLC Records
by E. O'Neill and R. Aluri
Report Number: OCLC/OPR/RR-80/3

ABSTRACT

This report describes an error-correcting algorithm that examines the subject headings in catalog records for common errors such as omission, addition, substitution, and transposition errors. If such errors are identified, the algorithm makes the needed corrections. The algorithm requires a subject heading authority file.

The subject heading authority file contains records representing valid subject headings. Each authority file record contains the subject heading, its primary key, and its reverse key. The primary key is derived from the subject heading by taking the initial letters or digits from the heading. The reverse key is formed by taking the last letters or digits, in reverse order, from the subject heading.

The error-correcting algorithm starts with a test subject heading whose validity is to be established. The subject heading under consideration will belong to one of the following classes: (1) valid subject heading which is included in the authority file; (2) valid subject heading which is not included in the authority file; and (3) invalid subject heading. The error-correcting algorithm derives the primary key of the test subject heading and searches the authority file for a record whose primary key matches exactly with that of the test key. If an exact match is found in the authority file, the test heading is assumed to be correct. If an exact match is not found, the algorithm identifies records from the authority file whose primary keys have the same initial characters as that of the test subject heading. The heading is then examined to see if, by making simple changes, it can be modified to match one of the valid records in the authority file. If modification does not produce a match, it is assumed that the error lies in the initial set of characters of the heading. Using the reverse key, the algorithm compares the heading to authority file records with similar endings. If no match can be found, even after modification, it is then assumed that the heading is one of questionable validity -- being either a valid heading with no corresponding record in the authority file or an invalid heading containing extensive errors. The algorithm separates the subject headings into groups of valid headings, corrected headings, and questionable headings that require manual examination.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of OCLC, Inc., and the numerous individuals associated with OCLC whose help was essential to the completion of this project. Specifically, the authors would like to thank Dr. James Rush, Dr. David Penniman, and Dr. Neal Kaske for their encouragement and administrative support of the project; Dr. Thomas Hickey for his extensive assistance in accessing the data base and in using the OCLC computer facilities; Mr. Robert Rose for his assistance in program development; Mr. Carl Anderson for his advice on subject cataloging; Ms. Peggy Zimbeck for her extensive editorial assistance; and Ms. Beckie Purdy for assistance in preparing the manuscript.

NOTE ABOUT THE AUTHORS

Dr. Edward T. O'Neill held a one-year appointment for 1978-1979 in the Research Department as OCLC's first Visiting Distinguished Scholar. During that time, he was on sabbatical leave from the School of Information and Library Studies at the State University of New York at Buffalo where he was an Associate Professor. Dr. O'Neill received his Bachelor of Arts degree from Albion College and his Bachelor of Science, Master cf Science, and Doctor of Philosophy degrees from Purdue University. After completing his graduate work, he joined the faculty at the State University of New York at Buffalo where he has held a variety of appointments, including Acting Dean and Assistant Dean of School of Information and Library Studies. In 1980 September, he joined the faculty at Case Western Reserve University as Dean and Professor of Library Science.

Mr. Rao Aluri was a Research Assistant in the Visiting Distinguished Scholar Program at OCLC during 1978-1979. During that time, he was a doctoral candidate in the Cooperative Doctoral Program of the Department of Higher Education and the School of Information and Library Studies, at the State University of New York at Buffalo. Mr. Aluri received his Bachelor of Science degree from Andhra University, Waltair, India, and his Master of Library Science degree from the University of Western Ontario, London, Ontario, Canada. Before starting his doctoral work, he was a reference librarian at the University of Nebraska at Omaha. In 1980 September, he joined the faculty of Emory University as an Assistant Professor of Library Science.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Page

Table

Figure

I. INTRODUCTION

The presence of errors in the on-line union catalogs of bibliographic utilities such as OCLC has an adverse effect on the utilities themselves and on the end users of their data bases. Bourne's analysis of the impact of spelling errors, although he was writing from the context of commercial bibliographic search systems such as SDC and BRS, is still valid for on-line union catalogs. [1] For the bibliographic utilities, the negative consequences are, following Bourne: "(1) extra computer time, storage space, and associated costs...; (2) damage to image/credibility/marketability [of the bibliographic utilities]; [and] (3) less effective service than is otherwise possible." The end users of the catalog records are forced to divert some of their resources in terms of personnel, time, and communication costs, to "cleaning up" the records before they can be used.

For OCLC users, errors in the subject heading fields currently are only a minor nuisance that can be overcome either by editing the record before producing cards or by simply ignoring the errors when the subject cards are filed. However, in the near future when computerized systems play a major role in providing subject access, these errors will have to be taken into consideration. Computers are not as forgiving of errors as are humans. With computerized subject access, errors in the subject heading fields will frequently result in the records being inaccessable and, depending on the retrieval technique employed, may make the search much more difficult. In view of these problems, bibliographic utilities have to devise effective and efficient means of identifying and correcting common errors in the subject headings.

A. Scope of Study

The algorithm described here is a by-product of a research project on the Library of Congress subject headings reported by O'Neill and Aluri. [2] The original project was undertaken to examine the distribution patterns of LC subject headings in the OCLC catalog records and to study the information content of subject headings assigned. During the course of this project, however, the presence of numerous misspellings and inconsistent spacing, punctuation, and capitalization practices could not be overlooked. The recognition of the need for correcting such common errors led to the design of the proposed error-correcting algorithm.

The original project on LC subject headings was conducted on a sample of 33,455 catalog records in the OCLC data base. The sample contained every full level nonjuvenile monographic record in the OCLC data base whose OCLC control number ended with "96," as of September 2, 1978. Of the 33,455 records in the sample, 7,490 were received from the Library of Congress through its MARC Cataloging Distribution Service and the remaining 25,965 were cataloged on-line by OCLC member libraries. A total of 50,213 subject headings occurred in the sample of which 47,036 were Library of Congress subject headings.

B. Subject Heading Errors

An examination of subject headings extracted from the 33,455 monographic records in the OCLC data base showed that a significant number of the headings contained various types of errors. The majority of the errors was typographical and fell into one of four major categories: [3,4]

(1) omission errors,
(2) addition errors,
(3) substitution errors,
(4) transposition errors.

"Antomy, Human" (instead of "Anatomy, Human") is an example of an omission error where a character was inadvertently dropped. "Geographty" (instead of "Geography") is an example of an addition error where an extraneous character was added. Substitution errors, as in "Hard-cord unemployed" (instead of "Hard-core unemployed"), have one character replaced by an incorrect character. "Commerical law" (instead of "Commercial law") illustrates a transposition error where a pair of characters is transposed. Although some subject headings contained multiple errors involving multiple characters, most of the incorrect subject headings contained only a single error involving a single character or one pair of characters.

The sample of subject headings also contained several spacing, punctuation, and capitalization inconsistencies. Examples are:

U.S.	U. S. (spacing)
Postage-stamps	Postage stamps (punctuation)
Congresses	congresses (capitalization)

While these inconsistencies are relatively unimportant in manual card catalogs, they become significant in a computerized catalog. For instance, computer software treats "O.T." and "O. T." as different character strings and hence as different headings.

Finally, the subject headings sample contained a large number of inconsistent abbreviations. Although, strictly speaking, abbreviations are not errors, the absence of standardization in their use may cause retrieval problems. For example, the subdivision "Description and travel" appeared in the sample in the following forms:

Descr. and trav.
Description and travel
Description & travel
Descr. & trav.
Descr. & travel
Desc. & trav.
Desc. & travel
Desr. & trav.

Each of these strings is distinct to the computer.

Typographical errors, variations in punctuation, and variations in the use of abbreviations become serious problems as the size of the 'ata base increases. It is conservatively estimated that 1% of OCLC records contain errors in their subject heading fields. Assuming the number of errors increases linearly with the size of the data base, one can estimate that when the OCLC data base grows to 10 million records, it could contain 100,000 catalog records with errors in subject headings alone. Many of these records may be inaccessible to the user through normal retrieval mechanisms. Consequently, these typographical and variation errors must be identified and corrected.

C. Objective of the Report

This report addresses the presence of errors and variations, and the need to correct and standardize subject headings for information retrieval. An error-correcting algorithm is described that automatically corrects a large percentage of the typographical errors. In addition, the algorithm identifies subject headings that may contain errors, regardless of their cause.

The proposed error-correcting algorithm is intended to be conservative. That is, it is designed to correct relatively simple errors and to identify complex errors for scrutiny by human editors. At the same time, the algorithm produces a list of corrected subject headings to permit the editors to check that the algorithm is not altering valid headings.

II. ERROR-CORRECTING ALGORITHM

A fairly large body of literature exists on the detection and automatic correction of spelling errors. [5-14] According to Zamora, the techniques described in the literature fall into three categories: "(1) isolation of low frequency words, (2) dictionary look-up, and (3) n-gram analysis, where an n-gram is a string of n characters extracted from a word." [15] The dictionary look-up method is the most appropriate for correcting spelling errors in subject headings in the records of the on-line union catalogs of bibliographic utilities. Here, the dictionary that could be used for subject heading verification and correction is a list of authorized subject headings. Because the on-line union catalogs of the bibliographic utilities are created and maintained through the cooperative efforts of a large number of libraries, there is already a need for the development of various authority lists (such as subject heading and name authority lists). Once such authorized lists are developed, they should logically be used in detecting and correcting errors. In this connection, Zamora's observation that "the dictionary look-up technique has the most favorable ratio of misspellings to words flagged when applied to the CAS (i.e., Chemical Abstracts Service) data base" is encouraging. [16]

Morgan describes the two stages of the error-correcting algorithm of the dictionary look-up technique. [17] In the case of subject headings, Morgan's algorithm begins with two key elements: (1) a test subject heading whose validity is under question, and (2) a valid subject heading that belongs to the authorized list of subject headings. The two basic stages of dictionary look-up technique, as described by Morgan are:

(1) selecting a subset of valid subject headings from the authority list, where the subset of subject headings contains nearly all headings of which the test heading may be a misspelling; and

(2) comparing, pairwise, each of the valid subject headings with the test subject heading to determine whether or not the test heading is a misspelling of the authorized heading.

Following Morgan, the error-correcting algorithm described in this report contains three elements:

(1) creation of a subject heading key corresponding to each of the valid and test subject headings, where the subject heading keys, rather than the subject headings themselves, are used in pairwise comparison between valid and test subject headings;

(2) creation of a subject heading authority file that would be the source of valid subject headings; and

(3) an error-correction routine that selects the subset of valid subject headings against which the test headings are compared, compares the

test subject headings with the valid subject headings for possible errors, and then corrects the test headings, if errors are detected.

The design of the error-correcting algorithm is based on the following observations:

(1) Over 90% of the subject headings in the OCLC records are Library of Congress (LC) subject headings. [18] LC subject headings are controlled vocabulary and form the authority list from which LC and most OCLC member libraries draw the headings for assignment to catalog records. In other words, the predominant use of LC subject headings limits both the number of subject headings and their variations which can occur in the OCLC records.

(2) If the subject headings are arranged according to their frequencies of occurrence, the headings which occur least frequently contain the largest number of typographical errors. [19]

(3) The typographical errors fall into one of the four categories identified in Chapter I of this report: dropped characters, excess characters, characters substituted for others, and pairs of transposed characters.

(4) In a majority of cases, subject headings contain only one error involving only one character or one pair of characters.

The first two observations are used to create an authority file consisting of 'good' subject headings. The last two observations are used to compare potentially erroneous headings with those in the authority file and to correct the errors.

A. Subject Heading Key

Much of the manipulation of subject headings is done on a key constructed from the subject headings. The subject heading key, which contains 28 characters, is made up of one character identifying the type of subject heading followed by 27 characters derived from the heading. Topical subject headings are assigned the identifying character "1," geographic headings the character "2," etc., as shown in Table 1. The derived portion of the key contains the first 27 characters of each subject heading.

Table 1. Types of Subject Headings and Subdivisions Identified in the Keys

First Character of a Key	Type of Subject .Heading or Subdivision
1	Topical Subject Heading
2	Geographic Subject Heading
3	Personal Name Subject Heading
4	Corporate Name Subject Heading
5	Conference/Meeting Subject Heading
6	Uniform Heading Subject Heading
X	General Subdivision
Y	Period Subdivision
Z	Place Subdivision

All letters are capitalized to eliminate the differences caused by variations in capitalization. If the subject heading contains numeric characters, all occurrences of the digit "1" are converted to alphabetic character "L." This is done to compensate for the common confusion between the digit "1" and the lowercase letter "l." The subject heading key ignores special characters, punctuation, spacing, and capitalization to compensate for minor variations in the subject headings. The key thus constructed eliminates a large number of common variations in the subject headings. Therefore, the key can be used to group together many of the variants of a subject heading. Examples of variant forms of subject headings and their keys are:

Greco-Turkish War, 1921-1922.	1GRECOTURKISHWARL92LL922
Greco-turkish war, 1921-1922:	
Greco-turkish War, 1921-1922.	
Greco-turkish War, 1921 - 1922.	

Freedman in Beaufort co., S.C.	1FREEDMANINBEAUFORTCOSC
Freedman in Beaufort Co., S.C.	
Freedman in Beaufort co., S. C.	

The digit "1" in the first-character position in the preceding keys identifies the keys as topical subject headings. Blanks are used at the end to fill the key out to 28 characters. In the case of subject headings containing subdivisions, separate keys for main headings and subdivisions are maintained. For example, if the subject heading is:

"650 b̸Ø English fiction $y 19th century $x History and criticism"

the following three keys are derived:

1ENGLISHFICTION
YL9THCENTURY
XHISTORYANDCRITICISM

As with the main headings, the first character in the subdivision key identifies the type of subdivision. However, for subdivisions, an alphabetical character is used. In the above examples, the "X" and "Y" preceding the second and third keys identify general and period subdivisions respectively. Table 1 shows the identification characters used in the keys and the corresponding types of subject headings or subdivisions.

A reverse key also is derived from the subject heading for use by the error-correcting routine. Reverse keys are made up of the last 14 nonblank characters, in reverse order, from the primary 28-character subject heading key. For example, if a subject heading is "Self-instruction," its primary key would be "1SELFINSTRUCTION" and its reserve key would be "NOITCURTSNIFLE." The primary subject heading key is used to locate subject headings in the authority file and to check for errors in the second half-segment of the heading; the reverse key is used to check for errors in the first half-segment of the heading.

For very long headings, some characters will be dropped in forming the keys. For example, the subject heading "Information storage and retrieval systems" would have as its primary key "1INFORMATIONSTORAGEANDRETRIE" since only the first 27 valid characters from the heading could be used. The length of this key would be 38 since the dropped characters would be counted The reverse key would be formed from the last 14 valid characters from the heading. For this example, the reverse key would then be "SMETSYSLAVEIRT."

B. Authority File

The authority file is created based on the assumption that frequently occurring subject headings will be valid. It is further assumed that the catalog records issued by the LC MARC Distribution Service (LC records) are less likely to have typographical errors than those contributed by the OCLC member libraries (contributed records). Once these assumptions are accepted, 'good' subject headings can be defined as those whose frequency of occurrence in LC and contributed records equals or is greater than an arbitrary number, while, at the same time, occurring at least a set number of times in LC records. All subject headings which satisfy these requirements are placed in the authority file. The more stringent the requirements for inclusion of subject headings in the authority file and the larger the number of subject headings on which the authority file is based, the more confident one can be that these subject headings are free from typographical errors. However, it is not always necessary to create an authority file by this method. For instance, an organization, such as LC, might distribute a machine-readable authority file of subject headings. Such a file, once carefully checked for errors, would also be suitable for this algorithm.

Each record in the authority file consists of information on a main subject heading or subdivision. Figure 1 shows an authority record for the subject heading, "Copyright, International." The first 28 characters in the authority record represent the primary key for the subject heading. The next 14 characters in the record represent the reverse key. Following the reverse key are three fields which show: (1) the number of nonblank characters in the primary key including any dropped characters, (2) the length of the subject heading or subdivision, and (3) a type of record indicator showing whether the record represents an entry for a valid heading or an entry for an abbreviated heading. When the type of record indicator is "Ø," it indicates that the record represents a valid heading. The last field in the authority record contains the actual subject heading or subdivision.

When the type of record indicator is "1," it indicates that the record represents an entry for an abbreviation. Entries for abbreviations act as "see" references. Figure 2 shows the authority record for an abbreviation. The first 28 characters represent the primary key for an abbreviated heading, e.g., "Hist. & Crit." which is a general subdivision. The primary key is followed by three fields as in the case of a valid heading. The type of record indicator, however, is "1," showing that the entry is for an abbreviated heading. The type of record indicator in this case is followed by the primary key for the unabbreviated version of the heading.

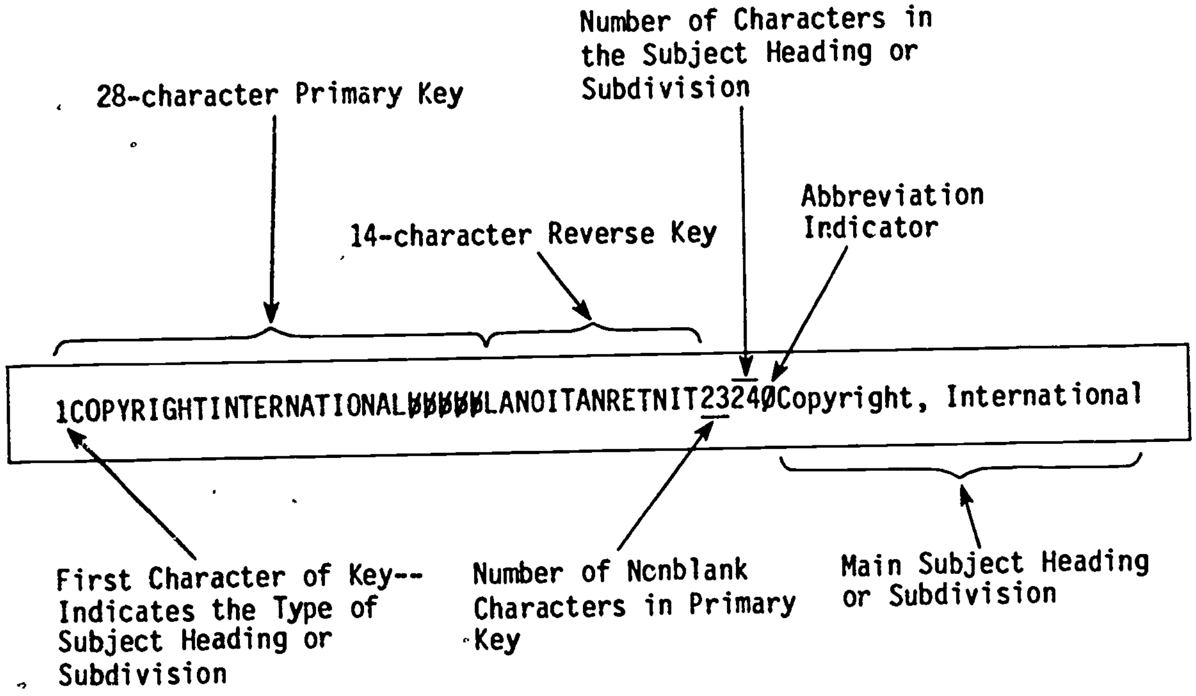

Number of Characters in the Subject Heading or Subdivision
28-character Primary Key
Abbreviation Indicator
14-character Reverse Key
1COPYRIGHTINTERNATIONALƀƀƀƀƀLANOITANRETNIT2324ØCopyright, International
First Character of Key--Indicates the Type of Subject Heading or Subdivision
Number of Nonblank Characters in Primary Key
Main Subject Heading or Subdivision

Figure 1. Authority File Record

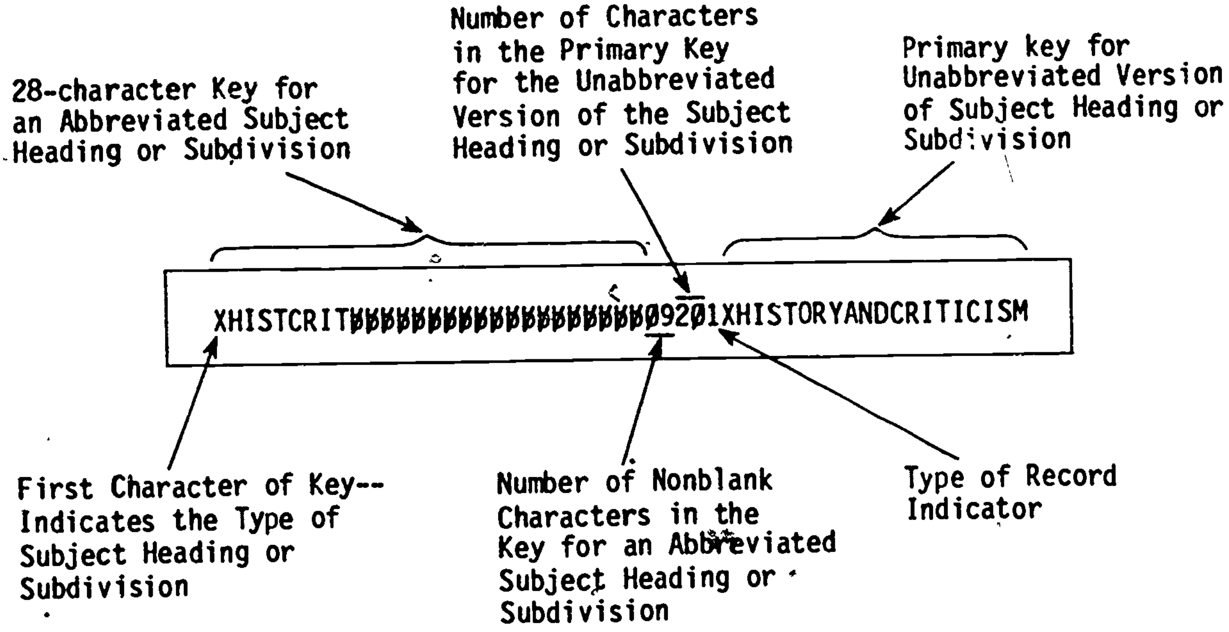

Number of Characters in the Primary Key for the Unabbreviated Version of the Subject Heading or Subdivision
Primary key for Unabbreviated Version of Subject Heading or Subdivision
28-character Key for an Abbreviated Subject Heading or Subdivision
XHISTCRITƀƀƀƀƀƀƀƀƀƀƀƀƀƀƀƀƀƀƀØ92Ø1XHISTORYANDCRITICISM
First Character of Key--Indicates the Type of Subject Heading or Subdivision
Number of Nonblank Characters in the Key for an Abbreviated Subject Heading or Subdivision
Type of Record Indicator

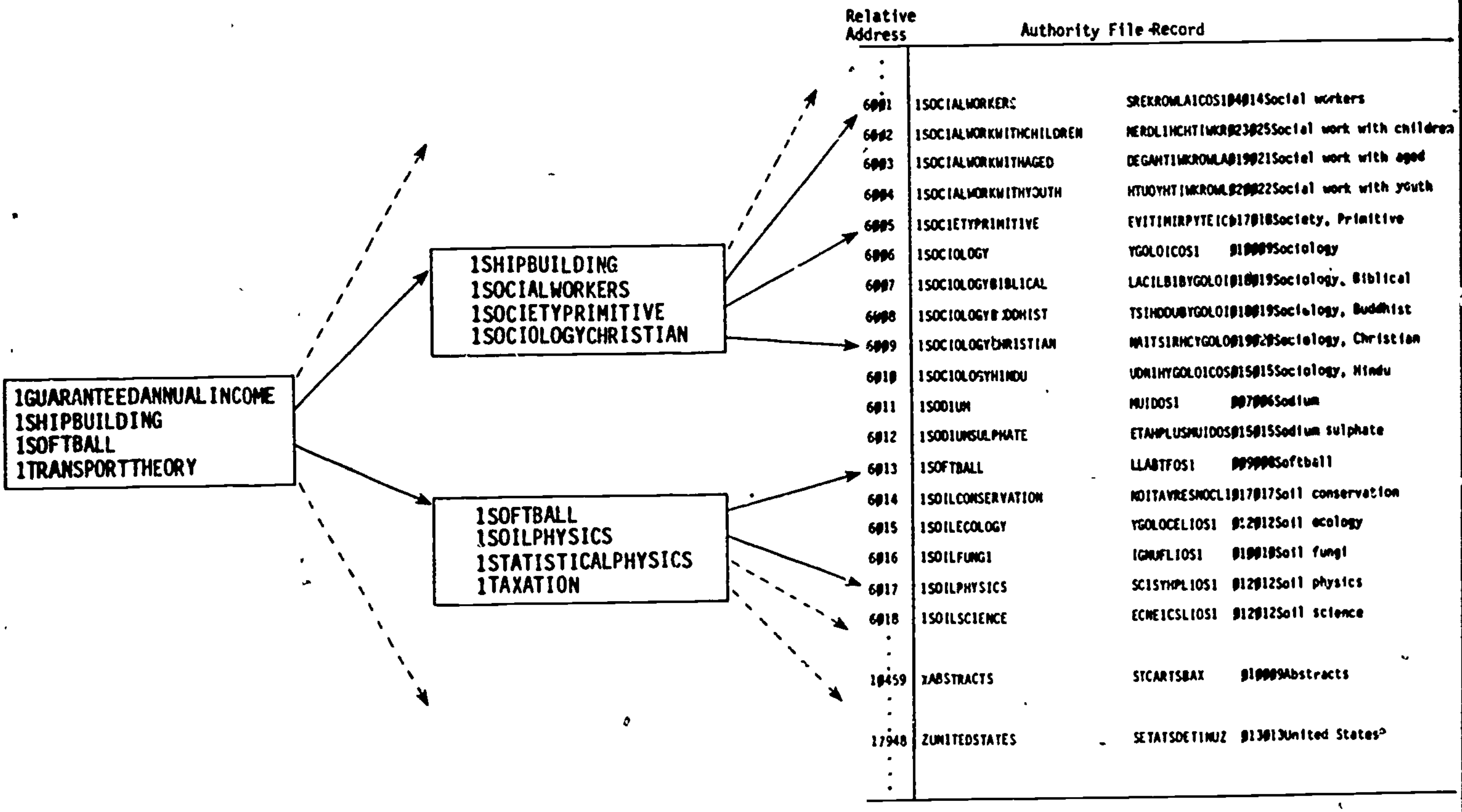
Figure 2. Authority Record for an Abbreviated Subject Heading or Subdivision

The authority records for abbreviations, in contrast to those for valid subject headings and subdivisions, have to be manually introduced into the file. This, however, should not present a serious problem as it is not difficult to compile a list of commonly occurring abbreviations in cataloging records. When the error-correcting algorithm comes across a key for an abbreviated heading, that key can be automatically replaced by the key for the unabbreviated version of the heading.

The authority file is an indexed file sequenced on its primary key (Figure 3). The index sequential organization brings together subject headings and subdivisions starting with the same initial characters. Records in the authority file can be accessed directly using the complete primary key. To determine if a given heading is in the authority file, the key for the heading would be derived. If a record with exactly the same key exists in the authority file, the corresponding record would be retrieved. When no match is found, it would be known that either the heading is new or that the heading is invalid.

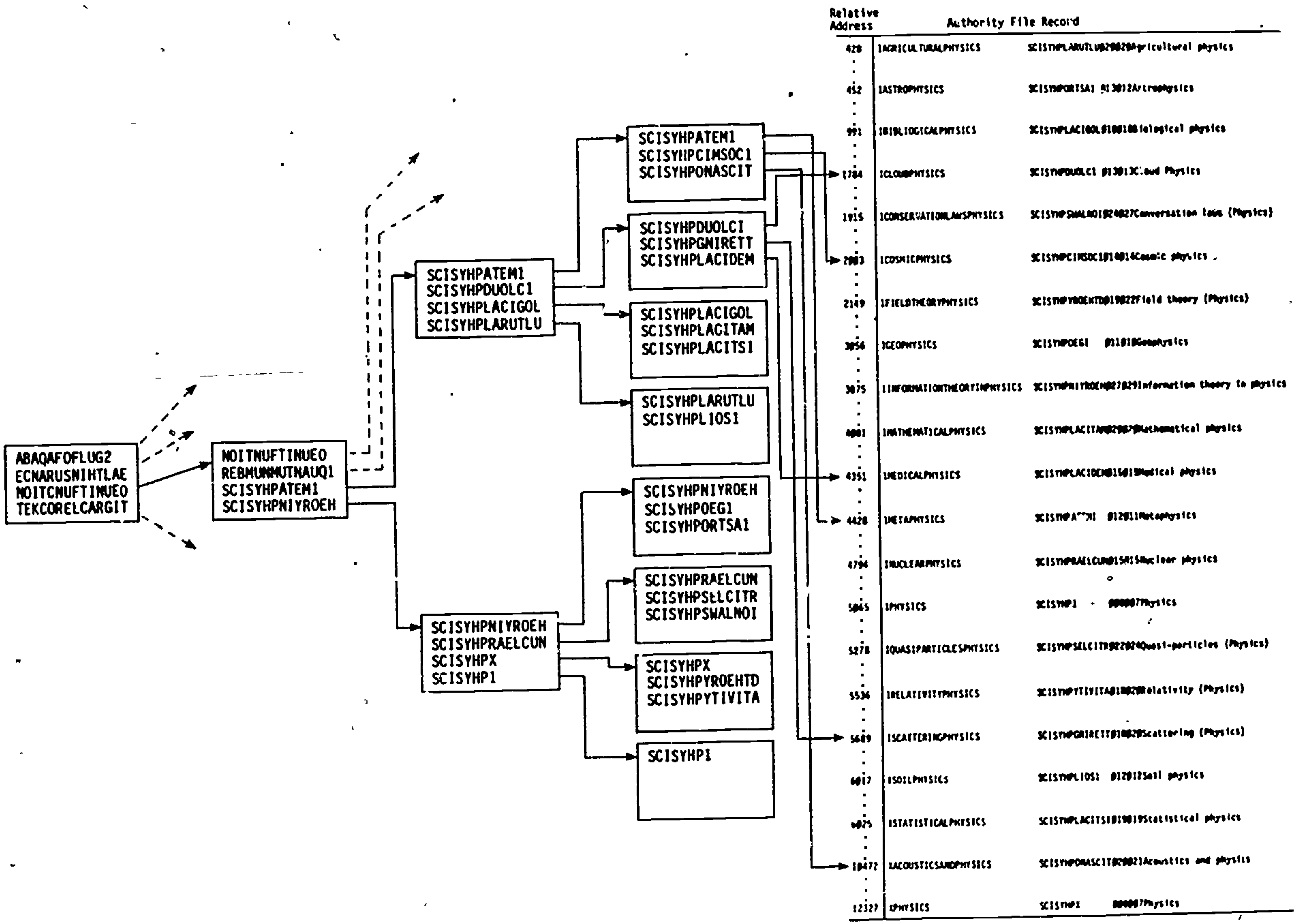
The authority file can also be accessed directly using only the initial portion of the primary key. If the authority file shown in Figure 3 was read using the key "1SODIU," no exact match would be found but the file would be positioned so that subsequent sequential read operations would retrieve the records corresponding to the headings "Sodium," "Sodium sulphate," "Softball," etc.

If only the end of a heading is known, the primary key index is of no assistance. To access the authority file by the reverse key, a second nonsequential index to the authority file is required (Figure 4). This nonsequential index permits access by the reverse key or by any portion of the reverse key. If the authority file shown in Figure 4 is read using the key "SCISYHPD," no exact match would be found but the file would be positioned. Subsequent sequential reads through the nonsequential index would retrieve the authority records for "Cloud physics," "Scattering (Physics)," "Medical physics," etc. When it is assumed that there is an error in the first half of the heading, the nonsequential reverse key index must be used to access the authority file.



Relative Address	Authority File Record	
.		
6ØØ1	1SOCIALWORKERS	SREKROWLAICOS1Ø14Ø14Social workers
6ØØ2	1SOCIALWORKWITHCHILDREN	NERDLIHCHTIWKRØ23Ø25Social work with children
6ØØ3	1SOCIALWORKWITHAGED	DEGAHTIWKROWLAØ19Ø21Social work with aged
6ØØ4	1SOCIALWORKWITHYOUTH	HTUOYHTIWKROWLØ2ØØ22Social work with youth
6ØØ5	1SOCIETYPRIMITIVE	EVITIMIRPYTEICØ17Ø18Society, Primitive
6ØØ6	1SOCIOLOGY	YGOLOICOS1 Ø1ØØØ9Sociology
6ØØ7	1SOCIOLOGYBIBLICAL	LACILBIBYGOLOIØ18Ø19Sociology, Biblical
6ØØ8	1SOCIOLOGYBUDDHIST	TSIHDDUBYGOLOIØ18Ø19Sociology, Buddhist
6ØØ9	1SOCIOLOGYCHRISTIAN	NAITSIRHCYGOLOØ19Ø2ØSociology, Christian
6Ø1Ø	1SOCIOLOSYHINDU	UDNIHYGOLOICOSØ15Ø15Sociology, Hindu
6Ø11	1SODIUM	MUIDOS1 ØØ7ØØ6Sodium
6Ø12	1SODIUMSULPHATE	ETAHPLUSMUIDOSØ15Ø15Sodium sulphate
6Ø13	1SOFTBALL	LLABTFOS1 ØØ9ØØ8Softball
6Ø14	1SOILCONSERVATION	NOITAVRESNOCL1Ø17Ø17Soil conservation
6Ø15	1SOILECOLOGY	YGOLOCELIOS1 Ø12Ø12Soil ecology
6Ø16	1SOILFUNGI	IGNUFLIOS1 Ø1ØØ1ØSoil fungi
6Ø17	1SOILPHYSICS	SCISYHPLIOS1 Ø12Ø12Soil physics
6Ø18	1SOILSCIENCE	ECNEICSLIOS1 Ø12Ø12Soil science
.		
1Ø459	XABSTRACTS	STCARTSBAX Ø1ØØØ9Abstracts
.		
17948	ZUNITEDSTATES	SETATSDETINUZ Ø13Ø13United States
.		

Figure 3. Sequential Index Organized by Primary Key



Relative Address	Authority File Record	
420	1AGRICULTURALPHYSICS	SCISYHPLARUTLU020020Agricultural physics
452	1ASTROPHYSICS	SCISYHPORTSA1 013012Astrophysics
991	1BIOLOGICALPHYSICS	SCISYHPLACIGOL018018Biological physics
1784	1CLOUDPHYSICS	SCISYHPDUOLC1 013013Cloud Physics
1915	1CONSERVATIONLAWSPHYSICS	SCISYHPSWALNOI024027Conservation laws (Physics)
2003	1COSMICPHYSICS	SCISYHPCIMSOC1014014Cosmic physics
2149	1FIELDTHEORYPHYSICS	SCISYHPYROEHTD019022Field theory (Physics)
3056	1GEOPHYSICS	SCISYHPOEG1 010010Geophysics
3875	1INFORMATIONTHEORYINPHYSICS	SCISYHPNIYROEH027029Information theory in physics
4001	1MATHEMATICALPHYSICS	SCISYHPLACITAM020020Mathematical physics
4351	1MEDICALPHYSICS	SCISYHPLACIDEM015015Medical physics
4428	1METAPHYSICS	SCISYHPATEM1 012011Metaphysics
4794	1NUCLEARPHYSICS	SCISYHPRAELCUN015015Nuclear physics
5065	1PHYSICS	SCISYHP1 000007Physics
5278	1QUASIPARTICLESPHYSICS	SCISYHPSELCITR022024Quasi-particles (Physics)
5536	1RELATIVITYPHYSICS	SCISYHPYTIVITA018020Relativity (Physics)
5689	1SCATTERINGPHYSICS	SCISYHPGNIRETT018020Scattering (Physics)
6017	1SOILPHYSICS	SCISYHPLIOS1 012012Soil physics
6025	1STATISTICALPHYSICS	SCISYHPLACITSI019019Statistical physics
10472	1ACOUSTICSANDPHYSICS	SCISYHPONASCIT020021Acoustics and physics
12327	1PHYSICS	SCISYHP1 000007Physics

Figure 4. Nonsequential Index Organized by Reverse Key

C. Error-correcting Procedure

The error-correcting procedure is performed on subject headings in the unchecked subject headings file. This unchecked file includes all those subject headings that must be tested for accuracy. The procedure consists of two operations: (1) check to see if an exact match is found, and (2) if not, make corrections to the headings, if pos.ible.

Using the error-correcting procedure, the subject headings under consideration are compared with those in the authority file to detect and correct errors. If there is an exact match between a subject heading in the unchecked file and one in the authority file, the heading in the unchecked file is accepted as valid. If no match is found, the heading in the unchecked file is examined for errors. If the algorithm fails to find a match between the heading in the unchecked file and those headings in the authority file even after corrections are made for possible typographical errors, the unchecked file heading is transferred to a "questionable" subject headings file for manual review. Figure 5 presents a diagrammatic representation of this procedure.

The subject headings in the unchecked file are matched with those in the authority file through two operations. In the first operation, characters in the first half of the test heading are assumed to be correct and the remaining characters are examined for errors. In the second operation, the process is reversed; the characters in the second half of the heading are assumed to be correct and the first half of the heading is checked for errors.

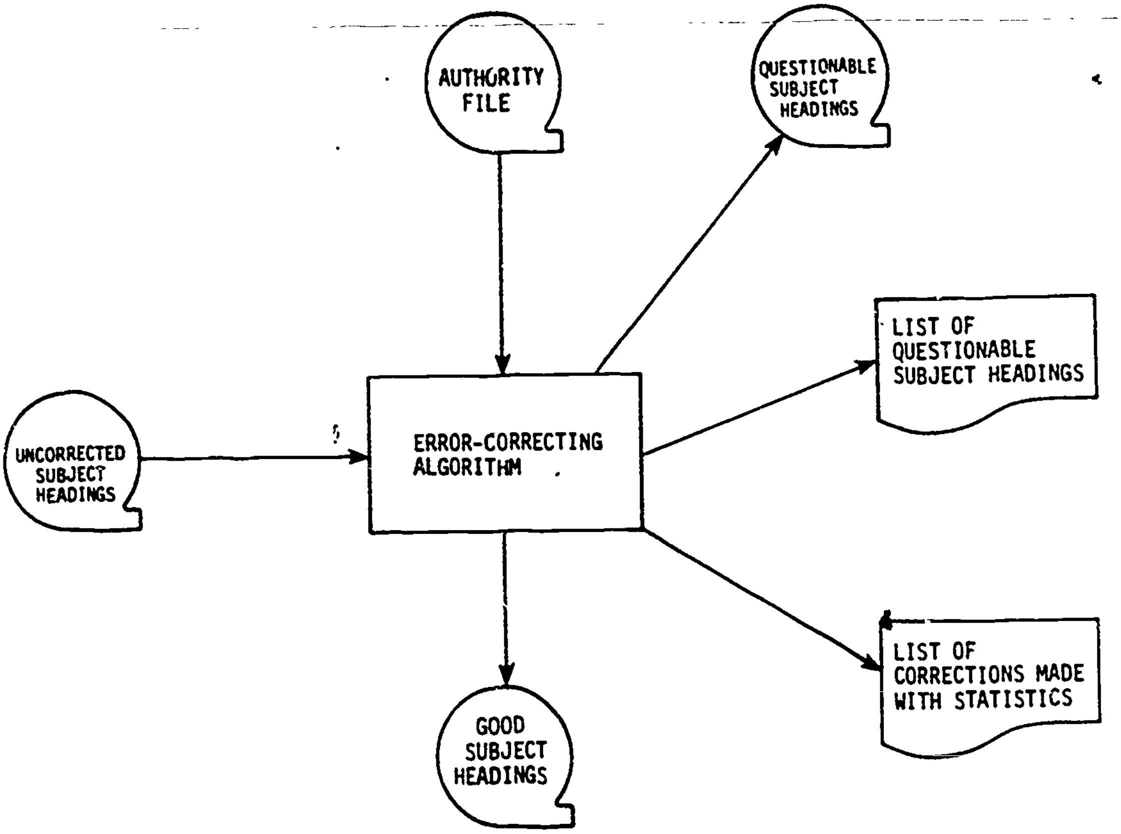

AUTHORITY FILE
QUESTIONABLE SUBJECT HEADINGS
LIST OF QUESTIONABLE SUBJECT HEADINGS
UNCORRECTED SUBJECT HEADINGS
ERROR-CORRECTING ALGORITHM
LIST OF CORRECTIONS MADE WITH STATISTICS
GOOD SUBJECT HEADINGS

Figure 5. Error-correcting Algorithm

1. Check for Errors in the Second Half of the Subject Heading

When checking for errors in the second half of a heading, a certain number of characters counted from left to right, depending upon the length of the key of the test subject heading, are assumed to be free from typographical errors. These characters are referred to as the "truncated key." The truncated key is used as an entry point into the authority file to locate the relevant subject headings. Then, the keys of these potentially relevant subject headings are matched with those of the test subject heading to identify typographical errors.

The truncated key is created from the original key of the test subject heading using the formula:

$$\text{Length of the truncated key} = \frac{\text{Key length} - 1}{2}$$

This formula is not used when the key contains 6 or fewer characters. If the key contains an even number of characters, truncated key length is obtained by rounding down the value obtained from the preceding formula. That is, if the length of the key is 10 characters, the formula gives the length of the truncated key as (10 - 1) / 2 = 4.5 which is then rounded down to 4. This means that the truncated key of the test subject heading contains 4 characters that are assumed to have no typographical errors.

To illustrate this truncation further, consider the 11-character key "1INVENITONS" of a test subject heading. The truncated key consists of 5 characters, "1INVE." These 5 characters, assumed free from typographical errors, are used to identify the potentially relevant subject headings in the authority file. By comparing the relevant headings identified by the truncated key with the test subject heading, the error-correcting algorithm checks for any spelling error in the last 6 characters, "NITONS," of the test subject heading.

Not all subject heading keys whose initial characters agree with the truncated key of the test subject heading are checked for errors. The only subject heading keys checked for errors are: (1) those whose initial characters are the same as those of the truncated key; and (2) those whose total lengths are within one character of the lengths of the keys of the test subject headings. The length of the subject heading keys to be checked is restricted because the error-correcting algorithm attempts to correct only the following types of errors:

(1) one excess character,
(2) one dropped character,
(3) a character incorrectly substituted,
(4) a transposition error.

In the case of the error of an excess or dropped character, keys to be compared have either one character more or one less than those of the keys of the test subject headings. In the case of the substitution or transposition error types, lengths of the keys of test subject headings are equal to those of the correct subject headings. If the lengths are equal, the keys are compared to determine the number of characters which do not match. If only one character does not match, then it is assumed to be a character incorrectly substituted. If two adjacent characters do not match, the test key is a candidate for a transposition error. [20] When more than two characters do not match, no automatic attempt is made to correct the heading.

As an illustration, let us again consider the test key, "1INVENITONS." Its truncated key is "1INVE." This truncated key identifies the following keys in the authority file as potentially relevant:

KEY	LENGTH
1INVENTIONS*	11
1INVENTORIES*	12
1INVENTORS*	10
1INVERSE	8
1INVERTEBRATES	14
1INVESTIGATION	14
1INVESTIGATIONS	15
1INVESTMENT*	11
1INVESTMENTS*	12

Since the length of the key of the test subject heading is 11 characters, only those keys whose lengths are 10, 11, or 12 characters are targeted for comparison with the test key. The target keys in the preceding list are identified by asterisks (*). The error-correcting algorithm compares "1INVENTIONS" and "1INVESTMENT" (each 11 characters) with the test key "1INVENITONS" for possible replacement and transposition errors. Similarly, the algorithm compares "1INVENTORIES" and "1INVESTMENTS" (each 12 characters) with "1INVENITONS" for a dropped character in the test key. The algorithm then compares "1INVENTORS" (10 characters) with "1INVENITONS" for an added character in the test key.

When the error-correcting algorithm discovers that a test subject heading differs from that of an authority file heading only in numeric characters, the algorithm will not alter the test heading. The reason is that the algorithm takes advantage of the natural redundancy in the subject headings. However, this redundancy does not exist in the case of numeric characters. For instance, changing the following subject headings from one to the other would result in incorrect headings:

IBM 360 (computer)	IBM 370 (computer)
Piano music (3 hands)	Piano music (4 hands)
United States-Economic Policy-1961	United States-Economic Policy-1971

In any case, given the widespread occurrence of such headings which differ only slightly in numeric characters and their unpredictability, the error-correcting algorithm does not change the numeric characters.

There may be some headings for which the primary key logically should have more than 28 characters. The number of nonblank characters in the key before truncation is included as part of the authority record. Whenever this value exceeds 28, the primary key will be incomplete. Since it will always have the initial characters and the character count, the incomplete key does not pose any problems in identifying the target keys. However, the incomplete key cannot be compared to the test key. For this purpose, the full key must be rederived from the subject heading or subdivision contained in the authority record.

Thus, the error-correcting algorithm compares all the subject headings in the test file with those in the authority file and splits the test file into two separate files: (1) valid headings, and (2) questionable headings. Valid headings are those for which there is a corresponding heading in the authority file and hence which are assumed to be free from errors; or headings for which, after correcting a typographical error, the algorithm found a match in the authority file. The questionable headings file consists of those headings for which no match could be found in the authority file even after correcting potential typographical errors. The headings in this file are then checked using the reverse key.

2. Check for Errors in the First Half of the Subject Heading

The algorithm described in the previous section would not work if the error occurs in the initial characters of the subject headings. If the key of the test subject heading is "1INEVNTIONS" instead of "1INVENITONS," the program would have difficulty in identifying the potentially relevant keys in the authority file. For this purpose, the reverse keys for all headings in the authority file are utilized. For example, for the keys "1INVENTIONS" and "1INVESTIGATIONS," corresponding reverse keys are "SNOITNEVNI1" and "SNOITAGITSEVNI1." These reverse keys form the basis for correcting errors which occur in the first half of the test subject heading.

If the key of a test subject heading is "1INEVNTIONS," the algorithm reverses the key to "SNOITNVENI1." The remaining correction procedure for deriving a truncated key, identifying the potentially relevant keys in the authority file, identifying the target keys, and performing the final correction process is the same as that described in the previous section. The truncated reverse key is "SNOIT" which identifies the following entries as potentially relevant:

PRIMARY KEY	REVERSE KEY	LENGTH
1DISLOCATIONS	SNOITACOLSID1	13
1INVESTIGATIONS	SNOITAGITSEVNI	15
1DIFFUSIONOFINNOVATIONS	SNOITAVONNIFON	23
1MEDICALINNOVATIONS	SNOITAVONNILAC	19
1EDUCATIONALINNOVATIONS	SNOITAVONNILAN	23
1AGRICULTURALINNOVATIONS	SNOITAVONNILAR	24
1INJECTIONS*	SNOITCEJNI1	11
1FUNCTIONS*	SNOITCNUF1	10
1HYPERFUNCTIONS	SNOITCNUFREPYH	15
1INJUNCTIONS*	SNOITCNUJNI1	12
1INVENTIONS*	SNOITNEVNI1	11
XINVENTIONS*	SNOITNEVNIX	11

Since the length of the test subject heading key is 11 characters, only those keys whose lengths are 10, 11, or 12 characters need to be compared with the test key. Once the target keys (marked with asterisks in the preceding list) are identified, the procedure proceeds exactly the same way as when the target keys were identified using the primary keys. After changing the "EV" in the text key to "VE," a match would be found and the correct form of the heading would be assumed to be "Inventions."

D. Testing the Algorithm

To test the error-correcting algorithm, a COBOL program was developed for the Sigma 9 computer at OCLC. The test was limited to form subdivisions since a relatively complete list of form subdivisions had been compiled as a part of the study on subject heading patterns. [21] The list was available in machine-readable form and was used as an authority file for form subdivisions. Form subdivisions extracted from bibliographic records were then checked using the error-correcting algorithm. The final version of the algorithm described above successfully corrected all omission, addition, substitution, and transposition errors. Subdivisions containing abbreviations were also successfully changed when a record for the abbreviated subdivision was included in the authority file. Form subdivisions containing more serious errors or multiple errors were identified but were not changed. There were no cases where a valid subdivision was modified.

III. LIMITATIONS OF THE METHOD

The success of the error-correcting algorithm depends on the comprehensiveness of the authority file. If the authority file is incomplete, the routine may change correct subject headings to other headings. Examples of such troublesome headings are:

Adaption	Adoption
Painting	Paintings

The greater the number of subject headings in the authority file, the greater the probability of avoiding erroneous corrections. For instance, if both "adaption" and "adoption" are included in the authority file, an erroneous correction would not occur.

There are a number of instances where the subject headings contain multiple errors as well as those involving more than one character or a pair of characters. Examples of such errors are:

Distribution (Probablilty theory)
Education accountanility

The algorithm described herein does not attempt to correct such errors.

Despite these limitations, the proposed algorithm would eliminate numerous typographical errors that occur in the subject headings of the OCLC catalog records. Furthermore, this error-correcting algorithm will work with any field (e.g., author) for which an authority file can be created.

REFERENCES

1. Bourne, Charles P. Frequency and impact of spelling errors in bibliographic data bases. Information Processing & Management. 13:1-12; 1977.

2. O'Neill, Edward T.; Aluri, Rao. Subject heading patterns in OCLC monographic records. OCLC/RDD/RR-79/1. Columbus, Ohio: OCLC, Inc.; 1979 ERIC ED 183 167.

3. Tagliacozzo, Renata; Kochen, Manfred; Rosenberg, Lawrence. Orthographic error patterns of author names in catalog searches. Journal of Library Automation. 3:93-101; 1970 June.

4. Damerau, F. A technique for computer detection and correction of spelling errors. Communications of the ACM. 7:171-176; 1964 March.

5. Alberga, Cyril N. String similarity and misspellings. Communications of the ACM. 10:302-313; 1967 May.

6. Blair, Charles R. A program for correcting spelling errors. Information and Control. 3:60-67; 1960 March.

7. Cornew, R.W. A statistical method for spelling correction. Information and Control. 12:79-93; 1968.

8. Damerau, F., op.cit.

9. Davidson, L., Retrieval of misspelled names in an airline passenger reservation system. Communications of the ACM. 5:169-171; 1962 March.

10. Glantz, H.T. On the recognition of information with a digital computer. Journal of the ACM. 4:178-188; 1957 April.

11. Morgan, H.L. Spelling correction in systems programs. Communications of the ACM. 13:90-94; 1970.

12. Muth, F.E., Jr.; Tharp, A.L. Correcting human error in alphanumeric terminal input. Information Processing and Management. 13:329-337; 1977.

13. Riseman, E.M.; Ehrich, R.W. Contextual word recognition using binary diagrams. IEEE Transactions on Computers. C-20:397-403; 1971.

14. Zamora, Antonio. Automatic detection and correction of spelling errors in a large data base. Journal of the American Society for Information Science. 31:51-57; 1980 January.

15. Ibid., p. 52.

16. Ibid, p. 53.

17. Morgan, op.cit., p.91.

18. O'Neill and Aluri, op.cit., p.4.

19. Zamora, op.cit., p.52

20 Morgan, op.cit., p.91

21. O'Neill and Aluri, op.cit., p.70