Analytical Study of Typographical Errors in OPACS and Corrective Measures

Surendran Cherukodan **G Santhosh Kumar** **Sheeja N K** **Humayoon Kabir S**

Abstract

The present study is an attempt to highlight the problem of typographical errors in OPACS. The errors made while typing catalogue entries as well as importing bibliographical records from other libraries exist unnoticed by librarians resulting the non-retrieval of available records and affecting the quality of OPACs. This paper follows previous research on the topic mainly by Jeffrey Beall and Terry Ballard. The word "management" was chosen from the list of likely to be misspelled words identified by previous research. It was found that the word is wrongly entered in several forms in local, national and international OPACs justifying the observations of Ballard that typos occur in almost everywhere. Though there are lots of corrective measures proposed and are in use, the study asserts the fact that human effort is needed to get rid of the problem. The paper is also an invitation to the library professionals and system designers to construct a strategy to solve the issue.

Keywords: OPAC, Information Retrieval, Authority Files, Typographical Errors

1. Introduction

Catalogues mirror a library's collection and Web based Online Public Access Catalogues (OPACs) mirror a collection worldwide. OPACs appeared in the 1980s followed by Web-based OPACs in the late 1990s. Web OPACs are a natural progression in technological development and could be termed to be an advanced second generation OPAC (Hildreth, 1991). They are advancement on traditional OPACs serving as the gateway to the resources not only held by the particular library but also to the holdings of other linked libraries and further to regional, national and international resources (Ramesh Babu and O'Brien, 2000). There are several factors that affect the quality of a library OPAC and typographical error is one of the important factors. The errors shall block access to the available documents. While research has been done on how OPACs are used, little research has been done on how spelling errors affect the information retrieval (Willson & Given, 2008). The present study is focusing on the issues of cataloguing errors and how these affect the retrieval of documents and the quality of OPACs. The paper also attempts to suggest some solutions to avoid errors in online catalogues.

2. Typographical Errors

Typographical error in online catalogues is a problem shared by all libraries (Eyler 2004). Typographical errors (typos) in bibliographic and authority records are a barrier to effective information retrieval in libraries. Typos impede information access by causing inaccurate or incomplete search results in online catalogue searches and in searches in other library databases (Beall 2004). Typographical errors are the mistaken addition, deletion, transposition or substitution of letters or other characters in a word, or the addition or deletion of spaces within or between words (Gardner, 1992). Spelling is an important literacy skill, crucial to

successful searching OPACs (Willson and Given, 2008). The spelling mistakes occurred during data entry work is left unnoticed by library professionals and it affects the standard of online systems. A simple typo can mislead the user to think that the library does not own the material. The invisibility of available records on OPACs leads to the non-use of materials and it violates the second law of library science.

3. Review of Related Literature

There are several studies on typographical errors in OPACs. Beall (1991) published an article in American Libraries in which he gave examples of some odd words that were appeared wrongly on online catalogues. Ballad (2008) conducted a keyword inspection of the Adelphi University online catalogue in 1991 based on the work of Beall. He worked with word like "Commerical" and found more than 800 references even in OCLC catalogue. He looked through the entire keyword database, found the words that were typos, fixed them, and maintained a list of these words. He confirmed that most libraries have a typo problem because libraries receive their cataloging records from similar sources, such as OCLC or retrospective conversion vendors. Eyler (2004) followed Ballard's original list in 1998 and made over 1300 corrections to Scotty bibliographic records by calling up the typos in the keyword ("Word") option in Scotty. Over the years, he looked for more misspellings and corrected around 4,500 records.

Beall (2004) evaluated several studies reported on the impact of typographical errors in bibliographic databases, and on methods of reducing them and identified that published studies had not covered the topic of typographical errors in authority records. The Library of Congress Authority File (LCAF) is the set of name, title and subject authority records created by the Library of Congress and by the co-operating libraries. He noted that the launch of the OCLC Connexion added significant new functionality to searching the LCAF. With Connexion, librarians can use OCLC to search the Library of Congress Authority File by keyword. This keyword search capability makes it possible to systematically find typographical errors in authority records.

The study by Willson and Given (2008) was focusing on the spelling mistakes done by users while searching OPACs. They observed that spelling was a topic that has received little attention in library and information studies. It affects both information retrieval and search behaviour. This study is important for librarians because the effect of misspelling either by libraries or by searchers can seriously affect the information retrieval.

4. Scope and Limitations of the Study

The current study is on a problem that affects all types of libraries. Though the list of words generated by Ballard for examining errors in OPACs is a long one, the authors selected only one word to inspect and report the possibility of errors in OPACs. The study does not attempt to depict the rate of errors statistically. The study primarily examines whether the problem of typos still exists in library OPACs. The authors examined various local, national and international OPACs belonging to public, national, academic and

special libraries. The study is limited to the typos associated with the term 'management'. However, it can be generalized that library OPACs suffer from various errors associated with other terms.

5. Methodology

The study was conducted by searching several OPACs using a single word. Terry Ballard listed 8160 misspellings that are likely to be found in our OPACs. Based on the number of hits a particular misspelling got in OhioLINK, he has categorized them into five sections. Section A contains 117 terms with 100 hits having highest probability. Section B contains 1313 terms inviting 16-99 hits having high probability. Section C involves 1143 terms with moderate probability showing 8-15 hits. Section D has 3242 terms and 2-7 hits with low probability. The Section E contains 2341 terms with only one hit categorized as lowest probability. The present study selected the word 'management' from Ballard's list to check its different variations of use in OPACs. The authors searched the word management in a mistaken form by addition, deletion, transposition and substitution of letters. Search was conducted on a number of local, national and international OPACs and it was found that the word is wrongly entered in online catalogues of libraries irrespective of type, size and region. The OPAC search was followed by literature survey on the web and on online databases. For the purpose of showing examples, five national libraries and five university libraries were selected. The numbers of instances of errors corresponding to each library were excluded since it was not a particular library based study.

6. Analysis

The following five national libraries were selected to show examples of error instances in OPACs in the term 'Management'.

1. The British Library (BL)
2. The Library of Congress (LOC)
3. The National Library of India (NLI)
4. The National Library of Australia (NLA)
5. The National Library of South Africa (NLS)

The OPACs of these national libraries were examined from 24-12-2013 to 10-2-2013. Basic search was performed to retrieve documents having the wrong term for 'Management'. The possibility of missing and replacing letters in the term in ten different ways was examined. The retrieved documents were examined to ensure the error. There were considerable numbers of references to the wrong terms in all OPACs. The errors were present in the title, author, subject, series, notes, contents etc. The Table I shows the presence of errors in five national library OPACs.

Table 1: Presence of Error in Five National Library OPACs

Sr. No	Wrong Word	Presence of error				
		BL	LOC	NLI	NLA	NLS
1	Mamagement	Yes	Yes	Yes	Yes	Yes
2	Managemant	Yes	Yes	Yes	Yes	Yes
3	Managemeent	Yes	Yes	-	-	-
4	Managemet	Yes	Yes	Yes	Yes	Yes
5	Managemnt	Yes	-	Yes	-	Yes
6	Managenent	Yes	Yes	Yes	Yes	-
7	Managment	Yes	Yes	Yes	Yes	Yes
8	Mangement	Yes	Yes	Yes	Yes	Yes
9	Mnagement	Yes	Yes	Yes	Yes	Yes
10	Nanagement	Yes	Yes	-	Yes	-

The authors further checked some University Library OPACs to know the presence of errors in them. For this purpose, Top five Universities from the Times Education World Ranking of Universities 2012-2013 were selected to show examples. The California Institute of Technology, the top ranking institute as per the Times ranking was excluded as there were no typos for the variations of the term 'Management' in its OPAC. The next 5 universities are listed below:

1. Stanford University (SU)
2. University of Oxford (UO)
3. Harvard University (HU)
4. Massachusetts Institute of Technology (MIT)
5. Princeton University (PU)

Table II shows the presence of errors in five top universities in Times Education World Ranking. The presence of errors was similar to that of five national libraries shown above.

Table 2: The Presence of Errors in Five Top Universities

Wrong Word	Presence of error				
	SU	UO	HU	MIT	PU
Mamagement	Yes	Yes	Yes	Yes	Yes
Managemant	Yes	Yes	Yes	-	Yes
Managemeent	-	-	-	-	-
Managemnt	Yes	Yes	Yes	-	-
Managenent	-	Yes	Yes	-	-
Managment	Yes	Yes	Yes	Yes	Yes
Mangement	Yes	Yes	Yes	Yes	Yes
Mnagement	Yes	Yes	Yes	Yes	-
Nanagement	Yes	Yes	-	-	Yes

Among the ten wrong variations of 'Management', the majority of errors belong to **Mangemen**t and **Managment**. These two terms can be identified as having the highest probability to err.

7. Typos in Online Databases

When the instances of errors examined, it was found that in some libraries most of the errors were associated with journal articles. This led to an examination of online databases. Four online databases (ScienceDirect, ProQuest ABI/INFORM Complete, Taylor & Francis and Emerald) were examined with the typos belonging to the word management. Surprisingly, it was found that online databases are not free from typos. It establishes that the typos in the online sources invariably enter library catalogue records.

8. Corrective Measures

The maintenance of OPACs without typos is essential for several reasons. The authors contacted the California Institute of Technology (Caltech) Library to know the strategies adopted for an error free OPAC. Teague Allen, Cataloguing and Metadata Librarian replied that the library avoids typos and other errors by obtaining the largest part of their authority records on personal, corporate, and jurisdictional names from their national library's cooperative authority program, which is the Library of Congress' Name Authority Cooperative Program, or NACO. The cataloging librarians of Caltech are trained to the same standard, so the headings we produce locally are of equal quality. There is a separate program for subject headings, SACO (http://www.loc.gov/aba/pcc/saco/index.html) , from which they obtain all their subject headings. Beyond adherence to national standards, the other part of the strategy is careful attention to maintenance. He does not think that there's anything startling or unexpected to their strategy, but it does demand time and regular attention. Keeping in mind the response of an eminent librarian, we suggest the following methods;

8.1 Batch Processing

By definition, authority file has to support maintenance function to support manual and automatic error detection and correction (Burger, 1985). To achieve quality catalogue records the Library of Congress (LC) follows a correcting mechanism called "BatchCat". Upon the receipt of request for a change in authority records the LC handles it as quickly as possible and the changes will be propagated subsequently to the other repositories and catalogues. The Program for Cooperative Cataloguing (PCC) has also resulted in recommendations from several task groups to improve the authority files and database of bibliographic files. This requires at least two things, a request from a user and initiation of batch processing from the part of LC personnel.

8.2 Collaborative Methods

It is generally believed that automated routines can correct data without the intervention of an expert. But, the intellectual capacity of human beings can be applied to avoid many typos. The seminal works of Terry Ballard can be followed to trace typos. The use of e-mail lists or listservs, announcement lists, discussion list and notification lists were suggested by various authors. Ballard maintains a blog (**Typo of the Day for**

Librarians, 2103) which is a collection of alphabetical list of possible typos. Kent State University maintains notification lists for errors DEWEYERROR (Dewey decimal classification numbers error) and LCCERROR (LC Classification Error). The university also maintains PERSNAME-L for notifying issues related to personal names in bibliographic and authority records and SACOLIST for LC Subject Headings (LCSH). Fairclough (2013) gives a detailed description of the use of these notification schemes. Many library websites provide the facility for reporting errors. All these efforts mentioned are collaborative in nature and monitoring and maintenance of each of them become the shared responsibility of many people.

8.3 Google's Effort

The Oxford-Google digitization project has made available large volume of material in PDF to the public. From books.google.com one can search the full text of books and magazines that Google has scanned, converted to text using optical character recognition, and stored in its digital database. James and Weiss (2012) report the error rates in the metadata records of Google's digitization project. It is claimed that the errors found in this project is more than that found in typical online catalogues. The Google project is a fully automated system and encourages full text search rather than metadata. Though Google has invited lots of criticism for the quality of its metadata, it is expected that these errors could be reduced when the project matures. The project shall impart many ideas for library professionals on typos.

8.4 Web 2.0 Based Methods

New generation OPAC, named as OPAC 2.0 extensively uses Web 2.0 features to improve search and retrieval services. OPAC 2.0 addresses severity of the catalogue problem by providing intelligent external interfaces which can be used to search items using associations, context and spelling alternatives which are generated from the library's own catalogue. For example, AquaBrowser from Serials Solutions (2013) presents each search term as a "Word Cloud" of related terms (subjects, thesaurus terms and spelling variations). Terms more closely related appear close to the search term. Faceted navigation further exposes the library's data including catalogue and article content to the patrons.

8.5 Open Library

Open Library project is an open, editable catalogue of books. Instead of using explicit data records, every piece of data is represented by an address known as URI (Universal Resource Identifier). The architecture aligns with W3C web standards for the semantic web, and allows much more flexible searching and data mining than would be possible with a MARC record. The patron is welcome to add and modify any data pertaining to the books like Wikipedia documents evolving into a clean collection of bibliographic records.

9. Discussion

The study found that typographical errors pose great challenges in the online environment irrespective of strategies and steps by the library professionals. The present study has attempted to address the matter to invite the attention of professionals who may collectively find measures to solve it. The authors just

searched an Indian OPAC with the wrong words "Engin**e**ring" and "Test Book" and found many references to both. It suggests that the matter requires further examination by experts. The occurrence of typos is widespread and global.

Those libraries who initiate the work of automation may include the subject of typos in its outline. Since errors are there in the servers from where libraries import bibliographic details, the imported record can be verified along with doing change for localization. Above all, human examination for every online creation is required. Most of the errors in our catalogues are due to lack of scrutiny as we did for card catalogues.

10. Conclusion

This current study is not a comprehensive one on the subject. The purpose was to make an attempt to invite the attention of library professionals towards the issue. It needs further discussion and examination by scholars using different strategies. The list of probable misspellings by Ballad (2009) and Beall (2004) can be used for the purpose of locating errors in OPACs. It will help library professionals to shift from the process of accidentally identifying errors in OPACs to systematically discovering and solving the problem. The authors do not propose a single method to fix it. Though there are many corrective measures suggested by many, the issue is still there and librarians should look into the matter seriously. Automated mechanisms and Web 2.0 based solutions are emerging out. But we conclude that without human attention, the problem will stay with us.

References

1. BALLAD, Terry (2008). Systematic identification of typographical errors in library catalogues. Cataloguing and Classification Quarterly, 46(1), 27-33.

2. BALLARD, Terry (2009). Typographical Errors in library databases. available at: http://www.terryballard.org/typos/typoscomplete.html. (accessed on 15/1/13)

3. BURGER, Robert H. (1985). Authority Work: the Creation, Use, Maintenance and Evaluation of Authority Records and Files. Littleton, Colo.: Libraries Unlimited.

4. BEALL, Jeffrey. (1991). Ideas: the dirty database test. American Libraries, 22(3), 197.

5. Beall, Jeffrey. (2004). Using OCLC Connexion to find typographical errors in authority records. OCLC Systems and Services: International Digital Library Perspectives, 20(2), 71-75.

6. EYLER, Wendee. (2004). Paradise lost is found: Typological errors in online catalogs. Available at: http://associates.ucr.edu/feyl304.htm (Accessed on 9/12/2012)

7. FAIRCLOUGH, Ian, (2013). Collaborative Initiatives in Error Handling and Bibliographic Maintenance: Use of Electronic Distribution Lists and Related Resources, Cataloging & Classification Quarterly, 51(1-3), 265-290

8. GARDNER, S A. (1992). Spelling errors in online databases: what the technical communicator should know. Technical Communication, 39(1), 50-3.

9. HILDRETH, C.R. (1991). Advancing toward the E3 OPAC: the imperative path, In Van Pulis, N. (Ed.), Think Tank on the Present and Future of the Online Catalog: Proceedings, (pp. 39-48). January 11-12, Chicago: American Library Association, 1991.

10. JAMES, R. and WEISS, A. (2012). An Assessment of Google Books' Metadata, Journal of Library Metadata, 12(1),15-22

11. RAMESH BABU, B. and O'BRIEN, A. (2000). Web OPAC interfaces: an overview. The Electronic Library, 18(5), 316 – 330.

12. Serials solutions (2013). available at: http://www.serialssolutions.com (Accessed on 9/2/2013)

13. Typo of the Day for Librarians (2013) available at: http://libtypos.pbworks.com (Accessed on 9/2/2013)

14. WILLSON, Rebekah and GIVEN, Lisa M. (2008).The effect of misspellings on information retrieval in online public access catalogues. Proceedings of the 36th annual conference of the Canadian Association for Information Science (CAIS), University of British Columbia, 5-7 June 2008 in Vancouver.

About Authors

Mr. Surendran Cherukodan, Junior Librarian, Cochin University of Science and Technology, CUSAT
E-mail: scherukodan@gmail.com

Dr. G Santhosh Kumar, Assistant Professor, Department of Computer Science, CUSAT
E-mail: san@cusat.ac.in

Dr. Sheeja N K, Assistant Librarian, Cochin University of Science and Technology, COAST.
E-mail: nkscusat@gmail.com

Dr. Humayoon Kabir S, Associate Professor, Department of Library and Information Science, University of Kerala
E-mail: humayoonkabirs@yahoo.co.in