BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en Vol. 30 No. 1. 2025

UDC 025.4:004.65
DOI: 10.62660/bcstu/1.2025.59

Factors complicating the identification and processing of duplicates in bibliographic records: A theoretical perspective

Oleh Vasylenko*
Postgraduate Student
V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine
03187, 40 Academician Glushkov Ave., Kyiv, Ukraine
Lecturer
Ukrainian Institute of Arts and Sciences
08292, 14 Instytutska Str., Bucha, Ukraine
https://orcid.org/0000-0001-8498-2950

Abstract. This article examined the factors that create challenges in the process of identifying and processing duplicates in bibliographic records, which are a crucial component of the information systems of libraries, archives, and publishers. The study explored issues arising from typographical errors, variations in transliteration, the use of special characters, homoglyphs, differing word abbreviation rules, inconsistencies in author name spellings, and shortcomings in the application of standard identifiers such as ISBN and ISSN. Particular attention was given to the impact of discrepancies between international and local MARC standards – including Unimarc and Marc21 – on the creation and processing of bibliographic data. The analysis demonstrated that improper handling of bibliographic records can lead to degraded information retrieval quality for users, inaccuracies in source citations, and increased time expenditures for cataloguing and indexing. Furthermore, inconsistencies in standards impair the management of bibliographic data in multinational systems. The article also examined the consequences of these issues for bibliographic systems, including reduced search query accuracy, difficulties in data integration across catalogues, and increased time and resource costs for record processing. A set of solutions was proposed, including the adoption of unified record standards, the implementation of advanced adaptive search algorithms that account for linguistic and technical discrepancies, and enhanced authority control in bibliographic record creation. The findings have practical implications for information system developers, cataloguers, and library professionals, as they contribute to improving bibliographic databases, reducing duplicate records, and enhancing information retrieval quality for end users

Keywords: data processing; transliteration; library systems; information technology; authority control; cataloguing

INTRODUCTION

The identification and processing of duplicates in bibliographic records constitute a critical yet complex task in modern library science, information management, and digital technologies. The growing volume of information accompanying the development of electronic resources and databases presents researchers and practitioners with new challenges in ensuring the accuracy, reliability, and accessibility of bibliographic data. The problem is exacerbated by technical, organisational, and conceptual factors that lead to a significant number of duplicates and their improper handling. This, in turn, affects information retrieval quality,

Article's History: Received: 9.10.2024; Revised: 29.02.2025; Accepted: 17.03.2025

Suggested Citation:
Vasylenko, O. (2025). Factors complicating the identification and processing of duplicates in bibliographic records: A theoretical perspective. *Bulletin of Cherkasy State Technological University*, 30(1), 59-67. doi: 10.62660/bcstu/1.2025.59.

*Corresponding author

the efficiency of working with bibliographic records, and the overall performance of information systems. Contemporary research in this field highlights the necessity of a systematic approach to studying this issue, incorporating both theoretical foundations and practical solutions. Analysing the key factors influencing the identification and processing of duplicates is essential for developing effective algorithms, methodologies, and standards. This issue remains relevant not only for library systems but also for other domains dealing with large volumes of information, including publishing, education, and scientific research.

Recent studies on the identification and processing of duplicate bibliographic records focus on resolving complex issues related to metadata accuracy, data processing algorithms, and bibliographic resource integration. This review examines key publications that provide theoretical frameworks and propose practical solutions to address this problem. One seminal work is that of W. Walters & E. Wilder (2023), which investigates issues related to errors in bibliographic citations and the need to improve metadata accuracy in digital bibliographic records. Their approach emphasises the importance of implementing new data processing algorithms. Furthermore, the study proposes novel methods for detecting and eliminating duplicates in large databases, thereby enhancing the precision of bibliographic searches.

A significant portion of research concentrates on specific aspects of duplicate processing. For instance, S. McKeown & Z. Mir (2021) explore the effectiveness of various deduplication methods in systematic reviews, while B. Hammer *et al.* (2023) provide evidence for the necessity of duplicate removal to ensure the quality of scholarly works. Another key study by N. Borissov *et al.* (2022) describes automated algorithms that can facilitate this process, particularly in medical research. Data inconsistencies and duplication are primary factors negatively affecting the quality of bibliographic information. Similarly, H. Levenson *et al.* (2024) analyse how imperfections in bibliographic records can lead to losses in monograph programmes. From an algorithmic perspective, the work of B. Berger *et al.* (2021) advocates the use of Levenshtein distance for identifying similar entries in biological databases – a method that could be adapted for bibliographic applications. The publication by M. Guo *et al.* (2023) outlines strategies for data cleansing in real-world scenarios, emphasising the use of automated processes and standardised workflows. Other authors, such as C. Lunny *et al.* (2021), offer practical recommendations for managing duplicate research findings in systematic reviews. E. Benevento *et al.* (2022) investigate interactive data discovery processes in business environments, which could be applied to library systems. Meanwhile, P. Raja & K. Thangavel (2020) propose machine learning techniques for imputing missing data, which may prove useful in duplicate identification.

The objective of this study was to analyse the factors complicating the identification and processing of duplicates in bibliographic records, identify key challenges, and outline potential solutions within the context of modern digital technologies. The research employed analytical and comparative methods to uncover the peculiarities of duplicate occurrence in bibliographic systems. The informational basis comprised contemporary publications in library science and scholarly works addressing cataloguing and bibliographic record processing in digital environments. Special attention was given to analysing international standards such as MARC21 and their adaptation to Ukrainian library contexts.

Identification and Processing of Duplicates in Bibliographic Records

The primary challenge researchers face in identifying duplicates is the large volume of data and the complexity of processing varying record formats. As noted by P. Lizunov *et al.* (2021), ensuring the accuracy of bibliographic records requires the application of combined methods capable of identifying near-identical entries, particularly those with minor discrepancies in metadata or formatting. Their research highlights the importance of employing modern algorithms for detecting and processing duplicates, significantly enhancing search efficiency in electronic academic databases. The issue of data cleansing is also an integral part of duplicate processing. M. Gesicho *et al.* (2020) examine the process of cleaning medical data and propose strategies to improve data processing quality. Although their expertise pertains to medical systems, similar approaches can be applied to bibliographic records. Data purity is a necessary condition for ensuring the reliability of search results and the analysis of scientific information.

The analysis of bibliographic record editing issues in electronic catalogues of Ukrainian libraries is a crucial aspect of contemporary library science. O. Bruy (2015) and T. Bykova (2016) examine the challenges of editing bibliographic records in electronic catalogues, using Ukrainian libraries as a case study, and emphasise the difficulties in adapting international standards such as MARC21. Innovative methods based on artificial intelligence (AI) can also contribute to resolving the duplicate problem. R. Chauhan *et al.* (2023) describe the use of AI for detecting duplicates in software repositories, significantly reducing errors in identifying duplicate reports. However, integrating AI into infrastructure monitoring systems demonstrates considerable potential while simultaneously raising challenges regarding accuracy, ethics, and data interpretation (Alves *et al.,* 2024). Consequently, establishing ethical guidelines for AI usage becomes essential, particularly through the development of "ecological" impact declarations for sustainable development (Bogani *et al.,* 2023). These technologies can be adapted for

bibliographic records by employing specialised algorithms for duplicate detection, thereby improving the accuracy of large-scale database processing.

Furthermore, in software engineering, the use of word vector representations facilitates more precise identification of duplicate bug reports, enhancing performance in IT systems (Budhiraja *et al.,* 2018). The issue of visual similarity between Unicode characters is particularly critical for cataloguing and security systems, as homoglyphs can be exploited for deception (Coffman, 2021). Online services such as AMP-What enable the detection of duplicate characters, which is especially relevant for digital libraries (Discover your character: C, n.d.). Classical cataloguing principles remain foundational for constructing logical information structures and remain pertinent in the context of digital transformation in library science (Gorman, 2004). Another significant aspect is the application of modern methods to enhance metadata management, such as frameworks for improving information systems. C. Bockel-Rickermann *et al.* (2023) note that the use of such approaches in financial systems could be beneficial for developing analogous solutions for bibliographic records. This includes devising methods for automated metadata verification and correction, minimising human error and improving the efficiency of bibliographic systems. In software bug duplicate detection, effective feature extraction plays a key role in enhancing the accuracy of routing systems (Neysiani *et al.,* 2020).

Thanks to technological advancements, the automation of duplicate identification has become more accessible. E. Negro-Calduch *et al.* (2021) highlight the necessity of employing cutting-edge technologies to optimise electronic healthcare systems, which could also benefit library information systems. The use of methods such as data cleansing and AI algorithms significantly reduces error rates and enhances record processing quality. Duplicate identification in bibliographic systems requires a comprehensive approach that incorporates technological innovations, new data cleansing and validation methods, and the application of modern standards for integrating diverse information types. This ensures data accuracy and reliability, which are essential for effective search and management of bibliographic records. One of the key tasks is developing models that reduce errors associated with duplicates, necessitating advanced algorithms capable of processing large datasets. S. Sehra *et al.* (2020) provide an example of ensemble models for detecting duplicates in software bug reports. This study underscores the importance of combined approaches to improving duplicate detection accuracy in large-scale data systems. These models can be adapted for bibliographic records, as their underlying principles – particularly metadata classification and analysis methods – are universal and can be tailored to the needs of academic databases.

Additionally, T. Thwel & G. Sinha (2021) note that data deduplication approaches are becoming increasingly vital in the face of rapidly growing information volumes. They outline key strategies and concepts used to combat duplicates, including machine learning methods actively applied across various domains, including bibliographic record processing. Their work provides a detailed analysis of deduplication aspects that could inform the development of duplicate identification methods in bibliographic systems. Another critical dimension of the duplicate problem is the cost of processing "dirty" data. As L. Page (2005) notes, data errors can significantly increase processing costs and degrade the quality of information systems. This further underscores the need for implementing efficient algorithms to process bibliographic records and ensure their reliability. Overall, addressing the issue of duplicates in bibliographic records requires the development of new models that account for both technological innovations and the specificities of metadata processing in particular domains. These models should incorporate duplicate detection methods adaptable to modern data processing and search requirements in bibliographic systems.

Factors Complicating the Detection and Processing of Duplicates in Bibliographic Records

One of the key challenges in bibliographic record management is ensuring their uniqueness, which requires effective detection and elimination of duplicates. However, certain factors significantly complicate this process, reducing the accuracy of algorithms and automated data processing tools. Extraneous characters often go unnoticed during manual data entry or when importing records from different systems. For instance, an extra space at the beginning or end of a text string may be invisible to the user but is accounted for in automated string comparisons, causing identical records to be recognised as distinct (Hillmann, 2004).

The problem is exacerbated when multiple spaces are inserted between words or when punctuation marks are misplaced. For example, a book title may be recorded as "Bibliography: Fundamentals" or "Bibliography: Fundamentals", leading to mismatches in duplicate detection. To mitigate this issue, data-cleaning algorithms must be applied to remove extraneous characters prior to comparison (Lobuzina, 2013). When working with bibliographic records, the use of special characters – such as non-breaking spaces, tabulation marks, or non-standard encodings – also introduces complications. For instance, a non-breaking space, sometimes used to enhance text aesthetics, may not be supported by duplicate detection algorithms. As a result, even seemingly identical records may be treated as distinct due to differences in whitespace characters (Tillett, 1989). Furthermore, software such as MarcEdit may generate errors during MARC file validation if records contain tabulation marks or other invalid characters. This not only complicates file processing but also necessitates additional time for corrections.

Figure 1 illustrates an error in MarcEdit resulting from the use of tabulation.

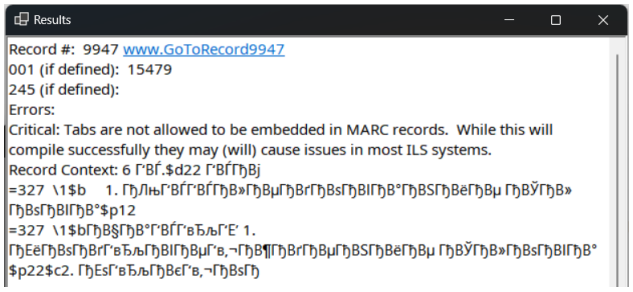

Figure 1. Error in MarcEdit due to tabulation
Source: compiled by the author

Typographical errors remain one of the most prevalent factors affecting the quality of bibliographic records. They may arise due to human error during data entry or imperfections in optical character recognition (OCR) tools. For example, misspelling an author's name, book title, or publication year may prevent duplicate identification or lead to incorrect merging. To minimise the impact of typographical errors, it is essential to implement spell-checking algorithms, machine learning techniques, and tools that compare records against authoritative databases. In multilingual bibliographic records, special characters are often replaced with visually similar alternatives or combinations. For example:

1. The letter Å may be replaced with A;
2. The symbol © may be rendered as (c);
3. The letter ß may be substituted with ss;
4. Fractions such as ½, may be written as 1/2.

These transformations are typically associated with several factors: the use of outdated character encodings, such as ASCII instead of Unicode, which supports a broader range of characters; a lack of knowledge or technical capacity to input special characters by cataloguers; or deliberate optimisation for compatibility with legacy search engines incapable of processing specific symbols correctly. While this approach may facilitate text processing, it complicates the accurate detection of duplicates, as a record containing the character "ß" and another in which it is replaced with "ss" would be treated as different entries. Homoglyphs are characters that appear visually identical but have different encodings. For instance, Cyrillic and Latin letters such as "c", "a", or "o" may look the same but are encoded differently in digital systems (Hammer, 2023). This introduces additional complexity in record comparison, as such characters are considered incompatible by search algorithms (Hopkinson, 2008).

Such errors may occur both during manual text editing and in the course of automated recognition of scanned documents. For example, the substitution of the Cyrillic character "і" with the Latin "l", or confusion involving apostrophes – such as the difference between the characters ' and ‘ or their omission altogether (Walters & Wilder, 2023). To mitigate these issues, it is necessary to implement algorithms that account for potential homoglyphs and to expand validation tools to detect and correct such instances in bibliographic records. These factors necessitate a meticulous approach to data entry standardisation and the adoption of contemporary technologies to improve the quality of bibliographic data processing. Table 1 presents examples of simplified and original formats of characters used in search algorithms.

Table 1. Simplified and original character formats

Simplified	Original character
a	a ą ą ä à á ą
c	c ć ċ
d	d ď
e	e ę ė é è
g	ġ ğ
h	h
i	i í ï
j	j ȷ
k	ĸ
l	l ļ
n	n
o	o o o ȯ ọ ỏ ơ ó ò ö
p	p
q	q
s	ş
u	ʊ u ü ú ù
v	ν v
x	x ҳ
y	y ý
z	ʒ ż

Source: compiled by the author

Homoglyphs of the letter C (Fig. 2), such as characters from other alphabets or graphically similar symbols (e.g., the Cyrillic letter "С" or the mathematical multiplication sign ×), constitute one of the factors complicating the identification and processing of duplicates in bibliographic records. They may lead to errors in search or record creation, as systems often fail to distinguish their differing semantic and graphical meanings. To address this issue, it is crucial to implement character normalisation technologies and data identity verification, which will reduce the number of false duplicates and enhance the accuracy of bibliographic databases.

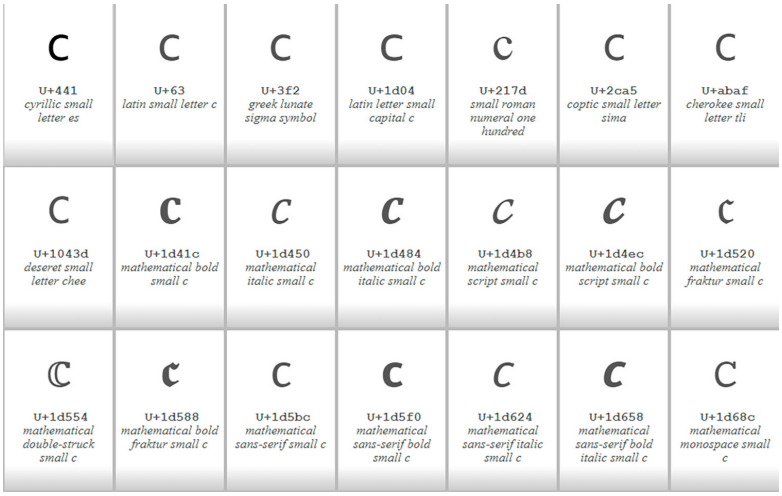

Figure 2. Homoglyphs of the letter C

Source: compiled by the author

The use of varying character cases, such as NASA and Nasa, typically does not affect duplicate detection, as most search algorithms identify such records as identical. However, when merging such records, an additional algorithm for selecting the correct spelling variant becomes necessary. This issue is particularly relevant in bibliographic records requiring precise representation of organisational names or other terms, where even minor case differences may carry significance (Borissov *et al.,* 2022). For instance, in organisational names, case usage may reflect official styling. Abbreviations are common in bibliographic records, yet their inconsistency leads to confusion. The same term may be abbreviated in multiple ways, creating potential difficulties in search and duplicate matching. For example, truncation-based abbreviations: інститут → інст.; or hyphen-based abbreviations: інститут → ін-*т* (McKeown & Mir, 2021). The more abbreviated a word is, the higher the likelihood of ambiguity, which may hinder contextual understanding and the identification of identical records. This is particularly problematic for bibliographic data reliant on the precise rendering of terms. The formulation of headings for authors and organisations varies depending on linguistic, cultural, or even institutional conventions. This may introduce additional challenges (Table 2):

Table 2. Simplified and original character formats

Difficulty	Description	Consequences
Transposed given name and surname	For example, Gresham Machen instead of Machen Gresham. Such variations complicate automated duplicate detection.	Hindered automated duplicate identification, increased errors in bibliographic records.
Name spelling variations	For instance, Smith і Smyth, or Ellen and Olena. Such discrepancies often arise from cross-linguistic conventions, alternative name forms, or typographical errors.	Increased duplicate entries due to name spelling mismatches.
Translation variants	Proper names may be rendered differently, e.g: Longbottom → Лонґботом, Долгопупс or Длиннопопп. Similarly Vyhmeister → Вихмейстер, Ваймейстер.	Search ambiguity, necessitating additional resources for cross-verification.

Source: compiled by the author

The absence of unified rules for recording identifiers is a significant issue. For instance, instead of *Taras Hryhorovych Shevchenko,* a record might only contain initials (*T. Shevchenko*), complicating record identification. Occasionally, only the first author or the responsible person is indicated, which limits search capabilities. The implementation of authority records (normative control) enables the standardisation of names, subject headings, and other terms used in bibliographic data. This not only facilitates searching and

reduces duplication but also enhances the accuracy and consistency of record creation. Standardising data entry, developing automated processing tools, and widespread adoption of normative control are key steps in addressing these challenges and improving the efficiency of working with bibliographic records.

Figure 3 illustrates a subset of author name variations recorded in the International Standard Name Identifier (ISNI) authority file. These variations demonstrate how differing transliteration approaches, linguistic discrepancies, and typographical errors affect the accuracy of bibliographic records. The lack of a unified standard for recording author names significantly complicates identification and duplicate processing. To mitigate this issue, it is advisable to employ authority files such as ISNI and implement algorithms that account for multilingualism and name variation. In gene data clustering methods, the use of fuzzy logic demonstrates increased robustness to data variability (Jahan & Hasan, 2021).

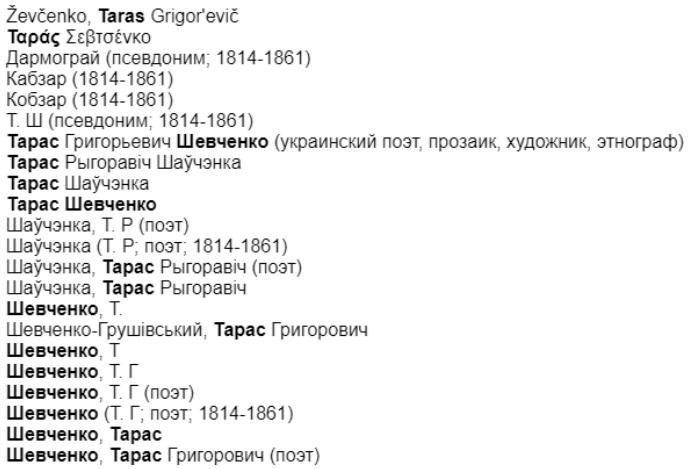

Figure 3. A subset of author name variations in the ISNI international authority file
Source: compiled by the author

Additionally, the use of different transliteration standards poses significant challenges for automated identification systems. For example, surname transliteration under the ISO standard may substantially differ from modern Ukrainian passport transliteration. This complicates the search process, particularly in multilingual databases where numerous spelling variants of the same name may exist. The issue is further exacerbated by the fact that for some languages, transliteration may only be normative in specific domains (e.g., passport records) without accommodating other specialised needs, such as library cataloguing or academic databases.

Different organisations or specialists may adopt distinct approaches to completing bibliographic records, influencing data structure and content. Consequently, information organisation structure remains pivotal in creating accessible and functional information systems (Joudrey & Taylor, 2017). For instance, some systems mandate all fields, including responsibility statements, whereas others only list the primary author or work title. Discrepancies in filling fields such as place of publication or publication year hinder record identification. Standardising these requirements would enhance the accuracy of automated data comparison. Even standard identifiers such as ISBN or ISSN do not always guarantee correct identification. For example, the same ISBN may be used for different works, particularly when publishers violate standards. The absence of an ISBN or ISSN in a record precludes automated verification. Duplicate searches often disregard information in linking fields, such as:

1. Unimarc 410 (*Series*), 411 (*Subseries*);
2. 461 (*Set*), 462 (*Subset*);
3. 482 (*Bound with...*), 454 (*Translated from...*).

Ignoring these fields may result in erroneous searches and record duplication. Another common issue arises when the primary ISBN field (e.g., 010 for Unimarc or 020 for MARC21) specifies: the ISBN of the original work instead of the translation; the ISBN of a book set instead of an individual volume. Such errors complicate precise record identification, as the system compares entries containing different data types. Depending on the standard, ISBNs may be recorded with hyphens (per DSTU and Unimarc) or without (MARC21). Consequently, the same number may be interpreted as two distinct records. In some cases, the ISBN field may contain additional information, such as clarifications in parentheses (ISBN 978-3-16-148410-0 (gift set)). Such practices violate standards and hinder automated analysis. These factors represent only a subset of issues affecting bibliographic record quality. Addressing them requires implementing data entry standards, developing effective automated processing tools, and enforcing quality control for bibliographic information.

CONCLUSIONS

To effectively resolve these issues, it is necessary to adopt standardised approaches to bibliographic record

completion, develop algorithms accounting for diverse standards, and refine normative control. The problem of bibliographic record duplication is systemic, stemming from technical, linguistic, cultural, and standardisation discrepancies. Primary negative factors include typographical errors, varying transliteration rules, lack of field completion standards, and unreliable identifiers such as ISBN. Thus, resolving these challenges necessitates implementing international standards for record unification, particularly MARC21 and IFLA LRM. The IFLA Library Reference Model (IFLA LRM) provides a conceptual framework for representing and accessing bibliographic information, aligning with modern information environment demands. More flexible search algorithms capable of accommodating multiple data representation variants should be developed. Normative control over bibliographic record creation must be reinforced, and artificial intelligence technologies for duplicate identification and processing should be advanced.

Future research prospects in this domain include the analysis of the effectiveness of existing duplicate detection algorithms across various bibliographic databases; the development of new record unification methodologies that account for multilingualism and diverse cultural contexts; the exploration of artificial intelligence integration for the automatic correction of typographical errors and transliteration; and a comparative analysis of the use of different identifiers, such as ISBN, DOI, and others, to enhance record accuracy.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Alves, V.H.M., Alves, V.A.M., & Cury, A.A. (2024). Artificial intelligence-driven structural health monitoring: Challenges, progress, and applications. In *New advances in soft computing in civil engineering. Studies in systems, decision and control* (pp. 149-166). Cham: Springer. doi: 10.1007/978-3-031-65976-8_7.

[2] Benevento, E., Aloini, D., & van der Aalst, W.M.P. (2022). How can interactive process discovery address data quality issues in real business settings? Evidence from a case study in healthcare. *Journal of Biomedical Informatics,* 130, article number 104083. doi: 10.1016/j.jbi.2022.104083.

[3] Berger, B., Waterman, M., & Yu, Y. (2021). Levenshtein distance, sequence comparison and biological database search. *IEEE Transactions on Information Theory*, 67(6), 3287-3294. doi: 10.1109/TIT.2020.2996543.

[4] Bockel-Rickermann, C., Verdonck, T., & Verbeke, W. (2023). Fraud analytics: A decade of research: Organizing challenges and solutions in the field. *Expert Systems with Applications*, 232, article number 120605. doi: 10.1016/j.eswa.2023.120605.

[5] Bogani, R., Theodorou, A., & Arnaboldi, L. (2023). Garbage in, toxic data out: A proposal for ethical artificial intelligence sustainability impact statements. *AI and Ethics,* 3, 1135-1142. doi: 10.1007/s43681-022-00221-0.

[6] Borissov, N., Haas, Q., Minder, B., Kopp-Heim, D., von Gernler, M., Janka, H., Teodoro, D., & Amini, P. (2022). Reducing systematic review burden using Deduklick: A novel, automated, reliable, and explainable deduplication algorithm to foster medical research. *Systematic Reviews*, 11, article number 172. doi: 10.1186/s13643-022-02045-9.

[7] Bruy, O. (2015). Marc21 format for the authoritative data - methodical elaboration and implantation in the library of NaUKMA. *Scientific Notes*, 22(1), 129-133.

[8] Budhiraja, A., Dutta, K., Shrivastava, M., & Reddy, R. (2018). Towards word embeddings for improved duplicate bug report retrieval in software repositories. In *Proceedings of the 2018 ACM SIGIR international conference on theory of information retrieval* (pp. 167-170). New York: Association for Computing Machinery. doi: 10.1145/3234944.3234949.

[9] Bykova, T.M. (2016). Specifics of editing bibliographic entries in the electronic catalogue: From experience of the Scientific Library of Odesa I.I. Mechnikov National University. *Bulletin of Odessa National University. Library Science, Bibliography, Book Studies*, 21(1), 137-149.

[10] Chauhan, R., Sharma, S., & Goyal, A. (2023). DENATURE: Duplicate detection and type identification in open source bug repositories. *International Journal of System Assurance Engineering and Management*, 14, 275-292. doi: 10.1007/s13198-023-01855-x.

[11] Coffman, A. (2021). *Unicode character look-alikes.* Retrieved from https://gist.github.com/StevenACoffman/a5f6f682d94e38ed804182dc2693ed4b.

[12] Discover your character: C. (n. d.). Retrieved from https://www.amp-what.com/unicode/search/c.

[13] Gesicho, M., Were, M., & Babic, A. (2020). Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: A case study of Kenya. *BMC Medical Informatics and Decision Making,* 20, article number 293. doi: 10.1186/s12911-020-01315-7.

[14] Gorman, M. (2004). Categories and normativity. In *Categories: Historical and systematic essays* (pp. 1-22). Washington: Catholic University of America Press.

[15] Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q., & Gao, R. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research*, 12, article number e44310. doi: 10.2196/44310.
[16] Hammer, B., Virgili, E., & Bilotta, F. (2023). Evidence-based literature review: De-duplication a cornerstone for quality. *World Journal of Methodology*, 13(5), 390-398. doi: 10.5662/wjm.v13.i5.390.
[17] Hillmann, D. (2004). *Metadata in practice*. Chicago: ALA Editions.
[18] Hopkinson, A. (2008). *Unimarc manual: Bibliographic format*. Munich: Walter de Gruyter.
[19] IFLA Library Reference Model (IFLA LRM). (2017). Retrieved from https://www.ifla.org/news/ifla-library-reference-model-lrm-march-2017-version-available/.
[20] Jahan, M., & Hasan, M. (2021). A robust fuzzy approach for gene expression data clustering. *Soft Computing*, 25, 14583-14596. doi: 10.1007/s00500-021-06397-7.
[21] Joudrey, D.N., & Taylor, A.G. (2017). *The organization of information*. Santa Barbara: Libraries Unlimited.
[22] Levenson, H.N., Amato, S., Bogus, I., Brody, F.E., Miller, M., & Nadal, L. (2024). Assessing bibliographic inaccuracy as a contributing factor for unintended loss in shared print monograph programs. *College & Research Libraries (C&RL)*, 85(7), article number 1021. doi: 10.5860/crl.85.7.1021.
[23] Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., & Serbin, O. (2021). Development of the combined method of identification of near duplicates in electronic scientific works. *Eastern-European Journal of Enterprise Technologies*, 4(4(112)), 57-63. doi: 10.15587/1729-4061.2021.238318.
[24] Lobuzina, K.V. (2013). Problems of creating comprehensive digital resources of historical, cultural and scientific heritage. *Special Historical Disciplines: Issues of Theory and Methodology*, 21, 121-129.
[25] Lunny, C., Pieper, D., Thabet, P., & Kanji, S. (2021). Managing overlap of primary study results across systematic reviews: Practical considerations for authors of overviews of reviews. *BMC Medical Research Methodology,* 21, article number 140. doi: 10.1186/s12874-021-01269-y.
[26] McKeown, S., & Mir, Z.M. (2021). Considerations for conducting systematic reviews: Evaluating the performance of different methods for de-duplicating references. *Systematic Reviews*, 10, article number 38. doi: 10.1186/s13643-021-01583-y.
[27] Negro-Calduch, E., Azzopardi-Muscat, N., Krishnamurthy, R.S., & Novillo-Ortiz, D. (2021). Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. *International Journal of Medical Informatics*, 152, article number 104507. doi: 10.1016/j.ijmedinf.2021.104507.
[28] Neysiani, S.B., Babamir, S.M., & Aritsugi, M. (2020). Efficient feature extraction model for validation performance improvement of duplicate bug report detection in software bug triage systems. *Information and Software Technology,* 126, article number 106344. doi: 10.1016/j.infsof.2020.106344.
[29] Page, L. (2005). The high cost of dirty data. *Materials Management in Health Care,* 14(11), 22-25.
[30] Raja, P., & Thangavel, K. (2020). Missing value imputation using unsupervised machine learning techniques. *Soft Computing*, 24(6), 4361-4392. doi: 10.1007/s00500-019-04199-6.
[31] Sehra, S.S., Abdou, T., Başar, A., & Sehra, S.K. (2020). Amalgamated models for detecting duplicate bug reports. In *Advances in artificial intelligence* (pp. 470-482). Cham: Springer. doi: 10.1007/978-3-030-47358-7_49.
[32] Thwel, T.T., & Sinha, G. (2021). *Data deduplication approaches: Concepts, strategies, and challenges*. Amsterdam: Elsevier Inc. doi: 10.1016/C2020-0-00104-0.
[33] Tillett, B.B. (1989). *Authority control in the online environment: Considerations and practices*. Santa Barbara: Libraries Unlimited.
[34] Walters, W., & Wilder, E. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, article number 14045. doi: 10.1038/s41598-023-41032-5.

Фактори ускладнення ідентифікації та обробки дублікатів у бібліографічних записах: теоретичний аспект

Олег Василенко

Аспірант
Інститут кібернетики імені В. М. Глушкова Національної академії наук України
03187, просп. Академіка Глушкова, 40, м. Київ, Україна
Викладач
Український гуманітарний інститут
08292, вул. Інститутська, 14, м. Буча, Україна
https://orcid.org/0000-0001-8498-2950

Анотація. У статті було розглянуто фактори, що створюють труднощі в процесі ідентифікації та обробки дублікатів у бібліографічних записах, які є важливою складовою інформаційних систем бібліотек, архівів та видавництв. Досліджено проблеми, спричинені друкарськими помилками, різними варіаціями транслітерації, використанням специфічних символів, омогліфами, різними правилами скорочення слів, варіаціями у написанні імен авторів, а також недоліками у використанні стандартних ідентифікаторів, таких як ISBN і ISSN. Особливу увагу було приділено впливу відмінностей у міжнародних та локальних стандартах MARC, Unimarc і Marc21 на створення та обробку бібліографічних даних. Проаналізовано, що неправильна обробка бібліографічних записів може призвести до зниження якості пошуку інформації користувачами, неможливості коректного цитування джерел та збільшення витрат часу на каталогізацію та індексацію. Більше того погіршується робота з бібліографічними даними у мультинаціональних системах через розбіжності в стандартах. У статті також проаналізовано наслідки цих проблем для функціонування бібліографічних систем, зокрема зниження точності пошукових запитів, ускладнення інтеграції даних між різними каталогами та підвищення витрат часу й ресурсів на обробку записів. Запропоновано низку рішень, серед яких: впровадження уніфікованих стандартів записів, використання сучасних алгоритмів адаптивного пошуку, що враховують мовні та технічні розбіжності, а також посилення нормативного контролю в процесі створення бібліографічних записів. Результати дослідження мають прикладне значення для розробників інформаційних систем, каталогізаторів і фахівців бібліотечної справи, адже сприяють вдосконаленню бібліографічних баз даних, зниженню дублювання записів і підвищенню якості пошуку інформації для кінцевих користувачів

Ключові слова: обробка даних; транслітерація; бібліотечні системи; інформаційні технології; нормативний контроль; каталогізація