Analysis of the deletions of DOIs
What factors undermine their persistence and to what extent?

Jiro Kikkawa, Masao Takaku, and Fuyuki Yoshikane
University of Tsukuba, Ibaraki, Japan
{jiro,masao,fuyuki}@slis.tsukuba.ac.jp

Abstract. Digital Object Identifiers (DOIs) are regarded as persistent; however, they are sometimes deleted. Deleted DOIs are an important issue not only for persistent access to scholarly content but also for bibliometrics, because they may cause problems in correctly identifying scholarly articles. However, little is known about how much of deleted DOIs and what causes them. We identified deleted DOIs by comparing the datasets of all Crossref DOIs on two different dates, investigated the number of deleted DOIs in the scholarly content along with the corresponding document types, and analyzed the factors that cause deleted DOIs. Using the proposed method, 708,282 deleted DOIs were identified. The majority corresponded to individual scholarly articles such as journal articles, proceedings articles, and book chapters. There were cases of many DOIs assigned to the same content, e.g., retracted journal articles and abstracts of international conferences. We show the publishers and academic societies which are the most common in deleted DOIs. In addition, the top cases of single scholarly content with a large number of deleted DOIs were revealed. The findings of this study are useful for citation analysis and altmetrics, as well as for avoiding deleted DOIs.

Keywords: Digital Object Identifier (DOI) · Scholarly Communication · Bibliometrics · Altmetrics.

1 Introduction

It is important to ensure persistent access to scholarly articles because many scholarly articles have been distributed on the Web. Digital Object Identifier (DOI) is the best-known persistent identifier of scholarly content and is an international standard. As of May 2022, 130 million DOIs have been registered by Crossref [5].

In the DOI system, a single and unique DOI is registered for each content item. A DOI name consists of a prefix, slash (/), and suffix. A prefix is assigned to a particular DOI registrant, such as publishing companies or academic societies. DOI registrants assign suffixes to their content and register DOIs through DOI registration agencies (hereinafter referred to as “RAs”). DOIs also provide hyperlinks (hereinafter referred to as “DOI links”) by adding DOI name
after “https://doi.org/.” Content holders, such as publishers and academic societies, maintain the relationship between the DOI and the original content’s URI. Through these functions, scholarly content can be consistently accessed using DOI links. Once registered, the DOIs are regarded as persistent and are not deleted.

Because of these characteristics of DOIs, citation indexes and altmetrics depend on them to identify unique scholarly articles. With the rapid increase in the number of scholarly articles, DOIs are playing an increasingly significant role in scholarly communication. However, in practice, DOIs and their artifacts (e.g., identifiers, metadata, and redirected URIs) may be deleted by content holders, such as publishers and academic societies, if something wrong happens, e.g., multiple DOIs were assigned to the same content item. These deletions cause problems not only for persistent access to scholarly content but also for bibliometric analysis, including citation analysis and altmetrics. In this study, we focus on the deletion of identifiers and metadata, and define DOIs for which identifiers and associated metadata cannot be retrieved as “Deleted DOIs.”

Deleted DOIs are an important issue in bibliometrics; however, little is known about their quantity and causes. Thus, we developed a methodology to identify deleted DOIs and analyze not only their quantity but also their causes, with a focus on the content holders and their patterns of DOI assignment.

The following research questions were addressed in this study:

**RQ1** How many deleted DOIs exist?
- We present the total number of deleted DOIs and classify deleted DOIs according to the redirect URIs and Crossref metadata.

**RQ2** Which document types are the most common in deleted DOIs?
- We investigate which document types have large numbers of deleted DOIs; i.e., we clarify whether these content items are single scholarly articles (such as a journal article, conference proceedings paper, or book chapter) or non-single scholarly articles (such as an entire book or a specific volume or issue of a journal). If a large number of deleted DOIs were registered for a single scholarly article, we analyze them in detail according to their metadata.

**RQ3** Which prefixes are the most common in deleted DOIs?
- We find publishers and academic societies with large numbers of deleted DOIs.

**RQ4** What are the most common changes in the suffixes of deleted DOIs?
- We investigate the pattern of changes in the suffixes of DOIs for the same scholarly content to determine the relationship between DOI assignment patterns and deleted DOIs.

The contributions of this study are twofold. (1) We present the overall picture of deleted DOIs by clarifying the number of deleted DOIs and the characteristics of their content. The potential impact of deleted DOIs for citation analysis and altmetrics is demonstrated. (2) By identifying the factors that cause deleted DOIs, we provide guidance for avoiding deleted DOIs and making the DOI system more stable.
2 Related Work

**Crossref DOI statistics.** Hendricks et al. [8] reported the statistics of Crossref DOIs in June 2019. More than 106 million Crossref DOIs had been registered, and the number of DOIs had increased by 11% on average over the past 10 years. As for the types of contents, 73% are journals, 13% are books, and 5.5% are conference papers and proceedings.

**Investigation of duplicated Crossref DOIs.** Tkaczyk [18] investigated Crossref DOIs not marked as an alias to other DOIs to consider their quantity and impact on citation-based metrics. Among DOIs randomly sampled from 590 publishers and academic societies with $\geq 5,000$ DOIs, 0.8% were duplicated, i.e., different DOI names but their metadata were the same or highly similar. The majority of them were caused by the re-registration of DOIs by the same publishers and academic societies. As for duplicated DOIs among different publishers and academic societies, one of the most frequent cases was content with DOIs initially registered by JSTOR and re-registered by new content holders.

**Incorrect DOIs indexed by scholarly bibliographic databases.** Several studies have revealed errors in DOIs indexed by scholarly bibliographic databases. Franceschini et al. [7] analyzed DOIs in the records of Scopus and found that multiple DOIs were incorrectly assigned to the same record as rare cases. Zhu et al. [19] analyzed DOIs in the Web of Science records. They reported not only “wrong DOI names” but also “one paper with two different DOI names.” The former are similar errors, as reported by Franceschini et al. [7]. The latter are classified into the following two cases: (1) there were both correct and incorrect DOIs in the records; (2) multiple correct DOIs were assigned to the same scholarly article.

**Analysis of persistence of Crossref DOIs.** Klein and Balakireva [12, 13] examined the persistence of Crossref DOIs by analyzing their HTTP status codes. They randomly extracted 10,000 Crossref DOIs and examined the final status codes for each DOI link by using multiple HTTP request methods. More than half of the DOI links did not redirect to the content when an external network from academic institutions was used. However, the errors of all the DOI links were reduced to one-third when an internal network from academic institutions was used. These results indicate that the responses for the same DOI can differ according to conditions such as the HTTP request methods and network locations, which implies a lack of persistence of DOIs.

**Analysis of the usage of DOI links in scholarly articles.** Regarding the usage of DOI links in the references of scholarly references, Van de Sompel et al. [16] examined references from 1.8 million papers published between 1997 and 2012. Consequently, they identified a problem that numerous scholarly articles were referenced using their location URIs instead of their DOI links.

As described previously, researchers have reported duplicated Crossref DOIs [18, 19], and some Crossref DOIs cause errors and are unable to lead to the contents [12, 13]. It has also been reported that DOIs included in the records of
scholarly bibliographic databases contain errors \cite{7, 19}. However, these studies were based on relatively small samples because a methodology to identify duplicated or deleted DOIs at a large scale has not been proposed. Thus, we propose a methodology for identifying deleted DOIs and conducting a large-scale analysis to determine how many deleted DOIs exist and what factors cause them.

3 Materials and Methods

3.1 Identifying Deleted DOIs

In this section, we describe the proposed method for identifying deleted DOIs, i.e., DOIs whose identifiers and metadata cannot be retrieved. Specifically, we extracted the DOIs that existed at a point in time and did not exist later by comparing the dump files of Crossref DOIs on two different dates. We then treated these DOIs as a candidate set of deleted DOIs.

First, we obtained the two datasets of all Crossref DOIs on different dates: March 2017 \cite{9} and January 2021 \cite{2, 10}. We unescaped and downcased to remove duplicates. The total numbers of unique DOIs were 87,542,203 and 120,684,617, respectively.

Next, we obtained the differences and product sets between the two datasets, as shown in Table 1. There were 711,198 Crossref DOIs that existed as of March 2017 but did not exist as of January 2021. Then, for the 711,198 DOIs, we (1) identified the RA, (2) examined the redirected URI of DOI links, and (3) obtained the Crossref metadata. For (1), we used a web API called “Which RA?” \cite{17}, which returns the RA name if the DOI exists and returns an error if it does not exist. For (2), we examined the HTTP header of each DOI link. For (3), we used the Crossref REST API \cite{3, 4}.

According to the results for (1) to (3), we classified the DOIs into the groups shown in Fig. 1. First, we classified them into “Non-existing DOIs,” “Crossref DOIs,” and “Non-Crossref DOIs” according to the results of “Which RA?”; these groups were derived from the results of “DOI does not exist,” “Crossref,” and RA names other than Crossref, respectively. Among these groups, Non-existing DOIs corresponded to deleted DOIs. In contrast, Non-Crossref DOIs were not deleted, because they were not included in the dataset as of March 2021, owing to the transfer of the RAs.
Fig. 1. Classification of the DOIs that existed as of March 2017 but no longer existed as of January 2021. Items with a red box correspond to deleted DOIs.

Next, we classified the Crossref DOIs into the groups “Defunct DOIs,” “DOIs without Redirects,” and “DOIs with Redirects” according to the results of each HTTP header for the DOI links. “Defunct DOIs” were the DOIs that redirected to the specific URI (https://www.crossref.org/_deleted-doi/) for the deleted content [15]. “DOIs without Redirects” were the DOIs without redirects when we accessed DOI links; thus, we could not reach the content via these DOIs. We treated defunct DOIs and DOIs without redirects as deleted DOIs.

Finally, the DOIs were classified into three groups according to the results of the Crossref REST API. They were all deleted DOIs, because they were unable to retrieve the Crossref metadata.

– **Alias DOI:** When multiple Crossref DOIs were assigned to the same content item, one was set as the Primary DOI, and the others were set as the Alias DOIs [6]. The Crossref REST API returned the error “Resource not found” for the Alias DOIs.

– **DOI with Deleted Description in Metadata:** Crossref metadata for a particular DOI or the corresponding Alias DOI contains the word “delete” in the value of “title” or “container title.” Moreover, the first author manually confirmed that the content was deleted by accessing the landing page.

– **Other DOI:** These DOIs had redirects that were not applicable to any of the conditions above.

When the DOIs were applicable to both the “Alias DOI” and “DOI with Deleted Description in Metadata” groups, we assigned them to the latter group.

### 3.2 Analysis Methods

The following four analyses were performed on the deleted DOIs identified by the method described in Section 3.1. Each analysis corresponds to an RQ.

1. **Basic Statistics of Deleted DOIs.** We investigated the number of deleted DOIs and their distributions using the groups presented in Section 3.1.

2. **Content Analysis Based on Crossref Metadata and DOI Links.**

   First, we determined which content types were common among the deleted DOIs
according to the value of "type" in the Crossref metadata. Next, we investigated the number of alias DOIs corresponding to a single Primary DOI to clarify the characteristics of deletion when multiple DOIs were registered to the same content. For example, when Alias DOIs "10.14359/15303," "10.14359/15304," and "10.14359/15305" corresponded to the same Primary DOI "10.14359/15306," the number of Alias DOIs was 3. We obtained the basic statistics of the Alias DOIs corresponding to the same Primary DOIs. We examined articles for cases in which many Alias DOIs were associated with a single Primary DOI.

(3) Prefix Analysis. We compared the prefixes of deleted DOIs and identified which publishers or academic societies had large numbers of deleted DOIs. In addition, we calculated the ratios $P_1$ and $P_2$ for each prefix, as follows.

$$P_1(X) = \frac{\text{total number of the Deleted DOIs whose Prefix was } X}{\text{total number of all Deleted DOIs}} \times 100$$

$$P_2(X) = \frac{\text{total number of the Deleted DOIs whose Prefix was } X}{\text{total number of all DOIs whose Prefix was } X \text{ as of March 2017}} \times 100$$

When both $P_1$ and $P_2$ are high in a prefix, it indicates not only that the percentage of deleted DOIs for this prefix in the total deleted DOIs is high but also that a large percentage of the DOIs registered by the registrant with this prefix are deleted. On the other hand, when $P_1$ is high and $P_2$ is low in a prefix, the percentage of deleted DOIs in the total deleted DOIs is high but a small percentage of the DOIs registered by the registrant with that prefix are deleted.

(4) Suffix Analysis. We analyzed the suffixes of the deleted DOIs when multiple DOIs were registered for the same content. Here, we set the target to the pairs of Alias DOIs and Primary DOIs. First, we classified the patterns of the pairs of Alias DOI and Primary DOI suffixes into the following categories: (A) only the suffix changed, (B) both the prefix and the suffix changed, and (C) only the prefix changed. Next, for (A) and (B), we analyzed the similarity of strings between the suffixes of the Alias DOI and Primary DOI. The Levenshtein distance [14] normalized by the string length of the suffixes was used as the similarity score. The following formula was used:

$$\text{Sim}(str1, str2) = 1 - \frac{\text{Levenshtein}(str1, str2)}{\text{Max}(|str1|, |str2|)}$$

where $\text{Sim}$ takes values between 0 and 1, with higher values indicating a higher degree of similarity between the two strings. We observed the distributions for the similarity score in groups (A) and (B). When the similarity scores of the majority in a group are high, it can be interpreted that there are many cases where slightly changed suffixes of Alias DOIs are set as those of Primary DOIs. In contrast, when the similarity score of the majority in a group is low, it can be interpreted that there are many cases where strings that differ significantly from the Alias DOI suffixes are set as the suffixes of Primary DOIs.
Finally, we analyzed the character-level differences between each pair of Alias and Primary DOIs to determine what types of changes were common. We used the package “diff-lcs” [20] from RubyGems to obtain the diffs of the Primary and Alias DOIs using the Diff::LCS.diff method. We retrieved the changes between each Primary DOI and Alias DOI as actions of adding or deleting the position and the corresponding characters. We analyzed which characters were added, deleted, or replaced (when both “delete” and “add” were present in the same position) for each pair. For example, in the case of the Alias DOI “10.14359/15303” and Primary DOI “10.14359/15306,” the change was to replace “3” with “6” once.

4 Results and Discussion

4.1 Basic Statistics of Deleted DOIs

RQ1 How many deleted DOIs exist?

Table 2 presents the number of deleted DOIs in each group. We identified 708,282 deleted DOIs by applying the method described in Section 3.1. Most deleted DOIs were Alias DOIs, accounting for 94.29% of the total. Therefore, the majority of the deleted DOIs were caused by the deletion of multiple DOI assignments for the same content. The percentage of deleted DOIs relative to all DOIs in 2021 was 0.59% (\( \frac{708,282}{120,684,617} \times 100 \)).

Table 2. Number of deleted DOIs in each group (n=708,282)

| # | Group                              | Count | %   |
|---|------------------------------------|-------|-----|
| 1 | Non-existing DOIs                  | 240   | 0.03|
| 2 | Defunct DOIs                       | 388   | 0.05|
| 3 | DOIs without redirects             | 693   | 0.10|
| 4 | Alias DOIs                         | 667,869 | 94.29|
| 5 | DOIs with deleted description on metadata | 1,144 | 0.16|
| 6 | Other DOIs                         | 37,948 | 5.36|

4.2 Content Analysis Based on Crossref Metadata and DOI Links

RQ2 Which document types are the most common in deleted DOIs?

Table 3 presents the document types of the deleted DOIs based on Crossref metadata. The top three types were journal articles (78.35%), proceedings articles (8.48%), and book chapters (6.77%). Thus, most of the deleted DOIs were individual scholarly articles rather than books or journals.

The minimum value, maximum value, median, and standard deviation of the number of alias DOIs corresponding to each primary DOI were 1, 1,474, 1, and 2.49, respectively. The majority of cases were those in which two DOIs were registered for the same content, and one of them was deleted because the median value was 1. On the other hand, focusing on the maximum value, there were cases where a large number of DOIs were assigned to a single content item.
Table 3. Document types of deleted DOIs (n=708,282)

| #   | Type            | Count   | %    | #   | Type            | Count | %   |
|-----|-----------------|---------|------|-----|-----------------|-------|-----|
| 1   | journal-article | 554,948 | 78.35| 10  | standard        | 70    | 0.01|
| 2   | proceedings-article | 60,092 | 8.48 | 11  | book-series     | 48    | 0.01|
| 3   | book-chapter    | 47,974  | 6.77 | 12  | report          | 40    | 0.01|
| 4   | dataset         | 23,784  | 3.36 | 13  | journal-issue   | 37    | 0.01|
| 5   | component       | 8,419   | 1.19 | 14  | proceedings     | 23    | 0.00|
| 6   | reference-entry | 6,611   | 0.93 | 15  | reference-book  | 21    | 0.00|
| 7   | book            | 3,094   | 0.44 | 16  | journal         | 15    | 0.00|
| 8   | monograph       | 2,340   | 0.33 | 17  | dissertation    | 10    | 0.00|
| 9   | other           | 754     | 0.11 | 18  | book-section    | 2     | 0.00|

Table 4 presents the 10 Primary DOIs with the largest numbers of associated Alias DOIs. The content item with the largest number of associated Alias DOIs was Volume 13 of the journal Energy Procedia (#1 in Table 4). This entire volume was retracted. These articles are different scholarly articles, and it is considered that the withdrawn information should be applied to the respective DOIs, rather than applying them as alias DOIs to a single primary DOI. In case #2, all articles published in Volume 11 of Energy Procedia were withdrawn. Case #3 involved abstracts of an international conference, where DOIs were registered to the abstracts of the presentations, but the publisher appeared to change the policy for registering DOIs to a set of abstracts. Cases #4–#9 were similar to #3 and all involved the Journal of the American College of Cardiology. Case #10 involved an abstract of an international conference published in another journal. According to these results, the Primary DOIs with large numbers of Alias DOIs corresponded to content such as retracted journal articles and abstracts of international conferences. This content would not be the target of citation analysis and altmetrics and did not have a critical impact on them. The Primary DOIs with large numbers of Alias DOIs were caused by the activities of publishers and academic societies rather than by the deletion of multiple DOIs mistakenly registered to the same scholarly articles.

4.3 Prefix Analysis

RQ3 Which prefixes are the most common in deleted DOIs?

Table 5 presents the top 10 prefixes of the deleted DOIs. $P_1$ and $P_2$ in Table 5 are the two ratios defined in Section 3.2. For example, the publisher Wiley registered 4,718,360 Crossref DOIs with the prefix “10.1002” as of March 2017, and 94,471 of them were Deleted DOIs. In this case, $P_1$ was 13.34% (= 94,471/708,282 * 100), and $P_2$ was 2.00% (= 94,471/4,718,360 * 100). As indicated by Table 5, the registrants with the largest numbers of deleted DOIs were major international commercial publishers, such as Wiley, Elsevier BV, and Springer Science and Business Media LLC (#1, #2, #9, and #10). These publishers registered numerous DOIs, but few of them were deleted, as the $P_2$
values for these publishers were low. In contrast, the “Society of Petroleum Engineers” (#5) had a remarkably high $P_2$ value of 87.62%; thus, most of the DOIs registered by this registrant with this prefix were deleted. IGI Global and the American Psychological Association (#7 and #4) had relatively high $P_2$ values of 17.51% and 6.74%, respectively. JSTOR (#3)—a digital library platform—was a registrant with many deleted DOIs. Regarding JSTOR, the deletions of DOI seem to be caused by the re-registration of DOIs to the same article published by content holders other than JSTOR. For these cases, it is necessary to assign a single DOI to the initial version of the article and avoid re-registration of the DOIs to the other versions. This problem would be caused by the registration agency improperly managing their content.

4.4 Suffix Analysis

RQ4 What are the most common changes in the suffixes of deleted DOIs?
Table 5. Top 10 prefixes of deleted DOIs (n=708,282)

| #  | Prefix      | Registrant                              | Count | P₁ (%) | P₂ (%) |
|----|-------------|-----------------------------------------|-------|--------|--------|
| 1  | 10.1002     | Wiley                                   | 94,471| 13.34  | 2.00   |
| 2  | 10.1016     | Elsevier BV                             | 85,643| 12.09  | 0.58   |
| 3  | 10.2307     | JSTOR                                   | 66,232| 9.35   | 2.82   |
| 4  | 10.1037     | American Psychological Association      | 51,846| 7.32   | 6.74   |
| 5  | 10.2523     | Society of Petroleum Engineers          | 35,499| 5.01   | 87.62  |
| 6  | 10.1163     | Brill                                   | 23,962| 3.38   | 2.68   |
| 7  | 10.4018     | IGI Global                              | 23,440| 3.31   | 17.51  |
| 8  | 10.1080     | Informa UK Limited                      | 21,134| 2.98   | 0.75   |
| 9  | 10.1007     | Springer Science and Business Media LLC | 19,807| 2.80   | 0.21   |
| 10 | 10.1111     | Wiley                                   | 18,743| 2.65   | 0.64   |

Table 6. DOI name change patterns (n=667,869)

| Pattern                                           | Count | %    |
|---------------------------------------------------|-------|------|
| Only the suffix changed                           | 465,789| 69.74|
| Both the prefix and the suffix changed            | 156,650| 23.46|
| Only the prefix changed                           | 45,430 | 6.80 |

Table 6 presents the DOI name change patterns for each pair of Alias and Primary DOIs. The most common pattern was “only the suffix changed” (69.74%), indicating that many DOIs registered under the same prefix were deleted when multiple DOIs were registered to the same content by the same registrant. In contrast, “both the prefix and the suffix changed” and “only the prefix changed” were cases of deletions in which the DOIs were re-registered with the content holder’s prefix when the content holder changed. For the prefixes “10.4018 (IGI Global)” and “10.1037 (American Psychological Association)” (#7 and #4, respectively, in Table 5), almost all the patterns were “only the suffix changes.”

Of the 156,650 cases in which both the prefix and the suffix changed, 56,491 were caused by the prefix “10.2307 (JSTOR).” The prefix “10.2307” was changed to “10.1090 (American Mathematical Society)” and “10.1080 (Informa UK Limited)” in 22,198 and 10,010 cases, respectively. These results match those of Tkaczyk [18] and indicate that a large amount of scholarly content provided by JSTOR was transferred to other registrants, and their DOIs were re-registered with the prefixes of new registrants, which is why the prefix “10.2307” is listed as #3 in Table 5.

Of the 45,430 cases of “only the prefix changed,” 35,176 involved the prefix “10.2523” of Alias DOIs and the prefix “10.2118” of Primary DOIs, and the corresponding registrant was the Society of Petroleum Engineers. This is why the prefix “10.2523” is listed as #5 in Table 5, and these cases were due to DOIs being re-registered by the same publisher.

Table 7 presents the distribution of similarity scores among suffixes for “only the suffix changed” and “both the prefix and the suffix changed.” The group
Table 7. Distribution of similarity scores between Alias DOI and Primary DOI suffixes

|                    | Only the suffix changed | Both the prefix and the suffix changed |
|--------------------|-------------------------|----------------------------------------|
|                    | Count | %    | Count | %    |
| 0 ≤ sim ≤ 0.1      | 9,033 | 1.94 | 15,942 | 10.18 |
| 0.1 < sim ≤ 0.2    | 35,253 | 7.57 | 80,420 | 51.34 |
| 0.2 < sim ≤ 0.3    | 20,058 | 4.31 | 34,385 | 21.95 |
| 0.3 < sim ≤ 0.4    | 9,539  | 2.05 | 13,278 | 8.48  |
| 0.4 < sim ≤ 0.5    | 27,074 | 5.81 | 6,493  | 4.14  |
| 0.5 < sim ≤ 0.6    | 27,783 | 5.96 | 2,243  | 1.43  |
| 0.6 < sim ≤ 0.7    | 26,331 | 5.65 | 878    | 0.56  |
| 0.7 < sim ≤ 0.8    | 57,866 | 12.42| 536    | 0.34  |
| 0.8 < sim ≤ 0.9    | 82,647 | 17.74| 2,361  | 1.51  |
| 0.9 < sim < 1.0    | 170,205 | 36.54| 114    | 0.07  |
| Overall            | 465,789 | 100.00| 156,650 | 100.00|

“only the suffix changed” had a large percentage of cases with Sim values of > 0.7, indicating that most of the cases had a high degree of similarity among suffixes. In contrast, the group “both the prefix and the suffix changed” had a large percentage of cases with Sim values of ≤ 0.3, indicating that the degree of similarity among the suffixes was low. This suggests that a slightly changed version of the suffix of the Alias DOI tended to be set as the suffix of the Primary DOI for the group “only the suffix changed.” In contrast, the group “both the prefix and the suffix changed” have a tendency that a significantly different suffix from the Alias DOI was set as the suffix of the Primary DOI.

Table 8. Three most frequent patterns for “only the suffix changed” (n=465,789). Red, blue, and purple text corresponds to deletion, addition, and replacement, respectively.

| #  | Pattern of changes in the suffix | Count | %   |
|----|---------------------------------|-------|-----|
| 1  | Delete a slash (/) once.         | 49,169| 10.56|
|    | /s12445-012-0033-7 → s12445-012-0033-7 |       |      |
| 2  | Add a hyphen (-) four times.     | 21,918| 4.71 |
|    | 9781591401087.ch001 → 978-1-59140-108-7.ch001 |       |      |
| 3  | Delete “2” twice, add “5” once, add “7” once, replace “8” with “9” once, and replace “6” with “3” once. | 18,607| 3.99 |
|    | 2214-8647_dnp_e1000010 → 1574-9347_dnp_e1000010 |       |      |
Table 8 presents frequent patterns for “only the suffix changed.” The most frequent pattern was the correction of “//” to “/” as a separator between the prefix and the suffix (10.56%). The second most frequent pattern was the correction of the ISBN in the suffix by adding hyphens as separators (4.71%). The third most frequent pattern was the replacement of the ISSN in the suffix with a new one (3.99%). According to these results, we must be careful not to set double slashes between the prefix and the suffix. When we apply other identifiers such as the ISBN and ISSN to suffixes, we may need to format or update them owing to changes in the ISBN or ISSN.

5 Conclusion

We identified and analyzed the deleted DOIs. As a result, we identified 708,282 deleted DOIs that existed in March 2017 and did not exist in January 2021 using the proposed method. The majority of these DOIs were individual scholarly articles such as journal articles, proceedings articles, and book chapters. The cases where many DOIs were assigned to the same content were retracted papers in a specific volume of a journal or abstracts of international conference proceedings. We revealed the factors that caused a large number of deleted DOIs. The findings of this study are useful for both considering the problems caused by deleted DOIs in citation analysis and altmetrics and assigning DOIs in a better way to avoid deleted DOIs.

The limitations of this study are outlined below, and directions for future work are suggested. First, we were unable to identify deleted DOIs that existed before March 2017, i.e., DOIs that had been deleted as of March 2017. We were also unable to capture the deleted DOIs after March 2021. Other DOIs whose contents were not reachable were beyond the scope of this study. The deleted DOIs in this study were defined only as deletions of identifiers and metadata. Therefore, we will expand the scope of the deleted DOIs and develop a methodology for identifying them beyond this definition. Second, we were unable and classify all the factors that caused deleted DOIs and classify them, because we only focused on the most frequent cases. Moreover, it is unclear whether the observed factors represented exceptional cases. Third, a quantitative analysis of the effects of deleted DOIs on citation analysis and altmetrics is important, e.g., examining whether Alias/Primary DOIs are included in the citation indexes and altmetrics. Recently, DOI assignments to the contents on preprint servers such as arXiv have been actively conducted [1], and there is a concern that the numbers of duplicated DOIs and deleted DOIs will increase. Analyzing these DOIs is beyond the scope of this study. However, it is important to understand the stability and persistence of the DOI system in the future. The dataset used in this study is available at Zenodo [11].

Acknowledgments. This work was partially supported by JSPS KAKENHI Grant Numbers JP21K21303, JP22K18147, JP20K12543, and JP21K12592. We would like to thank Editage (https://www.editage.com/) for the English language editing.
References

1. Cornell University: New arXiv articles are now automatically assigned DOIs | arXiv.org blog (2022), https://blog.arxiv.org/2022/02/17/new-arxiv-articles-are-now-automatically-assigned-dois/
2. Crossref: January 2021 Public Data File from Crossref. Academic Torrents. https://doi.org/10.13003/gu3dqnjyvg4
3. Crossref: Crossref Metadata API JSON Format (2021), https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md
4. Crossref: Crossref REST API (2021), https://api.crossref.org/
5. Crossref: crossref.org : : crossref stats (2022), https://www.crossref.org/06members/53status.html
6. Farley, I.: Conflict report - Crossref (2020), https://www.crossref.org/documentation/reports/conflict-report/
7. Franceschini, F., Maisano, D., Mastrogiacomo, L.: Errors in DOI indexing by bibliometric databases. Scientometrics 102(3), 2181–2186 (2015). https://doi.org/10.1007/s11192-014-1503-4
8. Hendricks, G., Tkaczyk, D., Lin, J., Feeley, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies 1(1), 414–427 (2020). https://doi.org/10.1162/qss_a_00022
9. Himmelstein, D., Wheeler, K., Greene, C.: Metadata for all DOIs in Crossref: JSON MongoDB exports of all works from the Crossref API. figshare (2017). https://doi.org/10.6084/m9.figshare.4816720.v1
10. Kemp, J.: New public data file: 120+ million metadata records (2021), https://www.crossref.org/blog/new-public-data-file-120-million-metadata-records/
11. Kikkawa, J., Takaku, M., Yoshikane, F.: Dataset of the deleted DOIs extracted from the difference set between Crossref DOIs as of March 2017 and January 2021. Zenodo (2022). https://doi.org/10.5281/zenodo.6841257
12. Klein, M., Balakireva, L.: On the Persistence of Persistent Identifiers of the Scholarly Web. In: Digital Libraries for Open Knowledge. pp. 102–115. Lecture Notes in Computer Science, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-54956-5_8
13. Klein, M., Balakireva, L.: An extended analysis of the persistence of persistent identifiers of the scholarly web. International Journal on Digital Libraries 23, 5–17 (2021). https://doi.org/10.1007/s00799-021-00315-w
14. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady 10(8), 707–710 (1966)
15. Smulyan, S.: Defunct DOI - Crossref (2020), https://www.crossref.org/_deleted-doi/
16. Van de Sompel, H., Klein, M., Jones, S.M.: Persistent URIs Must Be Used To Be Persistent. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 119–120. WWW ’16 Companion, International World Wide Web Conferences Steering Committee (2016). https://doi.org/10.1145/2872518.2889352
17. The International DOI Foundation: Factsheet DOI Resolution Documentation - 4. Which RA? (2020), https://www.doi.org/factsheets/DOIProxy.html#whichra
18. Tkaczyk, D.: Double trouble with DOIs - Crossref (2020), https://www.crossref.org/blog/double-trouble-with-dois/
19. Zhu, J., Hu, G., Liu, W.: DOI errors and possible solutions for Web of Science. Scientometrics 118(2), 709–718 (2019). https://doi.org/10.1007/s11192-018-2980-7
20. Ziegler, A.: halostatue/diff-lcs: Generate difference sets between Ruby sequences. (2022), https://github.com/halostatue/diff-lcs