Scientometrics (2019) 120:1427–1437
https://doi.org/10.1007/s11192-019-03162-4

Types of DOI errors of cited references in Web of Science with a cleaning method

Shuo Xu[1] · Liyuan Hao[1] · Xin An[2] · Dongsheng Zhai[1] · Hongshen Pang[3]

Received: 23 April 2019 / Published online: 11 July 2019
© Akadémiai Kiadó, Budapest, Hungary 2019

Abstract
Though the bibliographic databases, such as Web of Science (WoS), largely promote the development of scientometrics and informetrics, these databases are not free of errors. The main purpose of this work is to figure out which types of DOI errors of cited references exist, how often each type of errors occur, and whether it is possible to automatically correct these errors. After careful analysis, several classic DOI errors of cited references, such as prefix-, suffix- and other-type errors, are identified, Then, a cleaning method is put forward on the basis of regular expressions. Experimental results on the bibliographic data in the *gene editing* field from the WoS database indicate that our cleaning approach can improve largely the quality of DOI names of cited references.

Keywords DOI errors · Cleaning method · Web of Science · Cited references · Regular expression

This work was supported partially by the Social Science Foundation of Beijing [Grant Number 17GLB074] and Natural Science Foundation of Guangdong Province under Grant Number 2018A030313695.

✉ Xin An
anxin@bjfu.edu.cn

Shuo Xu
xushuo@bjut.edu.cn

Liyuan Hao
Leanne.H@qq.com

Dongsheng Zhai
zhaidongsheng@bjut.edu.cn

Hongshen Pang
phs@szu.edu.cn

[1] Research Base of Beijing Modern Manufacturing Development, College of Economics and Management, Beijing University of Technology, Beijing 100124, People's Republic of China

[2] School of Economics and Management, Beijing Forestry University, Beijing 100083, People's Republic of China

[3] Library, Shenzhen University, Shenzhen 518060, People's Republic of China

Introduction

With the establishment of digital object identifier (DOI) system in 1997 (Paskin 1999), managed by the International DOI Foundation (IDF) (Chandrakar 2006; Paskin 1999, 2010; Simmonds 1999), DOIs have been assigned uniquely to many digital objects, such as publications (Boundry and Chartron 2017; Gorraiz et al. 2016), illustrations or tables (Wang 2007), scientific data (Neumann and Brase 2014) and so on. The DOI name is a case-insensitive alphanumeric string, and consists of two parts separated by a forward slash (Sidman and Davidson 2001; Simmonds 1999): (a) a prefix beginning with the numeral 10 assigned by IDF or by DOI registration agencies, and (b) a suffix assigned by the registrants .

It is well known that comprehensive bibliographic databases, such as Scopus and Web of Science (WoS), largely promote the development of scientometrics and informetrics. However, one should keep in mind that these databases are not free of errors (Jacso 2006; Franceschini et al. 2013, 2014, 2016), though data quality has improved significantly over the past decade. Buchanan (2006) divided the database errors into two categories: (a) *author errors* and (b) *database mapping errors*. The illustrative examples are also given for each type of database errors in Buchanan (2006).

So far, errors have been found to happen to almost each field of publications, such as author names (Buchanan 2006), author address (Liu et al. 2018), publication year (Buchanan 2006), omitted citations (Franceschini et al. 2014), funding acknowledge (Tang et al. 2017), etc. The documents are also even missed from the bibliographic database (Krauskopf 2019). Of course, it is no exception for the DOI field. Franceschini et al. (2015) revealed that quite a few single DOI names were incorrectly assigned to multiple publications indexed in the Scopus database. The incorrect DOI names in the WoS database are also discovered by Zhu et al. (2019), Zhu et al. (2019) and Huang and Liu (2019).

By definition, each DOI name should be unique and must identify one and only one entity (Paskin 1999). Thus, one can utilize DOI names to identify and disambiguate the scientific publications. However, DOI errors present challenges for the data collection from different sources in order to avoid unwanted duplicate entries (Valderrama-Zurián et al. 2015), the application of new metrics, like altmetrics (Jobmann et al. 2014; Haustein et al. 2015), the accuracy of thematic structures extraction (Xu et al. 2018), and so on. In fact, apart from DOI errors described in Franceschini et al. (2015), Zhu et al. (2019) and Zhu et al. (2019), it remains unknown that whether there are other types of DOI errors, how often each type of errors occur, and whether it is possible to automatically correct these errors. In this work, various DOI errors of cited reference in the WoS database are deeply analyzed and a cleaning approach is put forward to alleviate the extent of DOI errors of cited references.

Dataset

The bibliographic data in the *gene editing* field was collected from the WoS core database on 25th January, 2018 from the library of Beijing University of Technology. The following search strategy is used in this study: "TS = (gene edit*) OR TS

Table 1 Distribution of number of publications over year for *gene editing* dataset

Pub. year	No. of pub.	Pub. year	No. of pub.	Pub. year	No. of pub.
2000	449	2006	208	2012	589
2001	443	2007	235	2013	747
2002	411	2008	266	2014	1093
2003	447	2009	320	2015	1572
2004	245	2010	380	2016	2637
2005	223	2011	456	2017	3277

Fig. 1 A snippet of the cited references in the WoS database

= (crispr) OR TS = (clustered regularly interspaced short palindromic repeats)". The language is limited to English, and the document type includes *article*, *proceedings paper* and *review*. The publication year spans from 2000 to 2017. The downloaded scientific publications contain the full records and resulting cited references in the tab-delimited file format. It is very surprised that two records with the same DOI (10.3389/FIMMU.2017.00351) in the retrieved results have different WoS IDs (WOS:000398414900001 and WOS:000399835000001). On closer examination by manual, these two records are found to refer to the same publication, so one record is removed directly. In total, the number of publications is 13,909 and Table 1 reports the distribution of number of publications over year.

To discover various DOI errors of cited references, one should delve into the reference list, i.e. CR field in the WoS database. Each cited publication in the reference list is usually only shown the following fields: the first author's name (family name and surname's initials), publication year, abbreviated publication venue (e.g., journal or conference name), volume number starting with the character "V", starting page number beginning with the letter "P", and DOI name with the prefix "DOI␣". If a field (such as *starting page number*) is unavailable for some articles, the related information is directly missed from the cited publication in the reference list. Here, "␣" denotes the whitespace character. These fields are separated by ",␣", and the cited references are delimited by ";␣". A snippet of the cited references is illustrated in the Fig. 1. It is worth mentioning that multiple DOI names can be attached to the same cited reference (cf. the second one in Fig. 1). In this case, multiple DOI names are enclosed by square brackets with the delimiter ",␣". If a cited reference has no DOI name, the prefix "DOI␣" is omitted directly (cf. the third one in Fig. 1).

According to whether the cited reference attaches a DOI name or not, the cited references are divided into two categories: the cited references with DOIs and those without DOIs. The number of the cited references with and without DOIs is 341,317 and 74,643 respectively. Due to the difficulty and workload of filling with the resulting

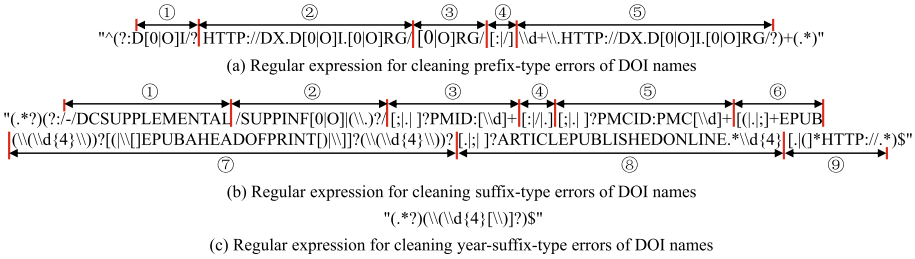


Fig. 2 Regular expressions for cleaning various DOI errors

DOI names for the latter, the cited references without DOIs are excluded from further analysis in this study.

Cleaning method

Through careful analysis, this study finds that various DOI errors of the cited references exist in the WoS database. That is to say, DOI names of cited references in the WoS database are contaminated to some extent. As a matter of fact, due to the variety of DOI errors, it is not trivial to clean automatically DOI names. To the best of our knowledge, no softwares public available can competent for this cleaning task until now. Hence, a method for cleaning DOI names is proposed in this work, as shown in Algorithm 1. On the basis of manual curation rules, this approach is made up of one procedure (*Cleaning*) and three functions (*JoinDoi*, *TrimDoi* and *IsBracketMatch*). To facilitate the understanding, many data types and built-in functions from Java programming language, shown in sans serif font family font style, are explicitly utilized here. For more elaborate and detailed description on these data types and built-in functions, we refer the readers to Java API reference.

The procedure *Cleaning* takes a cited reference (CR) field of an interested publication as input, splits it into multiple cited references (Line 2 in Algorithm 1), and then try to separate DOI name(s) from other information one by one (Line 3–14 in Algorithm 1). This study mainly focuses on various DOI errors, the cited references without the clue substring "␣DOI" are discarded directly. The cited references with DOI name(s) are further grouped into two cases: those with multiple DOI names (Line 7–10 in Algorithm 1) and those with single DOI name (Line 11 in Algorithm 1). Note that it is very possible that for the former case (multiple *literal* DOI names), only one DOI name is actually output (e.g., id = 1 and 5 in Table 5). The function *JoinDoi* devotes to removing the duplicate DOI names processed by the function *TrimDoi*.

The function *TrimDoi* tries to trim the DOI names by several regular expressions. Though most legal Unicode characters are allowed by ISO standard (ISO 26324:2012–Information and documentation–digital object identifier system), it is very seldom that DOI names contain whitespace characters. Exceptions are still found in the WoS database, such as id = 18 (2nd one) in Table 5. Hence, before cleaning further DOI names, all whitespace characters are removed (Line 32 in Algorithm 1), Then, prefix-type

Table 2 Mapping between regular expressions in Fig. 2 and DOI instances in Table 5

Reg. exp. (Panel no. in Fig. 2)	DOI instances (id in Table 5)	Reg. exp. (Panel no. in Fig. 2)	DOI instances (id in Table 5)
(a) ①	1 (2nd), 4, 5 (2nd)	(b) ④	23
(a) ②	2, 16 (1st)	(b) ⑤	23
(a) ③	3	(b) ⑥	24, 25 (2nd), 26
(a) ④	5 (1st), 19 (2nd)	(b) ⑦	27 (2nd), 28 (2nd)
(a) ⑤	16 (2nd), 17	(b) ⑧	29 (2nd)
(b) ①	8 (1st)	(b) ⑨	14, 15
(b) ②	9 (1st)	(c)	30 (2nd)
(b) ③	20 (1st), 21, 22		

Table 3 Distribution of various DOI errors in the *gene editing* dataset

Prefix-type errors	Suffix-type errors	Other-type errors	$\sum$
4992 (92.84%)	221 (4.11%)	164 (3.05%)	5377

(Line 33–36 in Algorithm 1), suffix-type (Line 37–44 in Algorithm 1) and other-type (Line 45–46 in Algorithm 1) errors of DOI names are cleaned sequentially with the regular expressions in Fig. 2. The prefix- and suffix-type errors are further grouped into several cases. For convenient understanding, Table 2 illustrates the mapping between regular expressions in Fig. 2 and DOI instances in Table 5.

In addition, the function *TrimDoi* (Line 45–46 in Algorithm 1) is also able to deal with several special cases, such as forward slash (id = 10 in Table 5), double underlines (id = 11 in Table 5), double dots (id = 12 in Table 5), XML tags (id = 13 in Table 5) and so on. In the end, if trimmed DOI names do not follow the specified characteristics (Sidman and Davidson 2001; Simmonds 1999) (Line 47 in Algorithm 1), trimmed DOI name and false status are returned. One can find the resulting instances in Table 5, e.g., id = 40–44 and 47. Otherwise, if trimmed DOI names end with hyphen or underline symbol (such as id = 32 (1st), 33 and 39 in Table 5), these DOIs are also illegal (Line 48–49 in Algorithm 1). Then, the function *IsBracketMatch* is used to check whether the involved brackets match in trimmed DOI names or resulting substrings excluding the last letter (Line 50–53 in Algorithm 1). Please refer to the DOI instances with id = 6 and 38 in Table 5 for more details. Since the functionality of *IsBracketMatch* is very simple and easy to implement, the corresponding pseudo-code is omitted in this study.

Table 4 Examples of various DOI errors in the WoS database

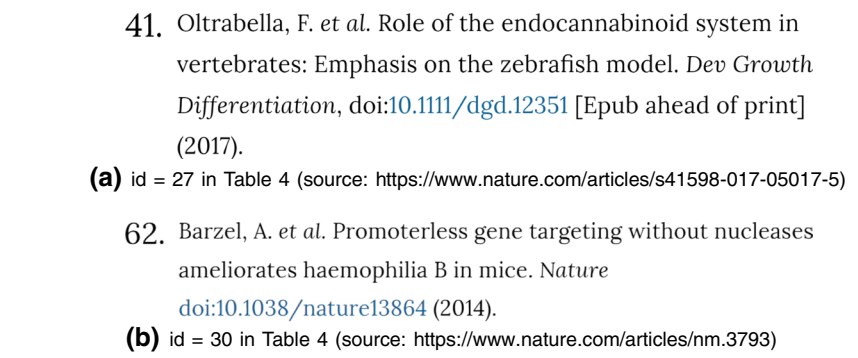
id	WoS id for citing article	Cited reference	DOI names after cleaning
1	WOS:000366706000001	Foldvari M., 2015, WILEY INTERDISCIP RE, DOI [10.1002/wnan.1361, DOI 10.1002/WNAN.1361.]	10.1002/WNAN.1361
2	WOS:000392772000006	Cheng Y., 1999, INT C INT SYST MOL B, V8, P93, DOI http://dx.doi.org/10.1007/11564126	10.1007/11564126
3	WOS:000358200100001	Breuninger S, J CLIN CELL IMMUNOL, DOI org/10.4172/2155-9899.1000264	10.4172/2155-9899.1000264
4	WOS:000416251400010	Husson SJ., 2012, WORMBOOK, DOI doi/10.1895/wormbook.1.156.1	10.1895/WORMBOOK.1.156.1
5	WOS:000402759900001	Egana M, 2009, CEUR WORKSHOP P, V496, DOI [:10.1038/npre.2009.4006.1, DOI 10.1038/NPRE.2009.4006.1]	10.1038/NPRE.2009.4006.1
6	WOS:000414888500006	DStrahl B. D., 2000, NATURE, V403, P41, DOI DOI 10.1038/47412)	10.1038/47412
7	WOS:000369840400002	Tas F, 2013, MOL CLIN ONCOL, V1, P788, DOI DOI 10.3892/MCO.2013.131]	10.3892/MCO.2013.131
8	WOS:000411518200001	Malvaez M, 2013, P NATL ACAD SCI USA, V110, P2647, DOI [10.1073/pnas.1213364110/-/DCSupplemental, 10.1073/pnas.1213364110]	10.1073/PNAS.1213364110
9	WOS:000416040200001	Zhang L, 2017, HEPATOLOGY, V65, P604, DOI [10.1002/hep.28882/suppinfo, 10.1002/hep.28882]	10.1002/HEP.28882
10	WOS:000389633400031	Srivastava K, 2012, PLOS ONE, V7, DOI [10.1371/journal.pone.0050966, 10.1371/journal.pone.\0050966]	10.1371/JOURNAL.PONE.0050966
11	WOS:000388946900001	Abouheif E, 2014, ADV EXP MED BIOL, V781, P107, DOI [10.1007/978-94-007-7347-9__6, 10.1007/978-94-007-7347-9_6]	10.1007/978-94-007-7347-9_6
12	WOS:000377106600003	Fux B, 2007, INFECT IMMUN, V75, P2580, DOI 10.1128/IAI..00085-07	10.1128/IAI.00085-07
13	WOS:000376667300001	Beker A. T., 2012, PLOS ONE, V7, DOI [10.1371/journal.pone.0048679, DOI 10.1371/JOURNAL.PONE.0048679, 10.1371/journal.pone.0048679<original-structure ref=info:doi/10.1371/journal.pone.0048679></original-structure>]	10.1371/JOURNAL.PONE.0048679
14	WOS:000281246700020	Khuri S., 1994, P 1994 ACM S APPL CO, P188, DOI DOI 10.1145/326619.326694.(HTTP://D0LACM.0RG/10.1145/326619.326694)	10.1145/326619.326694
15	WOS:000358198700087	Kocan KM, 2011, JOVE-J VIS EXP, V47, pe2474, DOI DOI 10.3791/2474.HTTP://WWW.J0VE.C0M/DETAILS.STP?	10.3791/2474
16	WOS:000369163300022	Seed KD, 2011, MBIO, V2, DOI [http://dx.doi.org/10.1128/mBio.00334-10, DOI 10.HTTP://DX.D0I.0RG]	10.1128/MBIO.00334-10
17	WOS:000365283800013	Qiu Y, 2014, CURR PLANT BIOL, V1, P6, DOI DOI 10.HTTP://DX.D0I.0RG/10.1016/J.CPB.2014.08.002	10.1016/J.CPB.2014.08.002
18	WOS:000312649800014	Bedell V. M., 2012, NATURE 0923, DOI [doi: 10.__1038/nature11537, DOI 10.1038/NATURE11537]	10.1038/NATURE11537
19	WOS:000313373700005	Eyzaguirre FC, 2009, REV MED CHILE, V137, P31, DOI [10.4067/S0034-98872009000100005, /S0034-98872009000100005]	10.4067/S0034-98872009000100005
20	WOS:000369199600009	Hall B, 2009, CURR PROTOC CELL BIO, V19, P1, DOI [10.1002/0471143030.cb1912s44PMID:__19731224, DOI 10.1002/0471143030.CB1912S44]	10.1002/0471143030.CB1912S44
21	WOS:000401419300022	Petzold G., 2016, NATURE, V532, P127, DOI DOI 10.1038/NATURE16979;PMID:26909574	10.1038/NATURE16979
22	WOS:000418494000022	Song J., 2016, NAT COMMUN, V7, P1, DOI DOI 10.1038/NC0MMS10548.PMID:26817820	10.1038/NCOMMS10548
23	WOS:000358707200001	Olofsson K, 2008, BIOTECHNOL BIOFUELS, V1, P7, DOI DOI 10.1186/1754-6834-1-7.PMCID:PMC2397418	10.1186/1754-6834-1-7
24	WOS:000399448400040	Ribas de Pouplana L, 2014, TRENDS BIOCHEM SCI, V39, P355, DOI DOI 10.1016/J.TIBS.2014.06.002.EPUB	10.1016/J.TIBS.2014.06.002
25	WOS:000323050700006	Hruscha A., 2013, J NEUROCHEM, DOI [10.1111/jnc.12198, DOI 10.1111/JNC.12198;EPUB]	10.1111/JNC.12198
26	WOS:000319248200012	Hassan A, 2012, BLOOD, DOI DOI 10.1182/BL00D-2011-12-396879.(EPUB	10.1182/BLOOD-2011-12-396879
27	WOS:000404970900020	Oltrabella F., 2017, DEV GROWTH DIFFER, DOI [10.1111/dgd.12351, DOI 10.1111/DGD.12351[EPUBAHEADOFPRINT](2017)]	10.1111/DGD.12351
28	WOS:000399344700015	Langie SA, 2016, BASIC CLIN PHARMACOL, DOI [10.1111/bcpt.12721, DOI 10.1111/BCPT.12721(2016)(EPUBAHEADOFPRINT]	10.1111/BCPT.12721
29	WOS:000418044000013	Kent WJ, 2002, GENOME RES, V12, P656, DOI [10.1101/gr.229202. Article published online before March 2002, 10.1101/gr.229202]	10.1101/GR.229202
30	WOS:000348974800014	Barzel A., 2014, NATURE, DOI [10.1038/nature13864, DOI 10.1038/NATURE13864(2014)]	10.1038/NATURE13864
31	WOS:000350839200006	Allers K., BLOOD, DOI [10.1182=blood-2010-09-309591, DOI 10.1182/BLOOD-2010-09-309591]	10.1182/BLOOD-2010-09-309591
32	WOS:000334853300003	Kasai T., 2001, LNCS, P181, DOI [DOI 10.1007/3-540-48194-X_, 10.1007/3-540-48194-X17]	10.1007/3-540-48194-X_17
33	WOS:000360277400001	Jensen NM, 2011, J BIOMED SCI, V18, DOI 10.1186/1423-0127-18-	10.1186/1423-0127-18-10
34	WOS:000288272900004	Wang JX, 2008, BIOCHEM CELL BIOL, V86, P157, DOI [10.1139/O08-008, 10.1139/O08-008]	10.1139/O08-008, 10.1139/008-008
35	WOS:000417079600003	Nishikawa T, 2014, MOL THER, V22, P2046, DOI [10.7038/mt.2014.128, 10.1038/mt.2014.128]	10.7038/MT.2014.128, 10.1038/MT.2014.128
36	WOS:000373787900001	Duan C., 2015, AM J PHYSIOL-CELL PH, V00315, DOI [10.1152/ajpcell.00315.2005, DOI 10.1152/AJCELL.00315.2015]	10.1152/AJCELL.00315.2005, 10.1152/AJCELL.00315.2015
37	WOS:000416225200012	Pai M, 2017, ELIFE, V6, DOI 10.7554/eLife.25956.001	10.7554/ELIFE.25956
38	WOS:000325073800018	Bikard D., 2013, NUCLEIC ACIDS RES, DOI DOI 10.1093/NAR/GKT520)(12	10.1093/NAR/GKT520
39	WOS:000357027000001	Zitnik M, 2014, BIOCOMPUT, V2013, P400, DOI DOI 10.1142/9789814583220_	10.1142/9789814583220_0038
40	WOS:000322292500007	Pereira IA, 2008, SWISS MED WKLY, V138, P534, DOI 2008/37/smw-12287	×
41	WOS:000418395600004	Frith J, 2008, TRANSFUS MED HEMOTH, V35, P216, DOI 000127448	10.1159/000127448
42	WOS:000369611900003	Sima H, 2010, IRAN J ALLERGY ASTHM, V9, P157, DOI 09.03/ijaai.157162	×
43	WOS:000377952700001	Bezuidt O, 2009, WORLD ACAD SCI ENG T, V58, P1169, DOI 10.1.1.193.5091	×
44	WOS:000184764800005	ROSA J, 1993, ENCYCL MED CHIR, DOI UNSP 13000S10	×
45	WOS:000320352400003	Flegal KM, 2005, JAMA-J AM MED ASSOC, V293, P1861, DOI [10.1001/jama.293.15.1861, 10.1001/jama.2009.2014]	10.1001/JAMA.293.15.1861, 10.1001/JAMA.2009.2014
46	WOS:000224082400021	Birklein F, 2000, ACTA NEUROL SCAND, V101, P262, DOI 10.1034/j.1600-0404.2000.101004262x./	10.1034/J.1600-0404.2000.09008.X
47	WOS:000376763300011	Hashimoto M., 2015, SCI REPORTS, V5, P5, DOI DOI 10.1038/	10.1038/SREP11315
48	WOS:000351183500182	FUJII W, 2013, NUCLEIC ACIDS RES, V41, P15514, DOI DOI 10.1093/NAR/	10.1093/NAR/GKT772
49	WOS:000415365700013	Bates Jane, 2017, Nurs Stand, V31, P31, DOI [10.7748/ns.31.27.31.s36, 10.7748/ns.31.35.31.s36, 10.7748/ns.31.21.31.s34, 10.7748/ns.31.36.31.s36, 10.7748/ns.31.22.31.s36, 10.7748/ns.31.50.31.s38, 10.7748/ns.31.51.31.s36, 10.7748/ns.31.44.31.s37, 10.7748/ns.31.25.31.s37, 10.7748/ns.31.19.31.s34, 10.7748/ns.31.24.31.s35, 10.7748/ns.31.23.31.s35, 10.7748/ns.31.41.31.s36, 10.7748/ns.31.52.31.s34, 10.7748/ns.31.34.31.s35]	10.7748/NS.31.27.31.S36, 10.7748/NS.31.35.31.S36, 10.7748/NS.31.21.31.S34, 10.7748/NS.31.36.31.S36, 10.7748/NS.31.22.31.S36, 10.7748/NS.31.50.31.S38, 10.7748/NS.31.51.31.S36, 10.7748/NS.31.44.31.S37, 10.7748/NS.31.25.31.S37, 10.7748/NS.31.19.31.S34, 10.7748/NS.31.24.31.S35, 10.7748/NS.31.23.31.S35, 10.7748/NS.31.41.31.S36, 10.7748/NS.31.52.31.S34, 10.7748/NS.31.34.31.S35

Table 4 (continued)

It is very possible that multiple citing articles cite simultaneously a same cited reference. But due to space limit, only one resulting WoS id is shown here

Table 5 The number of cited references with multiple DOI names in the *gene editing* dataset

No. of DOI names	2	3	4	5	8	15	∑
Before cleaning	9704	45	1	3	1	1	9755
After cleaning	1990	33	1	3	1	1	2029


41. Oltrabella, F. *et al.* Role of the endocannabinoid system in vertebrates: Emphasis on the zebrafish model. *Dev Growth Differentiation*, doi:10.1111/dgd.12351 [Epub ahead of print] (2017).

(a) id = 27 in Table 4 (source: https://www.nature.com/articles/s41598-017-05017-5)

62. Barzel, A. *et al.* Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. *Nature* doi:10.1038/nature13864 (2014).

(b) id = 30 in Table 4 (source: https://www.nature.com/articles/nm.3793)

Fig. 3 The snapshots of the cited references with id = 27 and 30 in Table 5

Results and discussions

Table 3 summarizes the distribution of various DOI errors in the *gene editing* dataset. From Table 3, one can see that the vast majority of DOI errors belong to the prefix-type error. In fact, the number of DOI errors with the prefix "DOI␣" is 4,968, which accounts for 92.39% DOI errors. Amongst the other errors, the number of the incoherently described DOI errors (not beginning with the prefix "10.") is 154, as reported in the supplementary material S1. To evaluate the performance of our cleaning method, the number of publications with multiple DOI names before and after cleaning is shown in Table 4. It is not difficult to see that the number of cited references with two and three DOI names is reduced drastically from 9,704 to 1,990 and from 45 to 33, respectively. This indicates that the quality of DOI names of cited references in the WoS database has been greatly improved. Please check the attached supplementary materials S1 and S2 for more details.

Algorithm 1 Cleaning the DOI names of the cited references in the WoS database

Precondition: $RECORD\_DELIMITER$ is the record delimiter, i.e., ";␣"
Precondition: $DOI\_DELIMITER$ denotes the DOI delimiter, i.e., ",␣"
Precondition: $regexPrefix$ is the regular expression for prefix DOI errors, i.e., Fig. 2(a)
Precondition: $regexSuffix$ is the regular expression for suffix DOI errors, i.e., Fig. 2(b)
Precondition: $regexYear$ is the regular expression for suffix year DOI errors, i.e., Fig. 2(c)

```
1: procedure Cleaning(String refList)
2:     String[] citedRefs ← refList.split(RECORD_DELIMITER)
3:     for i ← 0 to citedRefs.length() do
4:         int pos ← citedRefs[i].indexof(",␣DOI")
5:         if pos ≠ −1 then
6:             String doi ← citedRefs[i].substring(pos + ",␣DOI".length()).trim()
7:             if doi.startsWith("[") and doi.endsWith("]") then
8:                 doi ← doi.substring(1, doi.length() - 1)
9:                 doi ← JoinDoi(doi.split(DOI_DELIMITER))
10:                output i and doi
11:            else output i and TrimDoi(doi)
12:            end if
13:        end if
14:    end for
15: end procedure

16: function JoinDoi(String[] dois)
17:    Set<String> doiSet ← new LinkedHashSet<String>()
18:    for String doi: dois do
19:        tmp ← TrimDoi(doi)                ▷ tmp[0]: trimmed DOI name, tmp[1]: true or false
20:        if tmp[1] then doiSet.add(tmp[0])
21:        end if
22:    end for
23:    StringBuilder sb ← new StringBuilder()
24:    List<String> doiList ← new ArrayList<String>(doiSet)
25:    for int i ← 0 to doiList.size() - 1 do
26:        sb.append(doiList.get(i) + DOI_DELIMITER)
27:    end for
28:    sb.append(doiList.get(doiList.size() - 1)
29:    return sb.toString()
30: end function

31: function TrimDoi(String doi)
32:    String newDoi ← doi.toUpperCase().replaceAll("\\s+", "")
33:    Pattern patPrefix ← Pattern.compile(regexPrefix)
34:    Matcher matPrefix ← patPrefix.matcher(newDoi)
35:    if matPrefix.find() then newDdoi ← matPrefix.group(1)
36:    end if
37:    Pattern patSuffix ← Pattern.compile(regexSuffix)
38:    Matcher matSuffix ← patSuffix.matcher(newDoi)
39:    if matSuffix.find() then newDoi ← matSuffix.group(1)
40:    end if
41:    Pattern patYear ← Pattern.compile(regexYear)
42:    Matcher matYear ← patYear.matcher(newDoi)
43:    if matYear.find() then newDoi ← matYear.group(1)
44:    end if
45:    newDoi ← newDoi.replaceAll("\\\\", "").replaceAll("__", "_").replaceAll("\\.\\.", ".")
46:    newDoi ← newDoi.replaceAll("<.*?>.*?</.*?>", "").replaceAll("<.*?/>", "")
47:    if not newDoi.equals("") and newDoi.matches("10\\..*?/.*?") then
48:        if newDoi.matches(".*?[-|_]$") then return (newDoi, false)
49:        end if
50:        if IsBracketMatch(newDoi) then return (newDoi, true)
51:        else if IsBracketMatch(newDoi.substring(0, newDoi.length() - 1)) then
52:            return (newDoi.substring(0, newDoi.length() - 1), true)
53:        end if
54:    end if
55:    return (newDoi, false)
56: end function
```

It is worth noting that our cleaning method cannot conquer the following several situations: (a) if similar characters are confused with each other (Zhu et al. 2019), e.g. id = 7, 26 and 34 (1st) in Table 5, incorrect DOI names will be output, such as "10.3892/MC0.2013.131" (id = 7 in Table 5), "10.1038/NC0MMS10548" (id = 22 in Table 5), "10.1182/BL00D-2011-12-396879" (id = 26 in Table 5) and "10.1139/008-008" (the 2nd one with id = 34 in Table 5); (b) if multiple DOI names are incorrectly assigned to the same cited reference, our cleaning method cannot currently differentiate which one is correct, e.g. id = 34–36, 49 in Table 5 (DOI names with blue color are correct); (c) A DOI name assigned to the scientific publication cannot be resolved by the DOI system (http://dx.doi.org/), e.g. "10.1007/3-540-48194-X17" (id = 32 in Table 5) and "10.1093/NAR" (id = 48 in Table 5); (d) A DOI name assigned to the scholarly article is resolvable, but it is resolved to some knowledge unit by the DOI system. For instance, the DOI name "10.7554/ELIFE.25956.001" is not resolved to the cited reference *per se* by DOI system, but to its *abstract* (id = 37 in Table 5).

As a matter of fact, after our preliminary analysis, *unsupervised anomaly detection* algorithm (Goldstein and Uchida 2016) can be utilized to deal with most above cases. Let's take the prefix "10.7554/ELIFE." (id = 37 in Table 5) as an example. In our dataset, all DOI names with this prefix ends with five digital letters, except for the following three cited references: "10.7554/ELIFE.08716.001", "10.7554/ELIFE.11553.001" and "10.7554/ELIFE.25956.001". Of course, sometimes the truth is in the hands of a few people. For example, among DOI names with the prefix "10.3892/MC" in our dataset (id = 7 in Table 5), only one is found to be correct ("10.3892/MCO.2013.119"). Therefore, one should determine interactively whether or not the detected *abnormal* DOI names should be remained. Experimental results and insights from unsupervised anomaly detection algorithm will be described in another paper. However, as for the case with id = 49 in Table 5, any automatic approach seems be helpless, since all DOI names are resolvable and these publications are written by the same author (*Jane Bates*) and published in the same journal (*Nursing Standard*) with the same volume number (*31*) and page number (*31–31*),but with different issue number.

Conclusions

As noted by Zhu et al. (2019), there is no simple way to recognize and thus to evaluate the extent of DOI errors in the Web of Science database. After careful analysis on the bibliographic data in the *gene editing* field, several classic DOI errors of cited references, such as prefix-, suffix- and other-type errors, are identified. The other-type errors can be further divided into three subgroups: (a) those containing special characters (such as id = 10–14, 46 in Table 5), (b) incoherently described DOIs (such as id = 40–44, 47 in Table 5), and (c) those with incomplete suffix but with correct DOI prefix (i.e., "10.") (such as id = 32 (1st), 33 and 39 in Table 5) . Then, a cleaning method of DOI names is put forward on the basis of regular expressions in this work.

Though our cleaning approach can improve largely the quality of DOI names of cited references in the WoS database, several situations cannot still be conquered by our approach: (a) similar characters are confused with each other (Zhu et al. 2019), such as "O" versus "0", "b" versus "6" and "O" versus "Q"; (b) it is very difficult to distinguish the correct one from multiple DOI names assigned to the same cited reference; (c) A DOI

name assigned to some cited reference cannot be resolved by the DOI system; (d) A DOI name is resolvable, but points to some knowledge unit within the interested cited reference. According to our preliminary analysis, it seems that *unsupervised anomaly detection* algorithm (Goldstein and Uchida 2016) is able to deal with most above cases. In the near future, we will try this algorithm and report our insights in another paper.

In the meanwhile, this work argues that similar DOI errors should also exist for other bibliographic database. For example, wrong DOI names for id = 40–44 in Table 5 seems come from *MEDLINE* database, since the corresponding publications in these two database are assigned with the same incorrect DOI names. Therefore, it is not without reasons that the cleaning method proposed in this study should be applicable to other databases. Another interesting phenomenon, shown in Fig. 3, can be observed. The correct DOI names can be obtained from the detailed webpages of the resulting articles, though some noise information is followed. But, the correct DOI names and noise information are mixed in the WoS database.

From Table 5, in our opinion, it is very complex to figure out possible sources of errors due to the diversity of DOI errors. However, since high-quality bibliometric data is the stakeholders' ultimate goal, one feasible solution is to clean all DOI names of the cited references in the interested databases with our approach and then *unsupervised anomaly detection* algorithm (Goldstein and Uchida 2016). After these processing steps, for those publications with still more than one different DOI names or wrong DOI names, the WoS and other databases should recognize and keep the correct DOI names.

Supplementary material

S1 The cited references with the incoherently described DOI errors http://54xushuo.net/wiki/lib/exe/fetch.php?media=resources:papers:s1.xlsx

S2 The cited references with multiple DOI names before cleaning http://54xushuo.net/wiki/lib/exe/fetch.php?media=resources:papers:s2.xlsx

S3 The cited references with multiple DOI names after cleaning http://54xushuo.net/wiki/lib/exe/fetch.php?media=resources:papers:s3.xlsx

Acknowledgements Our gratitude goes to the anonymous reviewers and the editor for their valuable comments.

References

Boundry, C., & Chartron, G. (2017). Availability of digital object identifiers in publications archived by PubMed. *Scientometrics*, *110*(3), 1453–1469. https://doi.org/10.1007/s11192-016-2225-6.

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College and Research Libraries*, *67*(4), 292–303. https://doi.org/10.5860/crl.67.4.292.

Chandrakar, R. (2006). Digital object identifier system: An overview. *The Electronic Library*, *24*(4), 445–452. https://doi.org/10.1108/02640470610689151.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics. *Journal of the Association for Information Science and Technology*, *64*(10), 2149–2156. https://doi.org/10.1002/asi.22898.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics*, *8*(3), 751–765. https://doi.org/10.1016/j.joi.2014.07.003.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in indexing bybibliometric databases. *Scientometrics*, *102*(3), 2181–2186. https://doi.org/10.1007/s11192-014-1503-4.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. *Journal of Informetrics*, *10*(1), 174–182. https://doi.org/10.1016/j.joi.2015.11.006.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173.

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008.

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE*, *10*(5), e0127830. https://doi.org/10.1371/journal.pone.0120495.

Huang, M., & Liu, W. (2019). Substantial numbers of easily identifiable illegal DOIs still exist in Scopus. *Journal of Informetrics*,. https://doi.org/10.1016/j.joi.2019.03.019.

Jacso, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, *30*(3), 297–309. https://doi.org/10.1108/14684520610675816.

Jobmann, A., Hoffmann, C. P., Künne, S., Peters, I., Schmitz, J., & Wollnik-Korn, G. (2014). Altmetrics for large, multidisciplinary research groups: Comparison of current tools. *Bibliometrie-Praxis und Forschung*, *3*(1), 1–19. https://doi.org/10.5283/bpf.205.

Krauskopf, E. (2019). Missing documents in Scopus: The case of the journal enfermeria nefrologica. *Scientometrics*, *119*(1), 543–547. https://doi.org/10.1007/s11192-019-03040-z.

Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science-an explorative study. *Journal of Informetrics*, *12*(3), 985–997. https://doi.org/10.1016/j.joi.2018.07.008.

Neumann, J., & Brase, J. (2014). DataCite and names for research data. *Journal of Computer-Aided Molecular Design*, *28*(10), 1035–1041. https://doi.org/10.1007/s10822-014-9776-5.

Paskin, N. (1999). The digital object identifier system: Digital technology meets content management. *Interlending & Document Supply*, *27*(1), 13–16. https://doi.org/10.1108/02641619910255829.

Paskin, N. (2010). Digital object identifier (DOI) system. In A. Kent (Ed.), *Encyclopedia of library and information sciences* (3rd ed., pp. 1586–1592). Milton Park: Taylor and Francis.

Sidman, D., & Davidson, T. (2001). A practical guide to automating the digital supply chain with the digital object identifier (DOI). *Publishing Research Quarterly*, *17*(2), 9–23. https://doi.org/10.1007/s12109-001-0019-y.

Simmonds, A. W. (1999). The digital object identifier (DOI). *Publishing Research Quarterly*, *15*(2), 10–13. https://doi.org/10.1007/s12109-999-0022-2.

Tang, L., Hu, G., & Liu, W. (2017). Funding acknowledgement analysis: Queries and caveats. *Journal of the Association for Information Science and Technology*, *68*(3), 790–794. https://doi.org/10.1002/asi.23713.

Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre- Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, *9*(3), 570–576. https://doi.org/10.1016/j.joi.2015.05.002.

Wang, J. (2007). Digital object identifiers and their use in libraries. *Serials Review*, *33*(3), 161–164. https://doi.org/10.1016/j.serrev.2007.05.006.

Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics*, *117*(1), 61–84. https://doi.org/10.1007/s11192-018-2841-4.

Zhu, J., Hu, G., & Liu, W. (2019). DOI errors and possible solutions for Web of Science. *Scientometrics*, *118*(2), 709–718. https://doi.org/10.1007/s11192-018-2980-7.

Zhu, J., Liu, F., & Liu, W. (2019). The secrets behind Web of Science's search. *Scientometrics*, *4*, 1745–1753. https://doi.org/10.1007/s11192-019-03091-2.