Analysis of Crossref Reports in Order to Improve the Quality of Metadata of Scientific Publications

Aleksey Ermakov

Keldysh Institute of Applied Mathematics. Moscow, Russia

Abstract

Issues related to improving the quality of metadata of scientific publications placed in the Crossref bibliographic database are considered. All information contained in metadata obtained from publishers of scientific publications is analyzed by Crossref and displayed in various reports. Analysis of these reports gives publishers an idea of the completeness and correctness of the bibliographic data presented. The quality of metadata directly or indirectly affects the number of views and links to a publication, respectively, on the ratings of scientific publications, authors and organizations.

Keywords

metadata of publications, Crossref reports, citations, ratings of scientific publications

1. Introduction

The preparation of a scientific publication is inextricably linked with the topic and direction of research in which the authors work. As a rule, publications are a continuation of the research team's work in this direction and are based on previous research results carried out earlier either by the authors themselves, or by their scientific advisors or colleagues. This reliance in publications on prior scientific research results is accompanied by citation of relevant scientific materials. The correctness of citation is very important not only for the authors of a new scientific article, but also for the authors of the cited articles, as well as for their scientific publications, since it directly or indirectly affects the number of views and links to a publication, respectively, on ratings scientific publications, authors and organizations.

Publication metadata is needed to make bibliographic data about the article itself, authors, etc., identifiable, public and accessible to others.

All published scientific materials are assigned a digital identifier DOI (Digital Object Identifier), which provides a link (URL) to the permanent location of an object or information about it (metadata) on the Internet.

The overwhelming majority of authors have ORCID – this is a unique code (ID), which the author of scientific works receives to identify the results obtained and written works. The main task of ORCID is to unambiguously identify the researcher in all bibliographic databases.

Association Crossref [1], a member of which since 2016 is the Keldysh Institute of Applied Mathematics (KIAM), maintains a worldwide collaborative cross-citation service that acts as a gateway between publishers' electronic platforms. This service does not store full texts of scientific publications, but enters into its database information about the relationship of publications using DOI technology [2], as well as metadata of published scientific materials.

The tools under development by Crossref (and some other organizations such as Google Scholar, Scopus and Web of Science, which use different sources for their citation data) make it easier for both the author of the publication and readers to find, cite, evaluate, reuse re – the results of scientific research. In addition, the statistics of "reading" (transitions to publication), as well as back citations [3]

(links to publications referring to this article) in one way or another affect the rating of researchers and the Institute as a whole. Therefore, IPM, both as a source of scientific materials and as a research center, is important to monitor the correctness of the metadata of published scientific publications.

To check the quality of metadata, Crossref has developed a fairly large set of tools that help publishers evaluate and improve their metadata.

This paper reviews reports that any publisher or interested author can access on the Crossref website. Analysis of the information presented in these reports allows us to qualitatively assess the completeness of the content being uploaded related to published scientific materials and outline ways to improve the infrastructure of scientific publications and tools for interaction with Crossref.

2. List of publications with available for viewing metadata

Let's talk in detail about the Crossref reports using the example of some scientific publications of KIAM and Kazan Federal University (KFU). Using the KIAM publications as an example, we will show how a publisher can use Crossref reports to find errors and / or inaccuracies in publication metadata and make the necessary corrections. Using the KFU publications as an example, it will be shown how an interested author can discover inaccuracies or incompleteness of the metadata of the publication of interest. Further, the author can contact the publisher, pointing out the inaccuracies, and ask them to correct them.

Any scientific publishing house can study the Crossref reports in the same way, focusing on their scientific publications.

All organizations cooperating with Crossref are given the opportunity not only to see information about their publications, but also to make the necessary adjustments.

The information is presented in the form of three lists, sorted by alphabet:

• Magazines

https://www.crossref.org/06members/51depositor.html

• Conferences proceedings

https://www.crossref.org/06members/51depositorCP.html

• Books

https://www.crossref.org/06members/51depositorB.html

In Fig. 1 shows a fragment of the alphabetical list of world publishers of journals, among which the Keldysh Institute of Applied Mathematics (Keldysh Institute of Applied Mathematics) is represented by three journals:

• Mathematica Montisnigri (Mathematics of Montenegro);

• Mathematicheskoe modelirovanie (Mathematical modeling);

• Keldysh Institute Preprints.

In Fig. 2 shows a fragment of the alphabetical list of world publishers of conference proceedings, among which the Keldysh Institute of Applied Mathematics is presented by the materials of two conferences [4]:

• Futurity designing. Digital reality problems – 3 issues – 2018, 2019, 2020.

• Scientific Services & Internet – 5 issues – 2016 and 2017, and then combined into a series of 2018, 2019 and 2020.

In Fig. 3 shows a fragment of the alphabetical list of world magazine publishers, among which Kazan Federal University is represented by 11 magazines. Further, in section 6, we will dwell on one of them in more detail – the Russian Digital Libraries Journal [5].

In various Crossref tools (when loading content, defining a link, etc.), depending on the requirements of simplicity-complexity and efficiency-accuracy, one or another algorithm is used. But in any case, the author or employee of the publishing house is advised to independently check the correctness of the recognition result of bibliographic records and the accuracy of the reference obtained.

Figure 1: KIAM in the list of magazine publishers

Figure 2: KIAM in the list of conference proceedings publishers

Figure 3: KFU in the list of magazine publishers

3. Depositor's report

Perhaps to emphasize the fact that the publisher, by uploading metadata to Crossref, receives a certain benefit and some free services, the authors of the project began to use certain banking terms: the placement of content was called deposit, the publisher itself was called a contributor, etc.

Contributor reports for each publisher are used to verify basic information about DOI registrations. The reports are "tied" to three key lists of publishers supported by Crossref– the list of magazine publishers, the list of conference proceedings publishers, and the list of monograph (book) publishers (see Section 1). There are currently no contributor reports for other types of content, such as photos, pictures, videos, and audio files.

The index page is updated weekly. Header-level reports are updated as the metadata is updated. It is possible to obtain and analyze a report on all IPM publications (journals, conference proceedings, monographs), but we will consider one of the journals as an example.

Selecting the publisher Keldysh Institute of Applied Mathematics in the list of journals (see Fig. 1), and then selecting, for example, Mathematicheskoe modelirovanie, we will receive a detailed report, where for each DOI the owner prefix, timestamp, date of the last update of the record and the number of citations are indicated (according to Crossref data) of this publication (see Fig. 4).

Figure 4: Fragment of the Contributor's Report for the journal "Mathematicheskoe modelirovanie"

4. Conflict report

As you know, DOI is a unique identifier, so there should always be only one DOI for each content item. And the publisher will receive a conflict report if it has at least one DOI conflict.

It is important to fix these conflicts as soon as possible because they can lead to problems in the future. The presence of two different DOIs for the same content means that the researcher will not know which one to cite, thereby risking distorting the number of citations. In addition, the publisher may forget that he has two DOIs for one object, and update only one of them when the content changes. This means that anyone who uses a different DOI that has not been updated will click on the broken link. Therefore, bad metadata should be quickly eliminated and thus the problem should be solved.

A conflict report shows where two (or more) DOIs were submitted with the same metadata, or indicates that the publisher of a scientific publication might have duplicate DOIs when submitting metadata to Crossref.

All DOI conflicts associated with journal articles, conference proceedings or monographs are reported in a single Conflict Report on the Crossref website (Figure 5). If we have active conflicts, we will receive a reminder by email every month.

Figure 5: The journal "Mathematics Montisnigri" in the list of DOI conflicts

The DOI Conflict Report shown in Fig. 5, it is said about the presence of 15 conflict situations that arose when the metadata of the journal "Mathematics Montisnigri" was placed in Crossref. The first line in this report with an unspecified journal name and zero conflicts is a small glitch of the Crossref programmers, which we informed them about and with which they fully agreed.

And we need to understand the cause of the conflict and try to correct it. By clicking on the name of the journal, we get a full report on all 15 conflict situations. An example of one of them:

Created: 2020-03-19 14:23:13.0
ConfID: 5583368
CauseID: 1465359938
OtherID: 1462438643
JT: Mathematica Montisnigri
MD: Jokanović, 46 ,null,5,2019,A breaf survey on Armendariz and central Armendariz rings
DOI: 10.20948/mathmon-2019-46-1(Journal) (5583368-N)
DOI: 10.20948/mathmontis-2019-46-1(Journal)

As it turned out, the attempt made by the editorial staff of the journal to switch to new DOI identifiers simultaneously with the use of old identifiers is unacceptable from the point of view of DOI ideology and Crossref technology.

As you can see from the presented example, the first article of the 46th volume was uploaded twice with different identifiers (DOIs). Although for each content element there should always be only one DOI, since the presence of two different DOIs for the same content can mislead readers getting acquainted with the materials of this publication. Publishers themselves can get confused when changing metadata and re-uploading content.

Crossref offers 3 scenarios for correcting DOI conflicts:

Scenario 1: If you assigned two DOIs to different content items, but accidentally sent the same metadata for both of them. In this case, one of the DOIs has incorrect metadata. If you re-upload the corrected metadata of this DOI, the conflict will be resolved.

Scenario 2. If you have assigned two DOIs to the same content item. In this case, you can resolve the conflict by assigning one of the DOIs as the primary and the other as an alias. The DOI alias will be automatically redirected to the main DOI, so it will be enough to support only the main one.

Scenario 3: If two DOIs refer to different content items, but their metadata is so similar that a conflict has been noted. This happens when very little metadata is included in the items. It is best to register additional metadata to resolve the conflict. Alternatively, you can accept the conflict by removing the conflict status and setting it to resolved. This will not affect the metadata or DOI records, but will remove the conflicts from the conflict report.

For us, the simplest solution was to remove the "invalid" DOI. But you cannot delete DOI - this is the fundamental principle of DOI. Therefore, it was decided to follow the second path proposed by Crossref – to call the "wrong" DOI an alias for the "correct" one using the DOI administration subsystem (Fig. 6).

DOI	Submission	Status	Date	Type	Journal	Version	Year	Volume	Issue	Suppl.	Page	Author	Title	Sequence
10.20948/mathmon-2019-46-1	1462438643	Alias	28-Feb-21 06:25	FULL_TEXT	Mathematica Montisnigri	201912271540	2019	46			5	Jokanovi?	A breaf survey on Armendariz and cent...	
10.20948/mathmontis-2019-46-1	1465359938	Primary	28-Feb-21 06:25	FULL_TEXT	Mathematica Montisnigri	201912271530	2019	46			5	Jokanovi?	A breaf survey on Armendariz and cent...	

Figure 6: Change DOI status to resolve conflict

After changing the status of the "incorrect" DOI, any call to it will cause an automatic transition to the link specified in the primary ("correct") DOI.

Having thus updated the content of all "incorrect" DOIs and changed their status to "pseudonym", we excluded IPM publications from the list of DOI conflicts.

5. Fields or missing metadata report

The report on fields or missing metadata provides detailed information on the completeness of the metadata. Like the Contributor's Report, it is "tied" to one of the three key lists of publishers supported by Crossref – the list of magazine publishers, the list of conference proceedings publishers and the list of monographs (books) publishers (see Section 1).

This report can be accessed by selecting the icon (green right arrow) next to the scientific publisher's name in one of the above lists.

It should be noted that the sets of metadata transmitted for journal articles, scientific publications in conference collections and monographs are somewhat different. For any scientific publication, its title in Russian and English, a list of authors (for each author, the first name and surname in Russian and English, as well as ORCID), the year of publication, the number or range of pages are transmitted. For journals and collections, the volume number and / or issue number are added. In addition, over the past 2 years, we began to upload abstracts of scientific publications and bibliography to Crossref, which is very important for tracking mutual citation [3].

Consider the information on our journals presented in the reports on fields or missing metadata (Fig. 7).

Publication Title	Ignore Fields	V	I	P	A	S	V/I	V/I/P	V/I/P/A	P/A	T	N	F	missing iParadigms URLs	Total DOIs*
Keldysh Institute Preprints	none CHANGE	878	0	0	0	275	0	0	0	0	0	0	0	na	878
Matematicheskoe modelirovanie	none CHANGE	25	0	0	0	29	0	0	0	0	0	0	25	na	162
Mathematica Montisnigri	none CHANGE	0	89	0	2	21	0	0	0	0	0	2	1	na	89

Figure 7: Report on fields or missing metadata in KIAM journals

The table column headings in Fig. 7 (fields are marked in red; from the point of view of Crossref, they are incomplete or incorrectly filled in):
• Participation Title – title of the publication;
• Ignore Fields – an indication, the absence of which fields should be ignored;
• V = volume – volume number;
• I = issue – issue number;
• P = page – number or range of pages;
• A = author – author;
• S = single-author – single author;
• T = article title– the title of the article;
• N = no-first-name – author's name is not specified;
• F = first name initial only – only the initials of the author are specified as the name;
• U = missing iParadigms Url–Crossref has partnered (since 2008) with iParadigms LLC to offer its members – leading scientific and professional publishers – the ability to check originality using CrossCheck and iThenticate services. The CrossCheck database includes full-text journals from leading academic publishers and is growing rapidly as publishers participating in Crossref subscribe to the service.

As we can see, using the Field Report or Missing Metadata, Crossref emphasizes that although some bibliographic metadata is optional for content registration purposes, it is strongly recommended that all publishers register as complete metadata as possible for each registered item. And even highlights in red in the report fields for which, from the point of view of Crossref, the publisher grossly does not follow the specified recommendations.

Analyzing the information presented in Fig. 7, it can be noted that for all issues of the KIAM Preprints volume numbers are not assigned, for all volumes of the journal "Mathematics Montisnigri" issue numbers are not assigned. For the journal "Mathematical Modeling", the comments concern not all issues, but only several volumes of the initial period of assigning DOIs to scientific publications, when our journal was assigned DOIs by some external organizations, even incorrectly registered in Crossref.

It should be noted that we assigned DOIs to all articles of the "Mathematics Montisnigri" magazine. And, accordingly, we can make the necessary corrections and / or additions to the metadata sets loaded into Crossref. The situation with the journal "Mathematical Modeling" is more complicated – until 2020, as a result of competitions, the publication passed from hand to hand and received completely different DOIs. At the same time, metadata of scientific articles of the journal were shipped, completely different in completeness and correctness. And now KIAM, as a journal publisher, has neither the right nor the opportunity to correct or supplement anything in the information on the 2016–2019 issues.

Analyzing further Crossref's comments on registering publication metadata, we see marks in the column for assigning a single author (S = single-author) and some others.

Similar remarks are available on the metadata of conference proceedings and monographs. We are carefully examining these (and other) Crossref comments and, if possible, try to eliminate them by making the necessary edits to the metadata description.

However, the publisher of a particular magazine can refuse such obsession with Crossref by using the Change switch (second column, Fig. 7 – Ignore Fields), specifying which fields should be ignored (Fig. 8).

Figure 8: An indication of which fields are missing to ignore

6. DOI Scanner Report

The DOI Scanner report is executed only for magazines and, accordingly, is "tied" to the list of magazine publishers.

The DOI Scanner samples articles for each journal from a particular publisher to ensure that the DOIs you specify translate to the correct page. For each viewed journal, DOIs are selected, the number of which is approximately 5% of the total DOIs for the journal (up to a maximum of 50 DOIs). In Fig. 9 lists all the journals published by the Institute of Applied Mathematics, and for each the total number of DOIs and the date of the last scan are indicated.

Figure 9: Date of the last scan of the KIAM journals

You can get access to the details of the crawler's work for a given journal by selecting a date in the "Last Crawl date" column.

Figure 10: Date of the last scan of the KIAM journals

No errors, as follows from the scanner report in Fig. 10, there are no KIAM Preprints. At the same time, 10 DOIs were scanned (and validated). These fields are "clickable", so you can see the results of the scanning procedure in more detail (Figs. 11, 12).
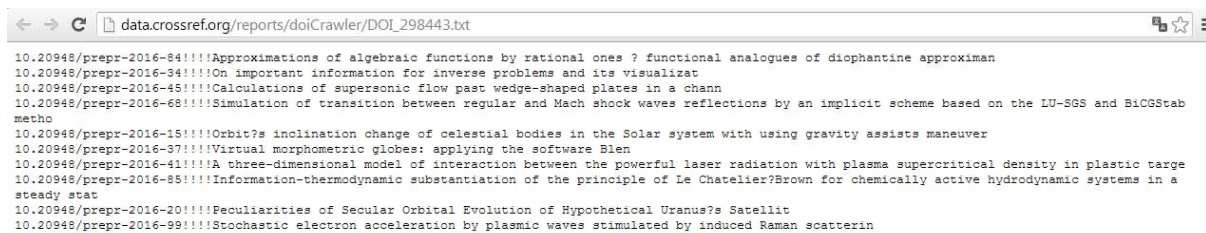
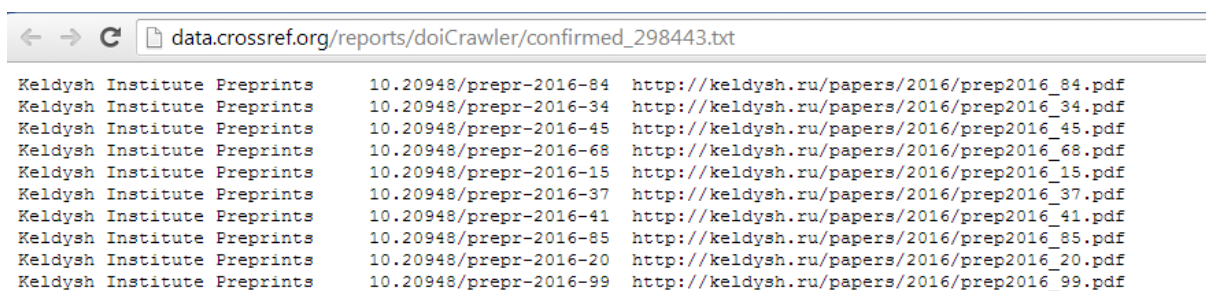**Figure 11**: List of selected DOIs for Crawling

Figure 12: Checked and confirmed DOIs

This function (DOI scanning), in our opinion, is quite interesting and useful, especially for those publications that have changed the database server or administration, since such changes could introduce serious errors in the procedure for depositing content and make publication materials inaccessible.

However, the implementation of this function by Crossref programmers seems a little strange and raises several questions:

• Why, for example, for KIAM Preprints, numbering 877 issues since the beginning of the DOI assignment, only 10 elements are checked, although the description of the algorithm says about 5% of the total number?

• Why is the Re-Crawl using the same sample as a few years ago in the previous scan? After all, the publisher is interested in tracking the correctness not only of old materials, but also of all publications of the studied journal.

So far, we have not received an answer to these questions from Crossref Support, but we continue to actively interact with it.

Another important addition is related to the fact that Crossref is one of the DOI recorders. And since DOI registrars do not exchange metadata, publishers associated with other registration services will not be able to use the Crossref tools.

7. Participation report

The authors of the project and the developers of Crossref urge publishers not only to post metadata of scientific publications, but to make them as complete as possible. In addition, Crossref encourages the scientific and publishing community to actively use the services offered, designed to analyze the completeness and correctness of the downloaded information, thereby, as it were, participating in the development and expansion of the range of these services.

For each publisher working with Crossref, there is a separate Participation report that shows what percentage of their contributing data is recorded for each of the ten key metadata elements. Participation reports show where there are gaps and what can be improved in terms of completeness of metadata.
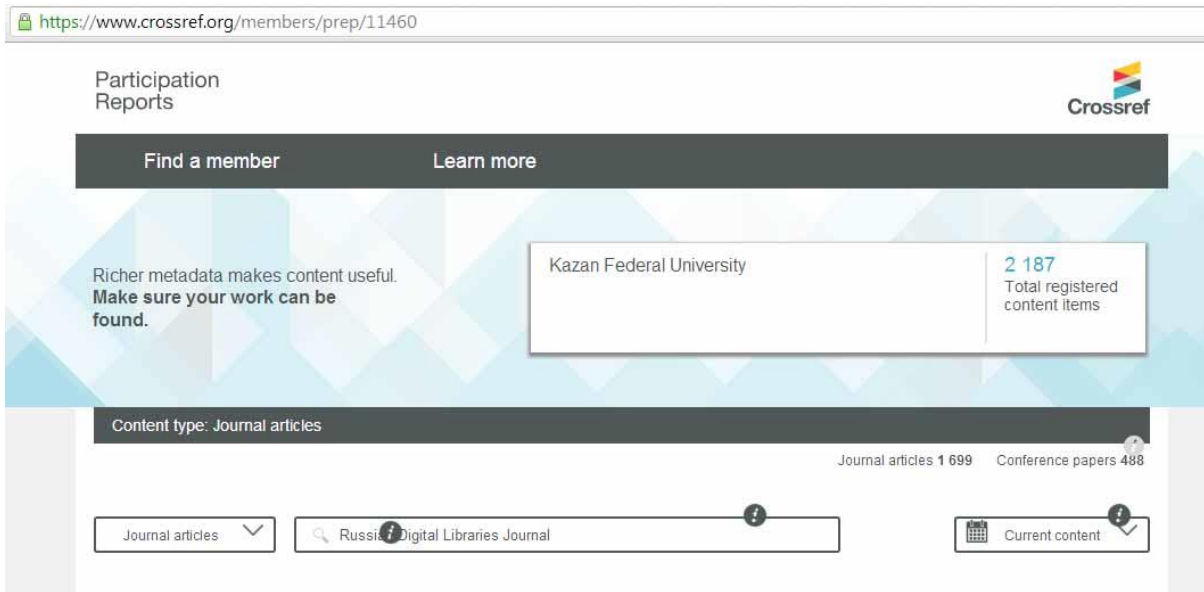
Figure 13: Participation report for scientific publications KFU

In Fig. 13 shows the title of the Report on the participation of the publisher of Kazan Federal University – the total number of content elements is 2187, including 1699 journal articles, 488 articles in conference proceedings. At the bottom of the title (Fig. 13) there are 2 menus – the choice of the type of scientific publication (left) and the choice of the analyzed period (all the time of deposit, the current period – the last 2 years, "old" materials – data downloaded more than 2 years ago). The central field of the title of the Participation Report allows you to enter the name of the journal, collection or even the title of the publication and analyze the completeness of the correspondingly loaded metadata.
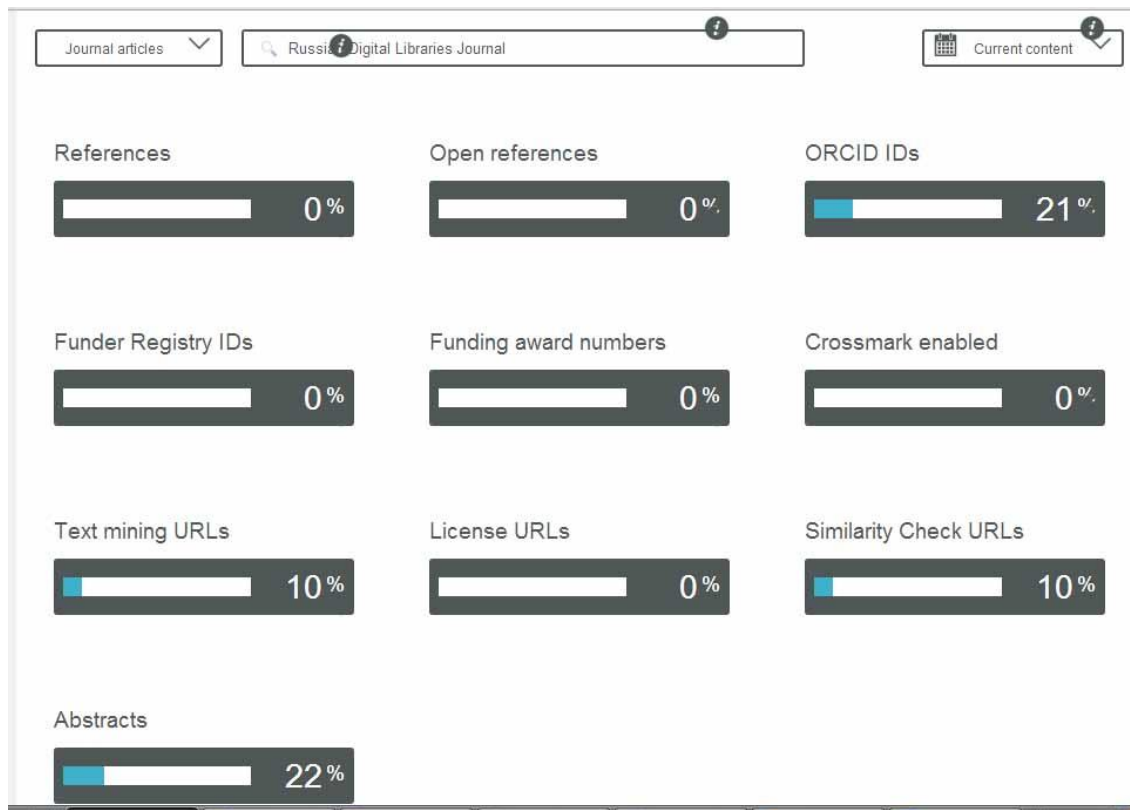

Figure 14: Percentages of the Russian Digital Libraries Journal Participation Report

The main part of the Participation Report for Russian Digital Library Journal is shown in Fig. 14 (the current time is set – the last 2 years):

• for 0% of publications, a list of references (References) is loaded;

• 0% of links are open (Open References) – available to all users of all Crossref services (there are no available links, since there are no lists of literature);

• ORCID is indicated for 21% of authors;

• for 0% of publications, the name and identifier (Funder Registry IDs) of the sponsor are indicated - at least one of the organizations that funded the study;

• for 0% of publications, the Funding award numbers are indicated;

• the percentage of content (in this case 0%) using the Crossmark service (Crossmark-enabled), which gives readers quick and easy access to the current status of a content item (as part of the publisher's policy for revisions, rebuttals, revocations and other updates);

• The percentage of registered content (in this case 10%) containing Text-mining URLs for a scientific publication – the automatic analysis and extraction of information from a large number of documents. At the moment, most scientific organizations in the world (and KFU, as it seems to us, including) are not interested in setting a special set of instructions, with the help of which someone for some reason will investigate their scientific materials;

• Percentage of metadata publications (in this case 0%) containing URLs that indicate the license (License URLs), which defines the conditions under which readers can access the content;

• percentage of publication metadata (in this case 10%), which includes URLs for checking similarity (Similarity Check URLs), for publications that cooperate with CrossCheck and iThenticate;

• 22% of publication metadata includes Abstracts, which gives a deeper understanding of the content of the work.

In our opinion, one should not chase 100% indicators, but at the same time it should be clear that a more complete and accurate filling of publication metadata in one way or another [6] affects the ratings of publications, authors and organizations. And the indication of grants and funds to support scientific activities has a positive effect on the relationship with these funds.

8. References

[1] Association Crossref. URL: https://www.crossref.org/about/.

[2] International DOI Foundation (IDF). URL: https://www.doi.org/.

[3] A. V. Ermakov, Bibliograficheskaya ssylka kak instrument avtora I chitatelya // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferencii (21–25 sentyabrya 2020 g., onlajn). M.: IPM im. M.V. Keldysha, 2020. S. 268–275. https://doi.org/10.20948/abrau-2020-55

[4] M. I. Slepenkov, Materialy konferencij v onlajnovoj biblioteke IPM im. M.V. Keldysha. Preprinty IPM im. M.V. Keldysha. 2020. № 18. 16 s. https://doi.org/10.20948/prepr-2020-18

[5] Russian Digital Libraries Journal. URL: https://elbib.ru.

[6] T. A. Polilova, Infrastruktura nauchnyh publikacij // Preprinty IPM im. M.V. Keldysha. 2009. № 15. 30 s. URL: https://library.keldysh.ru/preprint.asp?id=2009-15