Reflections on the Misuses of ORCID iDs

Miriam Baglioni[0000−0002−2273−9004], Andrea Mannocci[0000−0002−5193−7851], Paolo Manghi[0000−0001−7291−3210], Claudio Atzori[0000−0001−9613−6639], Alessia Bardi[0000−0002−1112−1292], and Sandro La Bruzzo[0000−0003−2855−1245]

ISTI-CNR, Pisa, Italy
{name.surname}@isti.cnr.it

Abstract. Since 2012, the "Open Researcher and Contributor Identification Initiative" (ORCID) has been successfully running a worldwide registry, with the aim of unequivocally pinpoint researchers and the body of knowledge they contributed to. In practice, ORCID clients, e.g., publishers, repositories, and CRIS systems, make sure their metadata can refer to iDs in the ORCID registry to associate authors and their work unambiguously. However, the ORCID infrastructure still suffers from several "service misuses", which put at risk its very mission and should be therefore identified and tackled. In this paper, we classify and qualitatively document such misuses, occurring from both users (researchers and organisations) of the ORCID registry and the ORCID clients. We conclude providing an outlook and a few recommendations aiming at improving the exploitation of the ORCID infrastructure.

Keywords: ORCID · Scholarly Communication · Open Science · Academia

1 Introduction

A precise and reliable identification of researchers and the pool of works they contributed to would greatly benefit scholarly communication practices and facilitate the understanding of science [4]. Several studies showed what can be achieved by pursuing researchers' productivity and affiliations: from understanding career trajectories and citation dynamics to analysing collaboration networks and migration pathways in academia [14, 15, 2].

Since 2012, ORCID [3], the "Open Researcher and Contributor Identification Initiative", has been running a worldwide registry[1], which mints alphanumeric iDs on behalf of registrant researchers, and maintains a core set of relevant information such as name, surname, affiliations, works, projects and so on in their so-called "ORCID profiles". ORCID's intended architecture figures the ORCID

[1] ORCID, https://orcid.org

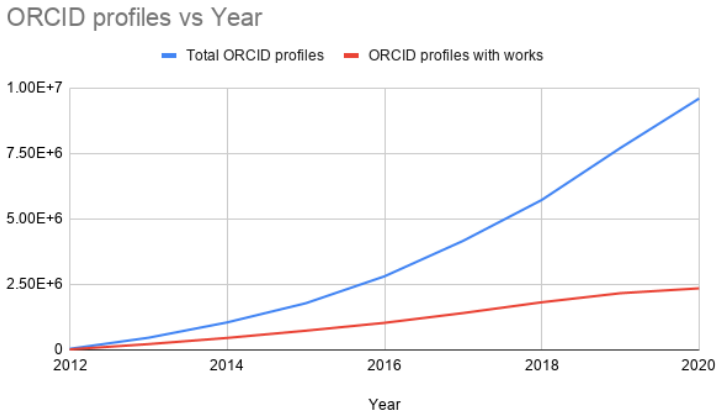

Fig. 1: Number of ORCID profiles per year.

registry on one side and *ORCID clients* on the other side. Clients include publishers, thematic and institutional repositories, data archives, such as CRIS systems and catalogues, hence services supporting the deposition of metadata and/or files relative to research outputs with referrals to authors' ORCID iDs.

Needless to say, ORCID popularity among organisations/researchers and clients has significantly increased over the years. Despite being still far from being adopted by the whole academic/research community, ORCID iDs are increasingly referenced upon deposition and metadata registration of new research products.

Unfortunately, the ORCID overlay suffers from a number of misuses which are today hindering its full potential. Some of these are strictly related to the misuse of the ORCID registry, while others can be identified in the misuse of ORCID referrals on the client-side. A clear example of the first class of misuse is given by the up-to-dateness of ORCID iDs. According to the official statistics[2], the whole dataset totals 10,324,141 registered, "live" (*sic*) profiles. However, many ORCID profiles - 7,593,481 (i.e., 73.5%) - exhibit no information about research outputs. Figure 1 reports the number of profiles registered every year (in blue) and the number of ORCID profiles with a nonempty work section (in red), as of October 2020. Moreover, a suspiciously increasing number of ORCID iDs seems not to have anything to share with academia and research in the first place, as the main purpose of such profiles appears to be related to URLs spamming and inducing visitors into click-baiting.

Examples of the second class of misuses occur instead whenever legitimate ORCID registrants and metadata curators, unwillingly or not, sometimes happen to make mistakes when depositing metadata of relevant research products

[2] ORCID official statistics, https://orcid.org/statistics

while linking them to the related ORCID iDs. Indeed, encountering profiles with mistakenly attributed works is not as infrequent as one might reckon.

It is therefore of pivotal importance to assess the presence of such inconsistencies as they can potentially be detrimental, undermine ORCID credibility in the eyes of the academic community, and ultimately hamper ORCID benefits and adoption at large.

In this paper, we qualitatively report on a selection of experienced issues that we drew from our direct experience while distilling the OpenAIRE Research Graph [8]. The Graph aggregates and deduplicates metadata from both ORCID registry and ORCID clients, links between research products metadata and ORCID iDs, thereby offering an overall map of the ORCID infrastructure as defined above. We report on and identify classes of ORCID misuses on the ORCID registry side and the ORCID clients side, with the intention of providing feedback and recommendations on how the related downfalls can be mitigated.

2 A report on ORCID misuses

The OpenAIRE project[3] aggregates metadata from over 12,000 sources, be they institutional/thematic Open Access repositories, publishers, registries and other aggregators, and builds the OpenAIRE Research Graph, redistributing it free of charge via periodic dumps on Zenodo [8]. An important component converging into the OpenAIRE Research Graph is DOIboost [7, 6]: a precomputed dataset containing Crossref[4] [5], Microsoft Academic Graph [11, 13, 12], ORCID and Unpaywall[5] [1]. Additionally, OpenAIRE features an algorithm for research outputs deduplication [9] so to reconcile different metadata descriptions of the same research output coming from different repositories.

Given the complexity and extreme heterogeneity of the aggregated information, OpenAIRE constitutes fertile ground for the identification of anomalies and misuse of ORCID. Presently, we target the following classes of misuses all emerged while tackling the consistency and quality of the information released in the OpenAIRE Research Graph. In the following, they are organised in two separated sections, one for ORCID misuse as a service and one for ORCID misuse at client-side. For each of them, we provide a generic description and report an example by linking to external services publicly accessible for open assessment.

2.1 ORCID registry misuse

Fake ORCID profiles. A fake ORCID profile is an ORCID profile whose registrant has nothing to share with academia and research, and whose existence on ORCID has the sole purpose of spamming links or inducing users into clickbaiting. Please have a look at the ORCID search for "bitcoin" (https://orcid.org/orcid-search/search?searchQuery=bitcoin) or the following ORCID profile:

[3] OpenAIRE project, https://www.openaire.eu
[4] Crossref, https://www.crossref.org
[5] Unpaywall, https://unpaywall.org

https://orcid.org/0000-0001-6997-9470. The profile here linked shows no worthy academic-related information whatsoever while offering a quite comprehensive collection of spam links to external websites, platforms, and services (e.g., Facebook, VK, Twitter, Youtube).

Poor quality ORCID profiles. While ORCID users have all the interest in keeping their profiles as informative and updated as possible, many of them still exhibit an unsatisfactory grade of information completeness despite being correctly referred from research outputs metadata. In fact, ORCID requires registrants to provide only their name and a valid email address; a family name is optional (see the ORCID search https://orcid.org/orcid-search/search?searchQuery =andrea), nor the current affiliation, thus yielding a significant amount of ambiguity. In 2016, ORCID was described as a mean to solve such uncertainty [10]; four years down the line, we can touch with our hands that ambiguity is an issue yet far to settle.

Overly-identified authors. Despite being against the whole philosophy of ORCID (i.e., unequivocally identify an individual in academia), some authors may, either intentionally or not, have registered to ORCID multiple times. Therefore, it might be the case that a publication gets deposited in two different archives using different ORCID iDs of a given author each time.

While finding duplicates in ORCID is, in general, as complex as finding author duplicates in research literature (i.e., none or limited ground truth), we can leverage the results yielded by deduplication algorithms [9] while producing the OpenAIRE Research Graph so to spot authors with multiple profiles as they use them in ORCID referrals. In fact, as soon as the OpenAIRE deduplication merges different ORCID referrals from the same research output, the various ORCID iDs used collide, and the author presents them all in the paper "reconciled" metadata. In this way, it is possible to detect these cases by checking research output metadata against ORCID profiles anagraphic information (e.g., checking name/surname correspondence). Indeed, this would be a lower bound estimation of the real number of duplicate ORCID profiles, as there are duplicates which we cannot spot out in this way. As an example, we report two individuals with two ORCID profiles: (https://orcid.org/0000-0002-0166-1973, https://orcid.org/0000-0002-2197-7270) and (https://orcid.org/0000-0003-4807-3623, https://orcid.org/0000-0002-7820-9889). Via manual inspection, we verified that they are not a case of homonymy.

Stale ORCID profiles. Besides referring its own ORCID iD when registering metadata of a publication, an author is, in general, advised to maintain up-to-date its own profile on the ORCID website as well. However, this practice, often time-consuming, is not encouraged nor enforced in any way.

Indeed, some authors appear to mint their own ORCID iD with the intention of using it in the future at deposition time, whilst updating the information in

their profiles as little and sporadically as possible. Most of the users prefer to take advantage of Crossref and publishers so to automatically update ORCID profiles on their behalf. As an example, the ORCID profile of the first author of the paper https://www.sciencedirect.com/science/article/abs/pii/S037842661830284X shows no curated work section (https://orcid.org/0000-0001-5001-2964).

2.2 ORCID clients misuse

ORCID clients offer users capabilities for (*i*) ingesting metadata (and/or files) where authors are associated to their ORCID iDs and (*ii*) adding ORCID iDs to authors of a metadata record. Such actions can be wizard-supported when clients implement direct connections with ORCID registry's APIs (see ORCID Search & Link Wizard[6]). Accordingly, users can unambiguously link their works to their profile while correctly referring the metadata record at hand with their own ORCID iD. Unfortunately, most ORCID clients only support a manual ingestion approach, which inevitably leads to a large number of mistakes. In the following we have identified two main classes: *non-existent ORCID iDs*, and *wrongly-attributed ORCID iDs*.

Non-existent ORCID iDs. As a matter of fact, manual ingestion is subject to human errors, whose common cases are *typos* and *misinterpretation*. The former is enough to lead to a (sometimes) well-formed, yet non-existent, ORCID iDs; for example, the author Jostein Askim in https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/E3W52Q (see the Metadata section) provides an ORCID iD where the last character is missing. Similarly, the article https://doi.org/10.31732/2663-2209-2019-53-159-168 reports a well-formed ORCID iD (0000-0001-2345-6754), which however does not resolve. Manifestations of the misinterpretation instead are, for example, emails or other author identifiers provided in place of ORCID iDs.

Wrongly-attributed ORCID iDs. A wrong attribution happens whenever an author of a given research output is attributed with a different ORCID iD. This misuse reflects the mistake happening during manual data entry/registration, when the registrant, for unknown reasons, mistypes or provides an irrelevant ORCID iD for one or more authors. As an example, the paper https://pubs.acs.org/doi/10.1021/acs.analchem.6b04010 shows two examples of wrong attribution: Zhaoyang Wu has the ORCID iD of a coauthor, and Ruqin Yu is attributed the ORCID https://orcid.org/0000-0002-7412-8360 of another researcher, not in the original author list. The first example identifies a subclass of wrongly attributed ORCID iDs, here referred as *shuffled*. Shuffled ORCID iDs occur whenever the authors of a given research output are attributed in the metadata the ORCID iDs of other coauthors. Shuffled ORCID iDs are not necessarily mutually exchanged

[6] ORCID Search & Link Wizard, https://support.orcid.org/hc/en-us/articles/360006973653-Add-works-by-direct-import-from-other-systems

in couples, and the shuffle can involve any number of authors participating in a publication (i.e., even one).

3 Discussion

Despite its adoption is far from being comprehensive, ORCID has become, without doubt, a central service in scholarly communication, and we are deemed to see its adoption increasing across all disciplines of science and research in the years to come. However, in its current status, ORCID presents several anomalies and misuses that, on top of being detrimental to scholarly communication practices, can undermine the service credibility and potentially hamper its adoption at large.

In the previous section, we have qualitatively reported on and documented a collection of issues in the usage of ORCID iDs to stress the need for future analysis so to exploit at best the benefits of the ORCID overlay. We intend to quantitatively extend the present study, so to provide a detailed outlook on the current incidence and impact of the aforementioned classes of misuse in ORCID. In this section, we elaborate on possible actions which could be undertaken in order to address such issues, making a distinction between ORCID registry misuses and ORCID clients misuses.

3.1 Mitigation of ORCID registry misuses

Four classes of misuses have been identified: fake ORCID profiles, poor-quality profiles, overly-identified users, and profile up-to-dateness.

In order to solve, or at least mitigate, the anomalies mentioned above, a combined set of actions could be taken, which mainly suggest the ORCID registry to be more "restrictive" and "rigorous" when approaching its users, i.e. researchers and organisations.

First of all, individual institutions and research organisations could, for example, take over control of the registration process to ORCID, as well as the dissemination of ORCID "philosophy" and best practices. The recent inclusion of institutional Identity Providers (IDPs) in ORCID login is undoubtedly a step forward in this direction; however, it does not suffice alone. Indeed, users still can register to ORCID providing a minimal amount of information, i.e., just a name and a valid, potentially non-institutional, email, or even register via Google and Facebook single sign-on. ORCID registrants could forget how they created their account in the first place and register afresh using one of the several other methods provided, quickly ending up in duplicates and fragmentation of information. Surely, enforcing the access to ORCID only via institutional IDPs would pose issues upon researchers relocation to another institution, and the consequent change of IDP, which certainly has to be handled accordingly. Nonetheless, we believe that this would dramatically improve the accuracy of the information contained in ORCID, and put a stop to the proliferation of fake profiles. Furthermore, an extensive analysis of the email domains of the existing user base

would enable ORCID to identify non-institutional ones and notify the authors for an action. At that point, non-compliant ORCID users could be flagged and deprecated (but never fully deleted, as they are PIDs[7]) if no action is taken within a given cool-down period. Indeed, a minority of rather uncommon independent researchers not affiliated to research institutions does exist. Such cases need to be handled with special care. One possible solution to this problem could consist in contacting coauthors or colleagues present on ORCID for an "endorsement".

To improve the quality and completeness of profiles, ORCID could ask registered users to simply provide more information, such as family name, alternate name forms and affiliations. Likewise, reserving other alphabets and charsets (e.g., Cyrillic, Greek, Arabic, Chinese and Japanese ideograms) only in the field reserved for other form names and using the main ones for transliteration into Roman alphabet would improve the clarity of the anagraphic information.

To disambiguate overly-identified users, ORCID could proactively identify potential duplicates thanks to AI/ML-driven techniques and/or by engaging the users directly involved. More generally, ORCID should initiate a quality program, similar to the one of Scopus, in order to clean up its record and ensure its reliability at a global level.

3.2 Mitigation of ORCID clients misuse

Manual (non-validated) ingestion of metadata at the ORCID client-side can greatly disrupt the benefits of the ORCID infrastructure by polluting the scholarly communication record. Potential causes are: non-existent ORCID iDs, and wrongly-attributed iDs. In order to mitigate these issues, ORCID clients should be equipped with tools that allow users, during metadata ingestion, to automatically recover ORCID iDs directly from the ORCID registry, systematically avoiding manual insertion. Such capabilities would exclude the non-existent ORCID iDs issue and certainly mitigate the wrongly-attributed issue as a whole (cases of homonymy still could be troubling) while completely avoiding the shuffle subset.

Acknowledgements

This work was co-funded by the EU H2020 project OpenAIRE-Advance (Grant agreement ID: 777541).

References

1. Else, H.: How Unpaywall is transforming open science. Nature **560**(7718), 290–291 (2018). https://doi.org/10.1038/d41586-018-05968-3

[7] https://support.orcid.org/hc/en-us/articles/360006896634-Removing-your-additional-or-duplicate-ORCID-iD

2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabási, A.L.: Science of science. Science **359**(6379), eaao0185 (mar 2018). https://doi.org/10.1126/science.aao0185
3. Haak, L.L., Fenner, M., Paglione, L., Pentz, E., Ratner, H.: ORCID: A system to uniquely identify researchers. Learned Publishing **25**(4), 259–264 (oct 2012). https://doi.org/10.1087/20120404
4. Haak, L.L., Meadows, A., Brown, J.: Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. Frontiers in Research Metrics and Analytics **3**(October), 1–7 (2018). https://doi.org/10.3389/frma.2018.00028
5. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, 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
6. La Bruzzo, S., Manghi, P., Mannocci, A.: Doiboost dataset dump (Dec 2019). https://doi.org/10.5281/zenodo.3559699, https://doi.org/10.5281/zenodo.3559699, When citing this dataset please cite this record in Zenodo and the relative article: La Bruzzo S., Manghi P., Mannocci A. (2019) OpenAIRE's DOIBoost - Boosting CrossRef for Research. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, doi:10.1007/978-3-030-11226-4_11
7. La Bruzzo, S., Manghi, P., Mannocci, A.: OpenAIRE's DOIBoost - Boosting Crossref for Research. pp. 133–143. Springer, Cham (jan 2019). https://doi.org/10.1007/978-3-030-11226-4_11
8. Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Löhden, A., Bäcker, A., Mannocci, A., Horst, M., Jacewicz, P., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., Ioannidis, A., Manola, N., Principe, P.: Openaire research graph dump (Nov 2020). https://doi.org/10.5281/zenodo.4279381
9. Manghi, P., Atzori, C., De Bonis, M., Bardi, A.: Entity deduplication in big data graphs for scholarly communication. Data Technologies and Applications **54**(4), 409–435 (jun 2020). https://doi.org/10.1108/DTA-09-2019-0163
10. Meadows, A.: Everything you ever wanted to know about ORCID... but were afraid to ask. College and Research Libraries News **77**(1), 23–30 (jan 2016). https://doi.org/10.5860/crln.77.1.9428, https://crln.acrl.org/index.php/crlnews/article/view/9428
11. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion. pp. 243–246 (2015). https://doi.org/10.1145/2740908.2742839
12. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020). https://doi.org/10.1162/qss_a_00021
13. Wang, K., Shen, Z., Huang, C., Wu, C.H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., Rogahn, R.: A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data **2**, 45 (dec 2019). https://doi.org/10.3389/fdata.2019.00045
14. Warner, S.: Author identifiers in scholarly repositories. Journal of Digital Information **11**(1), 1–10 (2010)

15. Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., Stanley, H.E.: The science of science: From the perspective of complex systems. Physics Reports **714-715**, 1–73 (nov 2017). https://doi.org/10.1016/j.physrep.2017.10.001