We Can Make a Better Use of ORCID: Five Observed Misapplications

RESEARCH PAPER

MIRIAM BAGLIONI
PAOLO MANGHI
ANDREA MANNOCCI
ALESSIA BARDI

*Author affiliations can be found in the back matter of this article

ABSTRACT

Since 2012, the "Open Researcher and Contributor ID" organisation (ORCID) has been successfully running a worldwide registry, with the aim of "providing a unique, persistent identifier for individuals to use as they engage in research, scholarship, and innovation activities". Any service in the scholarly communication ecosystem (e.g., publishers, repositories, CRIS systems, etc.) can contribute to a non-ambiguous scholarly record by including, during metadata deposition, referrals to iDs in the ORCID registry.

The OpenAIRE Research Graph is a scholarly knowledge graph that aggregates both records from the ORCID registry and publication records with ORCID referrals from publishers and repositories worldwide to yield research impact monitoring and Open Science statistics. Graph data analytics revealed "anomalies" due to ORCID registry "misapplications", caused by wrong ORCID referrals and misexploitation of the ORCID registry. Albeit these affect just a minority of ORCID records, they inevitably affect the quality of the ORCID infrastructure and may fuel the rise of detractors and scepticism about the service.

In this paper, we classify and qualitatively document such misapplications, identifying five ORCID registrant-related and ORCID referral-related anomalies to raise awareness among ORCID users. We describe the current countermeasures taken by ORCID and, where applicable, provide recommendations. Finally, we elaborate on the importance of a community-steered Open Science infrastructure and the benefits this approach has brought and may bring to ORCID.

CORRESPONDING AUTHOR:
Andrea Mannocci
ISTI-CNR, Pisa, IT
andrea.mannocci@isti.cnr.it

KEYWORDS:
ORCID; Scholarly Communication; Open Science; Disambiguation; Misapplication

TO CITE THIS ARTICLE:
Baglioni, M, Manghi, P, Mannocci, A and Bardi, A. 2022. We Can Make a Better Use of ORCID: Five Observed Misapplications. *Data Science Journal*, 20: 38, pp. 1–12. DOI: *https://doi.org/10.5334/dsj-2021-038*

1 INTRODUCTION

A precise and reliable identification of researchers and the pool of works and knowledge they contributed to would greatly benefit scholarly communication practices and facilitate the understanding of science (Haak et al., 2018). 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 (Warner, 2010; Zeng et al., 2017; Fortunato et al., 2018).

Since 2012, ORCID (Haak et al., 2012), the "Open Researcher and Contributor ID" organisation, has been successfully running a worldwide registry,[1] which allows researchers and collaborators, hereafter *ORCID registrants*, to mint alphanumeric iDs and maintain a core set of relevant information such as name, surname, affiliations, works, and projects in their so-called "ORCID records". The ORCID registry enables the decentralised growth of a non-ambiguous scholarly record, where every product of science, e.g., publications, datasets, software, is clearly attributed to an author via the ORCID iDs. To this aim, scholarly communication data sources – e.g., institutional and thematic repositories, publishers, Current Research Information Systems (CRIS) systems, data repositories, etc. – are increasingly including references to author ORCID iDs, hereafter *ORCID referrals*, in the bibliographic metadata of their research products.

The correct behaviour of ORCID registrants and the proper usage of ORCID referrals is key to deliver a sound and complete ORCID infrastructure, enabling two modes of consumption of the ORCID registry as depicted in ***Figure 1***:

- *ORCID record consultation*: users checking ORCID records via the ORCID portal (e.g., editors);
- *ORCID records batch consumption*: users downloading and processing ORCID content in batch for research purposes (e.g., Science of Science, scientometrics) or to provide scholarly communication added-value services (e.g., Scopus, Dimension, OpenAIRE) by integrating ORCID data with other scholarly datasets.

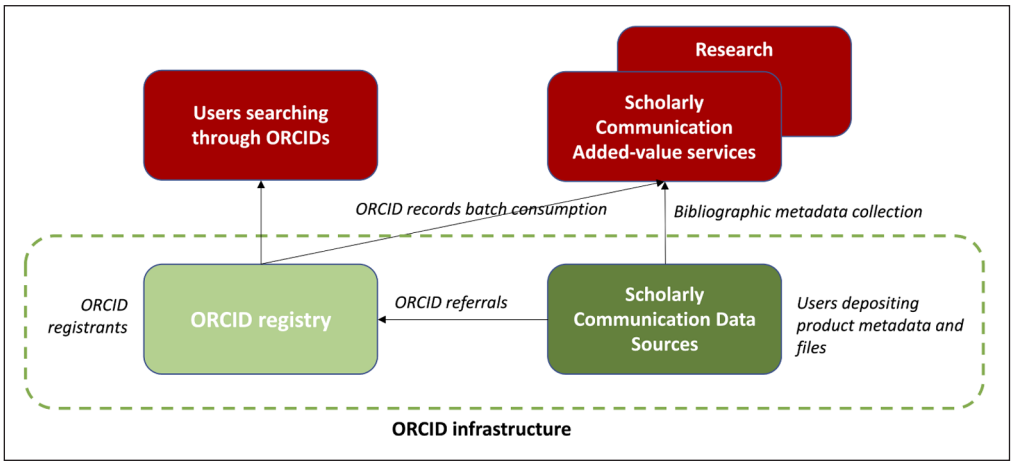


Figure 1 ORCID infrastructure: services and usages.

ORCID iDs are steadily growing over the years, as well as ORCID referrals upon deposition of metadata and files of new research products. ***Figure 2*** reports the number of records registered every year, as derived from the ORCID public data dump as October 2021.

Unfortunately, a portion of ORCID user base happens to use the service (intentionally or not) incorrectly, thus jeopardising the quality of the data in the registry and the scholarly record as a whole. Errors and anomalies ingested via batch consumption of ORCID are inevitably propagated downstream and can affect the quality of the intended reuse. Such cases, although far from being the norm, fuel the rise of detractors and scepticism about ORCID and unjustifiably cast a shadow over its uptake and success (Leopold, 2016; Frank Ritter, 2020; Teixeira da Silva, 2020, 2021a,b).

1 *https://orcid.org*.

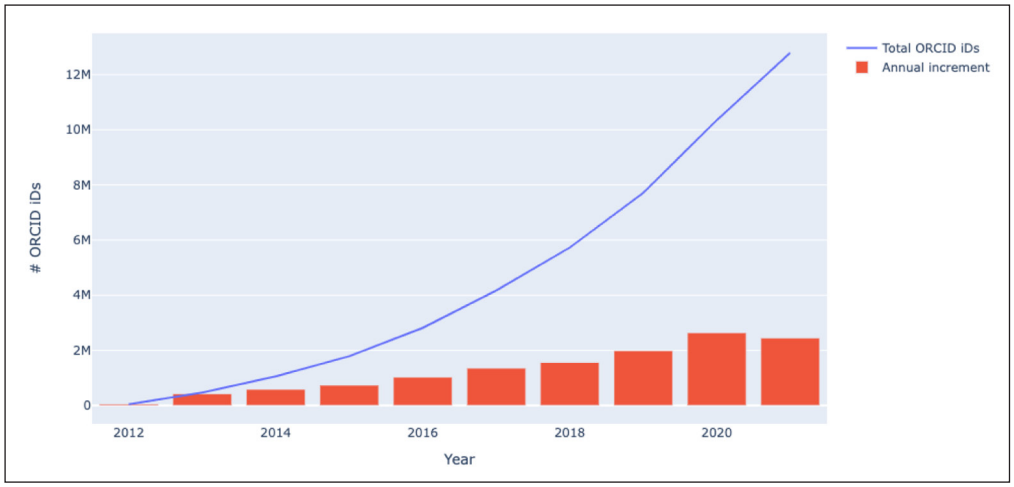


Figure 2 Number of ORCID records per year. The bar plot indicates the annual increment, while the line reports the total number of ORCID iDs through the years. (source: ORCID's public data file, October 2021).

Given the pivotal role of ORCID, essential for the scholarly communication infrastructure as a whole, this work aims at investigating the reasons, identifying the typologies, and possible countermeasures of such *ORCID misapplications*. To this aim, we leverage the OpenAIRE Research Graph[2] (Manghi et al., 2020a) (hereafter, the Graph), which is one of the core services of the OpenAIRE AMKE,[3] a not-for-profit legal entity operating an infrastructure that offers global services in support of Open Science scholarly workflows, and member of ORCID since 2020. The Graph is a knowledge graph that aggregates metadata from 96,514 scholarly sources (as of October 2021), comprising publications, research data and software repositories, publishers, and registries, including ORCID, ROR, re3data, OpenDOAR, Crossref, DataCite, providing an extensive collection of literature, datasets, software with ORCID referrals. Its aggregation and deduplication framework, delivers bibliographic records that include links between researcher names and ORCID iDs as collected from both ORCID records and other scholarly services and tools. As such, the Graph materialises the ORCID infrastructure described above and constitutes fertile ground to perform an investigation on such anomalies.

This paper is structured as follows. In Section 2, we identify and qualitatively document the misapplications related to ORCID registrants, e.g., users creating multiple ORCID iDs or fake ORCID iDs, and ORCID referrals, e.g., users of repositories specifying wrong ORCID iDs in the metadata while depositing research products. Section 3 describes the countermeasures taken by ORCID and, where applicable, provides recommendations to ORCID users, in order to further mitigate the side effects of ORCID misapplications. In addition, we provide some food for thoughts to the ORCID registry on how to further leverage on community-engagement as a way to deliver the optimal service for ORCID users. Making synergies between communities and services is the key towards a sustainable and community-driven Open Science infrastructure. Finally, in Section 4, we summarise and discuss possible future extensions of the present study.

2 A REPORT ON ORCID MISAPPLICATIONS

"ORCID registrants" and the "users referencing ORCID iDs from scholarly services and tools" are the two user groups that are imputable for the misapplications described in this paper. Their incorrect behaviour may undermine the quality of ORCID on one side and jeopardise the scholarly record on the other side. The consequences become particularly evident when the content of ORCID is aggregated and integrated with other scholarly data sources (i.e., batch consumption user group). Indeed, this work and its motivations find their roots in the processes of metadata aggregation behind the population of the OpenAIRE Research Graph.

The Graph is a service that aggregates from over 96,514 sources (institutional and thematic Open Access repositories, publishers, journals, registries, National aggregators, etc.) metadata records about publications, research data, research software, and other research products with

2 *https://graph.openaire.eu*.

3 *https://www.openaire.eu*.

semantic relationships among them and other research entities, such as organisations, funders, projects, and data sources themselves. Core data sources of the Graph are DataCite,[4] Crossref[5] (Hendricks et al., 2020), and thousands of institutional, thematic repositories and aggregators. Crossref is ingested via DOIboost (La Bruzzo et al., 2019a, b), a pre-computed dataset that enriches Crossref with the Microsoft Academic Graph (Sinha et al., 2015; Wang et al., 2019, 2020), Unpaywall[6] (Else, 2018), and dumps of ORCID records. In this context, ORCID is integrated as an "inverted list" of bibliographic records, each bearing ORCID iDs for the authors that claimed that record via their ORCID profile. All such records are harmonised and deduplicated (Manghi et al., 2020b) so as to build one bibliographic record out of the many describing the same research product but collected from different sources. Of particular interest to this investigation, is the fact that such "richer" metadata records, feature author/creator metadata which may bear a list of ORCID iDs, as collected from ORCID records and as collected from data sources (ORCID referrals). The Graph is redistributed free of charge via periodic dumps on Zenodo[7] (Manghi et al., 2020a) and made available via open APIs[8] for researchers, organisations, and companies, among them Elsevier's Scopus and the European Commission Participant Portal.

Given the number and variety of data sources it federates, including ORCID itself as a large number of ORCID referrals, the Graph constitutes fertile ground for the identification of anomalies and misapplications of ORCID. In the following, we describe two classes of misapplications that emerged while tackling the consistency and quality of the information released in the OpenAIRE Research Graph: *ORCID registry misapplications*, i.e., users abusing of or making mistakes while using ORCID, and *ORCID referral misapplications*, i.e., users making mistakes referring to ORCID iDs from scholarly communication services. We identify several specific misapplications for each class, which we describe and report via examples by linking to external services publicly accessible for open assessment, and by including screenshots and links to Web snapshots on Wayback Machine[9] whenever possible in order to keep a full track record of the results here described.

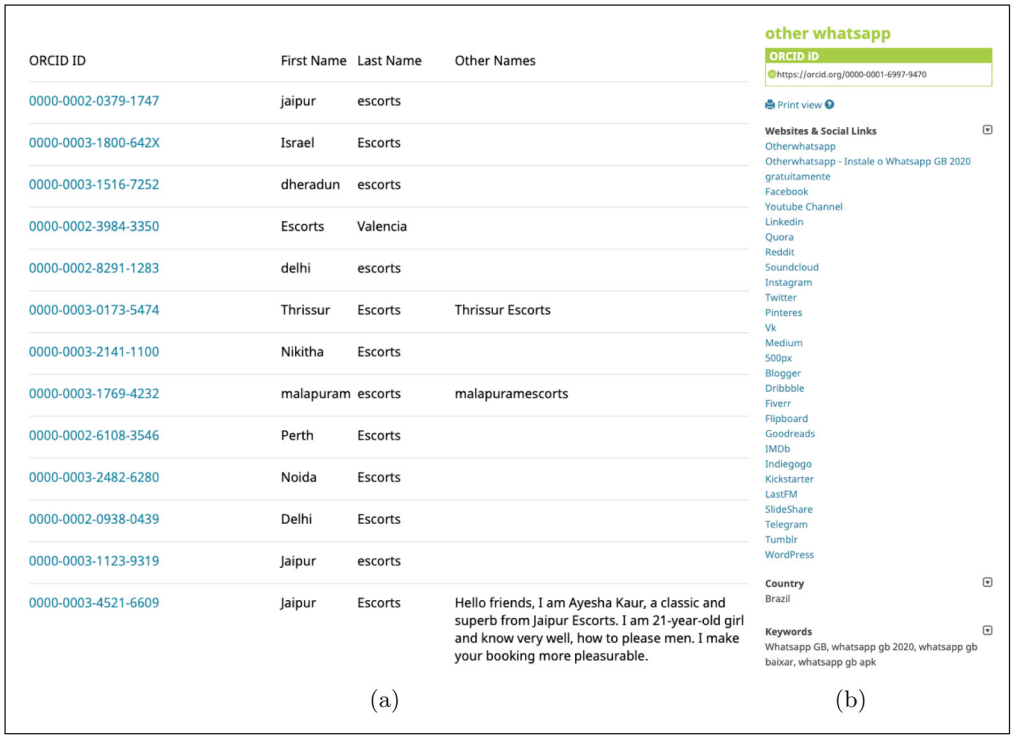
2.1 ORCID REGISTRY MISAPPLICATIONS

2.1.1 Fake ORCID records

A fake ORCID record is a record whose registrant has nothing to share with academia and research, and whose existence on ORCID just contributes to increasing the noise in the platform. While multiple intents can drive such malpractice, in the majority of the cases, it is clear that the main objective is to link to external domains to boost their ranking on search engines (i.e., link farming). As an example, have a look at the ORCID search for "escorts" reported in ***Figure 3a***.[10] As ORCID is on a constant arc of improvement, the query might return different results (possibly none); in such a case, the reader may experiment with any other common spam-like buzzword of her/his choice.

Most fake records show no worthy academic-related information while offering a collection of spam links to external websites, platforms, and services (e.g., Facebook, VK, Twitter, YouTube), as reported in ***Figure 3b***. This is a typical pattern witnessed in most open services online, which can be tackled via dedicated prevention, curation and detection activities. "Ironically, this exercise is largely futile, as links to other sites from ORCID records are tagged with NoFollow codes" in order to instruct Web crawlers (Shillum et al., 2021), *de facto* preventing this practice from being effective. Yet, the creation of fake ORCID records relentlessly continues and requires labour-intensive activities from the ORCID data curation team, which tackles the problem via heuristics and, only recently, by experimenting with machine learning techniques.

4 *https://datacite.org*.
5 *https://www.crossref.org*.
6 *https://unpaywall.org*.
7 *https://zenodo.org*.
8 *http://develop.openaire.eu*.
9 *https://web.archive.org*.
10 *https://orcid.org/orcid-search/search?searchQuery=escorts*.



ORCID ID	First Name	Last Name	Other Names
0000-0002-0379-1747	jaipur	escorts	
0000-0003-1800-642X	Israel	Escorts	
0000-0003-1516-7252	dheradun	escorts	
0000-0002-3984-3350	Escorts	Valencia	
0000-0002-8291-1283	delhi	escorts	
0000-0003-0173-5474	Thrissur	Escorts	Thrissur Escorts
0000-0003-2141-1100	Nikitha	Escorts	
0000-0003-1769-4232	malapuram	escorts	malapuramescorts
0000-0002-6108-3546	Perth	Escorts	
0000-0003-2482-6280	Noida	Escorts	
0000-0002-0938-0439	Delhi	Escorts	
0000-0003-1123-9319	Jaipur	escorts	
0000-0003-4521-6609	Jaipur	Escorts	Hello friends, I am Ayesha Kaur, a classic and superb from Jaipur Escorts. I am 21-year-old girl and know very well, how to please men. I make your booking more pleasurable.

(a)

(b)

Figure 3 Fake records returned by the query "escorts" (**a**) and details of a fake record (**b**).

2.1.2 Over-identified users

Despite being against the whole philosophy of ORCID (i.e., unequivocally identify an individual in academia), some users may, either intentionally or not, be over-identified, i.e., registered to ORCID multiple times. The automatic identification of such cases is not trivial, due to pseudonyms and to homonyms present in the data. However, a lower bound use case can be detected by identifying publication records claimed across different ORCID records, for users with the same name and surname. As an example, we report two individuals with two ORCID records, identified via the methodology above, and then manually verified.[11]

ORCID is taking all possible countermeasures by paving the ORCID registration procedure with controls and prompting the user with confirmation popups whenever any of the emails specified by the user or similar name/surname combinations are already present in the database. Still, as in the first example above, ORCID users may dodge such controls and create multiple iDs to perpetrate a misapplication of the registry. Not necessarily with malicious intent, as for example ORCID iDs are assigned by some universities to the newly employed researchers,[12] who carelessly and happily create a new one. In all the cases where this happens by mistake, ORCID users can declare the redundancy between their records, electing one of them as the main one, and redirecting the others to this (Haak, 2014).

It is worth noticing that pseudonyms are not to be considered a misapplication of ORCID, as authors have the right to decide about their identity or multiple ones.

2.1.3 Overloaded ORCID records

While ORCID iDs are in general meant for individual researchers and contributors, we encountered several ORCID records that, in our view, deliberately bend this principle and derail from the expected functioning of the service. In particular, we experienced the presence of records both for academic venues,[13] and research institutes and organisations.[14]

11 *https://web.archive.org/web/20210709111517/https://orcid.org/0000-0002-0166-1973*, *https://web.archive.org/web/20210709111621/https://orcid.org/0000-0002-2197-7270* and *https://web.archive.org/web/20210709111721/https://orcid.org/0000-0003-4807-3623*, *https://web.archive.org/web/20210709111825/https://orcid.org/0000-0002-7820-9889*.

12 *https://info.orcid.org/universities-now-creating-orcid-ids-for-their-researchers-and-scholars/*.

13 *https://orcid.org/orcid-search/search?searchQuery=jurnal)*.

14 *https://orcid.org/orcid-search/search?searchQuery=institute*.

Venues like these three Indonesian journals,[15] the Spanish CIC nanoGUNE Consolider, the Sethu Institute of Technology, or the Indian Institute of Foreign Trade[16] feature a portfolio of works which associates several distinct authors to the venues' ORCID iD (as reported in ***Figure 4***), rather than to the authors' ones (if available).

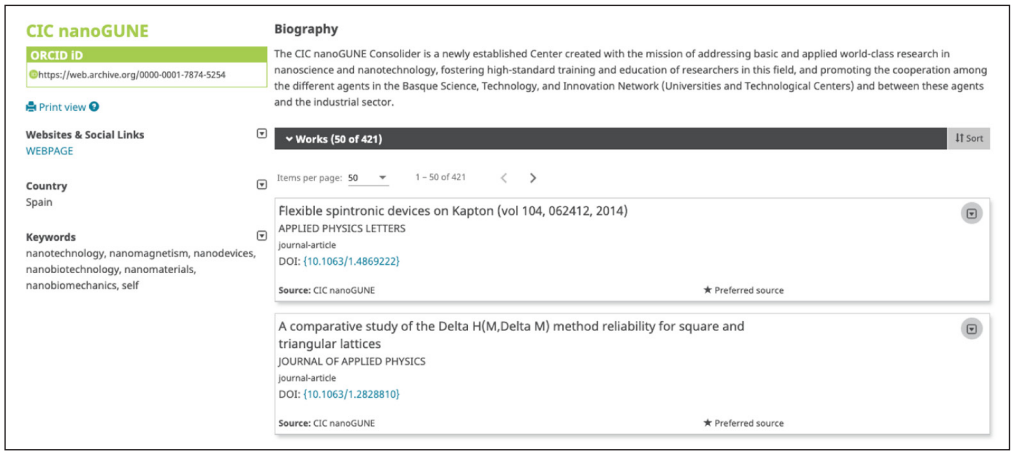


Figure 4 Overloaded ORCID record associated to a research organisation.

This, of course, on top of being a profoundly wrong practice according to ORCID's rules of participation, is also unnecessary as there are already registries purposely devised for journals and research organisations (e.g., Directory of Open Access Journal[17] (DOAJ) and Global Research Identifier Database[18] (GRID.ac), respectively, where incidentally neither the three journals, nor the Sethu institute can be found.

To the best of our knowledge, the accounts belonging to this type of misuse are limited in amount and thus are pretty subtle to identify, especially with an automated approach. Nonetheless, they are treated by ORCID as exceptional cases of impostors' accounts and deactivated whenever discovered.

2.2 ORCID REFERRAL MISAPPLICATIONS

ORCID referrals happen whenever scholarly communication services provide the users with capabilities for *i*) ingesting/curating metadata where individuals are associated to ORCID iDs and *ii*) adding ORCID works to an ORCID record (i.e., only for ORCID members[19]).

Such actions can be wizard-supported when referrals are implemented via direct connections/integrations with ORCID registry's APIs (see ORCID Search & Link Wizard[20]). Accordingly, users can unambiguously link their works to their record while correctly referring to the metadata record at hand with their own ORCID iD.

Unfortunately, not every ORCID referral comes from a tightly integrated workflow (i.e., with authenticated API calls). In fact, many ORCID referrals are just supported by a manual data entry approach where the user can freely provide the ORCID iDs intended to refer, which can lead to mistakes. Examples include Zenodo and Figshare,[21] which count around 4 million research outcomes combined, and most of the 1000 DSpace[22] installations worldwide, of which only a handful integrates ORCID API validation.

15 *https://web.archive.org/web/20210709110118/https://orcid.org/0000-0003-3880-4082*, *https://web.archive.org/web/20210709112555/https://orcid.org/0000-0002-9548-3943*, *https://web.archive.org/web/20210709110933/https://orcid.org/0000-0002-2653-6095*.

16 *https://web.archive.org/web/20210709123632/https://orcid.org/0000-0001-7874-5254*, *https://web.archive.org/web/20210709111056/https://orcid.org/0000-0002-8835-1336*, *https://web.archive.org/web/20210709113003/https://orcid.org/0000-0002-2546-3285*.

17 *https://doaj.org*.

18 *https://grid.ac*.

19 *https://orcid.org/members*.

20 *https://support.orcid.org/hc/en-us/articles/360006973653-Add-works-by-direct-import-from-other-systems*.

21 *https://figshare.com*.

22 *https://duraspace.org/dspace*.

Two main classes of misapplication derive from this scenario, and are described in the following: *non-existent ORCID iDs*, and *wrongly-attributed ORCID iDs*.

2.2.1 Non-existent ORCID iDs

Manual ingestion is prone to human errors, including *typos* and *misinterpretation*. The former is enough to lead to a (sometimes) well-formed, yet non-existent, ORCID iD; for example, an author of the paper[23] illustrated in ***Figure 5*** provides an ORCID iD where the last digit is missing (0000-0002-5675-693). Similarly, another article[24] reports a well-formed ORCID iD (i.e., 0000-0001-2345-6754), which however – for the time being – does not resolve[25]. Manifestations of the misinterpretation instead are, for example, emails or other author identifiers provided in place of ORCID iDs. All these are made possible by the absence of the most basic form validation countermeasures in Web UIs.

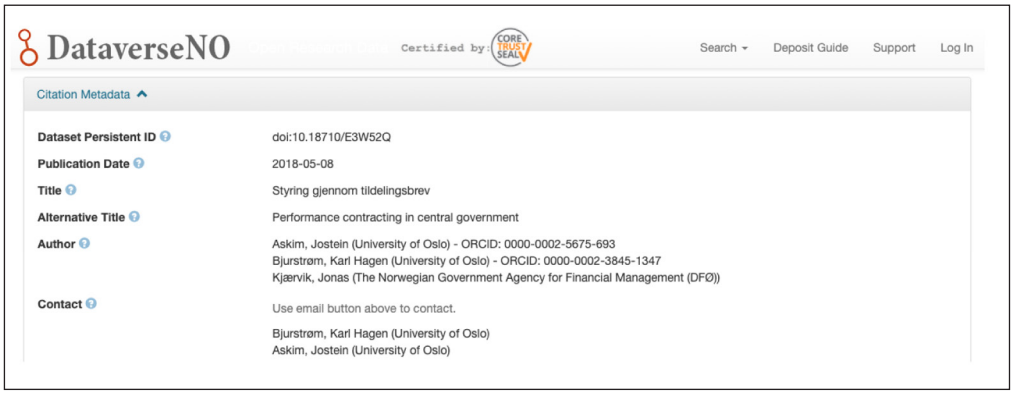


Figure 5 Referral to a syntactically wrong ORCID iD.

2.2.2 Wrongly-attributed ORCID iDs

A wrong attribution happens whenever an author of a given research output is attributed with a different, yet existing, ORCID iD. This misapplication 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. For example, another paper[26] shows two examples of wrong attribution: one author has been attributed the ORCID iD of another co-author, while another author is attributed the ORCID iD of another researcher, not in the original author list. The first example identifies a subclass of wrongly attributed ORCID iDs, here indicated as *shuffled*. Shuffled ORCID iDs are not necessarily mutually exchanged in couples, and the shuffle can involve any number of iDs among the authors participating in a publication (i.e., even one moving from the legitimate owner to a co-author).

2.3 PRELIMINARY QUANTITATIVE IMPACT ASSESSMENT

As mentioned above, the objective of such qualitative analysis is to identify and classify the reasons behind some of the anomalies derived by the misuse of ORCID and the misuse of ORCID referrals. In this section, however, we provide preliminary results for the quantitative impact of some of the witnessed misapplications, to strengthen our reflections and recommendations in the subsequent section. Where available, the numbers have to be considered "lower bounds", as it is technically non-trivial to match authors names/surnames, which are provided under diverse forms, and many of the use cases cannot be identified via simple author name matches (e.g., pseudonyms). We are currently developing anomaly detection techniques that *i*) exploit more advanced author name normalisation to improve name match confidence and *ii*) heavily rely on AI to identify anomalies based on "common behaviour".

23 *https://web.archive.org/web/20210715105512if_/https://dataverse.no/dataset.xhtml?persistentId=doi%3A10.18710%2FE3W52Q#datasetForm:tabView:metadataMapTab*.

24 *https://doi.org/10.31732/2663-2209-2019-53-159-168*.

25 *https://orcid.org/0000-0001-2345-6754*.

26 *https://web.archive.org/web/20210715104457/https://pubs.acs.org/doi/10.1021/acs.analchem.6b04010*.

The intention is to proactively identify the individual anomalies and report them to ORCID or scholarly communication services to improve the overall quality of the ORCID ecosystem.

By analysing the portion of the OpenAIRE Research Graph overlapping with the ORCID registry, where ORCID records have at least one associated work (live profiles), we could (or not) calculate the following quantitative impact, preliminarily confirming the qualitative impact on ORCID is rather low:

- *Fake ORCIDs*: fake ORCIDs cannot be identified using the Graph, bur rather detecting common fake users' behaviour via AI techniques currently being investigated in a separate analysis;
- *Over-identified users*: as a first analysis it impacts for the 0.1% of the live profiles; the technique is currently based on a name matching strategy across publication records claimed by different ORCID users with the same name/surname;
- *Overloaded ORCID records*: as a first analysis it impacts for the 0.00005% of the live profiles; the technique is currently based on a name matching strategy, which makes sure the ORCID user name is "very unlikely" an author of all publications claimed under the profile;
- *Non-existent ORCID iDs*: as a first analysis it impacts for the 0.3% of the ORCID referrals;
- *Wrongly-attributed ORCID iDs*: as a first analysis the 0.5% of the live profiles has been wrongly attributed at least once.

3 DISCUSSION

ORCID has become a central service in scholarly communication, deemed to be increasingly adopted across all disciplines of science and research in the years to come. Its role in enabling a connected scholarly communication record is crucial, and with it the importance of its correct usage. Our analysis highlighted why this is not always the case but also how, despite their minor impact in terms of coverage, the undermining anomalies may affect the experience of ORCID consumers, both humans and service providers, and bloom scepticism in the community.

This section elaborates on actions that could be undertaken to lower the effects of the misapplication identified above, again partitioning between actions related to the ORCID registry and actions related to ORCID referrals.

3.1 ORCID REGISTRY USERS MISAPPLICATIONS: FOOD FOR THOUGHTS

Ideally, a digital registry for "people" would *i*) provide a unique identifier to a person, for disambiguation and reference purposes, *ii*) establish the core of mandatory/optional metadata properties characterising a person, to support use-cases such as *search* (i.e., find an ID, given person details), *verification* (i.e., given a person claiming the ownership of an ID, assure that assertion) and *disambiguation* as requested by the registry consumers, and *iii*) forbid registration to non-legitimate users (like the UK Driver and Vehicle Licensing Agency[27] (DVLA) or any other registration office in real life), i.e., eligibility of the user ("is the user a researcher?") and user identity verification ("is the user who he/she claims to be?"). All registries address point *i*) above, but in general may vary in how they interpret and implement *ii*) and *iii*).

More specifically, for each issued identity, the registry mandates a given set of metadata properties that strongly characterise the underlying registrant. The nature of such properties, i.e., cardinality, semantics, and interpretation, has a direct impact on the spectrum of use-cases supported by the registry, but may potentially hinder the openness and uptake of the registry (e.g., stricter requisites may discourage or exclude users).

Furthermore, the accuracy of gatekeeping, i.e., data quality controls, which exclude errors, impostors, double registrations, and keep the registry up to date, usually come at a price. The trade-off is, therefore, between registry quality and operational costs. Such choices are not trivial and are addressed through a long-term strategy, defining policies that can vary over time depending on both internal and external new requirements and needs.

27 *https://www.gov.uk/government/organisations/driver-and-vehicle-licensing-agency*.

3.1.1 ORCID registry: policy and strategy

In the case of ORCID, the large number of users, the ever-increasing volume of new registrants, and the particular application domain (with independent and early-career researchers) posed the preconditions for an approach fully devoted to openness (and persistence) – ORCID's ten founding principles.[28] In particular, at the time of writing, ORCID requires registrants to provide only their first name(s) and a valid email address; family names are optional, as well as the current affiliation(s). With respect to gatekeeping, ORCID's policy opts for openness, poses trust on users, and supports pre-registration and post-registration quality check activities. The registration process prevents double-registrations by email and backup email validation, and checks for the existence of existing records already associated with them. After registration, data curation activities are regularly performed, proactively identifying fake ORCID records and duplicates, which can be semantically related to remove ambiguity.

3.1.2 ORCID registry: policy pros and cons

While, on the one hand, such an approach allows the onboarding of budding, early-career and independent researchers, on the other hand, the lack of gatekeeping undoubtedly lowers the guard to profiteers and other erroneous or malicious uses of the service – a (real) story highlighting the setbacks of this choice is portrayed in (Teixeira da Silva, 2020). In this work, we have identified three classes of misapplications, due to such a choice of openness and inclusivity (fake ORCID records, overly-identified users, and overloaded ORCID records), but also reported their impact is confined to a relatively low portion of the overall amount of ORCID records, indicating the choice of openness and trust to be overall successful.

Yet, two main drawbacks still surface. On the one hand, the end-user perception of a few mistakes is amplified and exposes ORCID to criticism, as previously shown. Moreover, batch consumers of ORCID data (e.g., researchers, data scientists, and services building monitoring tools on ORCID, such as the OpenAIRE Graph) need to realise the tooling necessary to purge ORCID records caused by such misapplications (e.g., deduplication, fake detection, filtering), so as to avoid propagating the very same anomalies downstream and suffer from similar misjudgements. On the other hand, reconciling author details in bibliographic data with the legitimate ORCID iD can become somewhat challenging when an ORCID record is scarcely populated or most of the information is set to private by the owner. These two behaviours are perfectly compliant to ORCID's intended usage, nonetheless effectively hinder two other common use-cases featuring an agent (e.g., an editor) willing to *i*) find the ORCID iD of a researcher given some personal details or *ii*) verify that a researcher is effectively the one associated to an iD at referral time.

3.1.3 Afterthoughts

This paper is the result of the analysis made necessary in OpenAIRE to generate a high-quality knowledge graph inclusive of ORCID records. The afterthoughts of such analysis resulted in the following two reflections about ORCID and possible extra community engagement activities:

> **Openness and gatekeeping policies** As we learnt from ORCID community engagement policies and past community (Armstrong et al., 2015; Meadows et al., 2017; 2019) and members (Meadows and Laurel, 2019) surveys, ORCID closely liaise with its community to the point that current policies seemingly look like a choral expression of people's expectations and requirements. From the aforementioned surveys emerges an overwhelming favour towards ORCID mandates (from 72% to 85.9%), which highlights the urgent demand for ORCID iDs of the scientific community, but also hints an increased awareness and commitment. Such interest may also hint an opening towards a change of policies, for example by enforcing more mandatory fields in the ORCID record, e.g., surname, affiliation – the recent adaptation of the ORCID search, today returning complete ORCID record first, de facto fosters this idea, sponsoring users showing commitment and fully-fledged adoption of the registry. Open consultations could timely test the evolving

28 *https://info.orcid.org/what-is-orcid*.

expectations of ORCID stakeholders and user groups so as to make sure that the ORCID policies and roadmap are transparent and fit to community expectations.

Collaborative data curation ORCID data batch consumers (researchers, data scientists, and service providers), working on identifying solutions to identify ORCID misapplications, may become a special class of ORCID community to engage with, contributing with data and methods to ORCID data curation activities. This work, which brought valuable feedback and data to the ORCID team, contributed to the curation of the ORCID registry, can be seen as an example of such fruitful collaborations. Similarly, other initiatives may be openly engaged into similar actions to improve the quality of one of the main central services in the Open Science commons.

3.2 MITIGATION OF ORCID REFERRAL MISAPPLICATIONS: RECOMMENDATIONS

Manual (non-validated) data entry of ORCID referrals – as in Zenodo (in ***Figure 6***) or Figshare, as well as in many institutional repositories – may introduce errors in the ORCID infrastructure thereby polluting the scholarly communication record. Potential causes are non-existent ORCID iDs and wrongly-attributed iDs. Both misapplications contribute to the diffusion of wrong information via metadata records first exposed by scholarly communication data sources, then propagated via a number of added-value services that are exploiting this data, such as metadata aggregators, catalogues, knowledge graphs, monitoring dashboards, etc. A typical scenario is one where research product records describing the same object are collected from multiple sources, bearing different ORCID iDs for the same authors. In order to provide a high-quality service, e.g., impact monitoring, proper solutions need to be engineered to clean up the data. As a matter of fact, no ORCID referral can be trusted unless it is issued by a data source offering ORCID validation mechanisms. While checking the validity of the ORCID iDs, carried out by matching with ORCID registry data via iDs and author name strings, the service provider must manage following scenarios:

- *non-existent ORCID iDs*: removal of ORCID referrals pointing to non-existent ORCID iD;
- *wrongly-attributed ORCID iDs*: identification of wrongly attributed ORCID iDs and investigating the possibility of an ORCID iD's shuffling within the record, then either remove the iDs or re-shuffle the iDs;

Figure 6 Unrestricted ORCID field in Zenodo.

The ORCID iD verification step must be performed for all ORCID referral occurrences and include code to handle cleaning/normalisation of ORCID referrals strings and author name string match. The accuracy of such actions highly depends on the quality and completeness of the metadata on both the referral side and the ORCID registry side. Name strings, for example, are provided in many forms and are in some cases incomplete (e.g., first name is the only mandatory field for ORCID registrants).

As recommended by ORCID, in order to mitigate these issues, scholarly communication data sources intending to refer ORCID iDs should be equipped with tools that support users, during metadata ingestion, to automatically recover iDs directly from the ORCID registry, systematically avoiding manual insertion. This can be achieved via integration with ORCID public APIs (free, limited) and member APIs (paid, unrestricted). Such capabilities would exclude the non-existent ORCID iDs issue and certainly mitigate the wrongly-attributed subset, by avoiding the shuffled ORCID iDs and reducing the misattributed ORCID iDs to the unlikely cases caused by homonymy.

Known examples of such integration are DSpace v5/v6 and ArXiv, which implemented the necessary plugins to ensure a fail-proof ORCID referral mechanism. In ArXiv, the user depositing a research paper is responsible only for her/his own details and, most importantly, the referral of her/his own ORCID iD (fetched automatically by linking the ArXiv account to ORCID one, via authentication). Additional metadata about other authors, including validated ORCID referrals, can only be deposited by the authors themselves, once notified about the deposition. The implementation of such countermeasures falls beyond the duties of the ORCID registry, which offers clear APIs and workflows for scholarly communication data sources to embrace the ORCID infrastructure. The development cost are on the data sources and, as in many other potentially beneficial efforts, resources may not be invested in this direction.

4 CONCLUSIONS

In this paper, we described five ORCID misapplications in the way users interact with the registry or specify ORCID referrals from scholarly communication data sources. The aim is to raise awareness about such misapplications, in order to further limit their occurrences, and shed some light on the extensive actions taken by ORCID to overcome them. As a result of this analysis, in collaboration with ORCID, we also highlight the benefit of engaging with user communities, i.e., researchers and services using ORCID data, to ensure transparency and fit-for-purposeness of the high-level guidelines and policies implemented by ORCID.

In the future, we plan to extend the present work by performing a quantitative analysis of the misapplications here described, and provide methodologies to practically flag the anomalies and contribute to the quality and correctness of the global research record. More specifically, we started an in-depth study to tackle the fake records phenomenon as a whole, understand its characteristics and dynamics, and leverage machine learning approaches to detect potentially fake accounts.

ACKNOWLEDGEMENTS

This work was co-funded by the EU H2020 project OpenAIRE-Advance (Grant agreement ID: 777541). We would like to thank the ORCID team for the openness and cooperation shown to improve this work and, most importantly, to enhance the quality of the ORCID data.

AUTHOR AFFILIATIONS

Miriam Baglioni *orcid.org/0000-0002-2273-9004*
ISTI-CNR, Pisa, IT
Paolo Manghi *orcid.org/0000-0001-7291-3210*
ISTI-CNR, Pisa, IT; OpenAIRE AMKE, Athens, Greece
Andrea Mannocci *orcid.org/0000-0002-5193-7851*
ISTI-CNR, Pisa, IT
Alessia Bardi *orcid.org/0000-0002-1112-1292*
ISTI-CNR, Pisa, IT

REFERENCES

Armstrong, D, Haak, L, Meadows, A and **Stone, A.** 2015. ORCID 2015 Survey Report (final). *Technical report*, ORCID.

Else, H. 2018. How Unpaywall is transforming open science. *Nature*, 560(7718): 290–291, ISSN 14764687. DOI: *https://doi.org/10.1038/d41586-018-05968-3*

Fortunato, S, Bergstrom, CT, Börner, K, Evans, JA, Helbing, D, Milojevi, S, Petersen, AM, Radicchi, F, Sinatra, R, Uzzi, B, Vespignani, A, Waltman, L, Wang, D and **Barabási, A-L.** Mar 2018. Science of science. *Science*, 359(6379): ISSN 0036-8075. DOI: *https://doi.org/10.1126/science.aao0185*

Haak, L. January 2014. Managing duplicate orcid ids. *https://info.orcid.org/managingduplicate-orcid-ids*.

Haak, LL, Fenner, M, Paglione, L, Pentz, E and **Ratner, H.** Oct 2012. ORCID: A system to uniquely identify researchers. *Learned Publishing*, 25(4): 259–264. ISSN 09531513. DOI: *https://doi.org/10.1087/20120404*

Haak, LL, Meadows, A and **Brown, J.** 2018. Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. *Frontiers in Research Metrics and Analytics,* 3(October): 1–7, ISSN 2504–0537. DOI: *https://doi.org/10.3389/frma.2018.00028*

Hendricks, G, Tkaczyk, D, Lin, J and **Feeney, P.** 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies,* 1(1): 414–427. DOI: *https://doi.org/10.1162/qss_a_00022*

La Bruzzo, S, Manghi, P and **Mannocci, A.** January 2019a. OpenAIRE's DOI-Boost - Boosting Crossref for Research. In Manghi, P, Candela, L and Silvello, G (eds.), *Digital Libraries: Supporting Open Science*, pages 133–143, Cham: Springer International Publishing. ISBN 978-3-030-11226-4. DOI: *https://doi.org/10.1007/978-3-030-11226-4*

La Bruzzo, S, Manghi, P and **Mannocci, A.** December 2019b. Doiboost dataset dump. DOI: *https://doi.org/10.5281/zenodo.3559699*

Leopold, SS. May 2016. Editorial: ORCID is a Wonderful (But Not Required) Tool for Authors. *Clinical Orthopaedics & Related Research,* 474(5): 1083–1085. ISSN 0009-921X. DOI: *https://doi.org/10.1007/s11999-016-4760-0*

Manghi, P, Atzori, C, Bardi, A, Baglioni, M, Schirrwagen, J, Dimitropoulos, H, Bruzzo, SL, Foufoulas, I, Löhden, A, Bäcker, A, Mannocci, A, Horst, M, Jacewicz, P, Czerniak, A, Kiatropoulou, K, Kokogiannaki, A, Bonis, MD, Artini, M, Ottonello, E, Lempesis, A, Ioannidis, A, Manola, N and **Principe, P.** November 2020a. Openaire research graph dump. DOI: *https://doi.org/10.5281/zenodo.4279381*

Manghi, P, Atzori, C, Bonis, MD and **Bardi, A.** Jun 2020b. Entity deduplication in big data graphs for scholarly communication. *Data Technologies and Applications,* 54(4): 409–435. ISSN 25149288. DOI: *https://doi.org/10.1108/DTA-09-2019-0163*

Meadows, A, Armstrong, D, Laurel, H and **Wilkinson, LJ.** November 2017. ORCID 2017 Community Survey Report. Technical report, ORCID.

Meadows, A and **Laurel, H.** March 2019. ORCID 2018 Member Survey. Technical report, ORCID.

Meadows, A, Laurel, H, Sherlock, N and **Gruber, B.** 2019. ORCID 2019 Community Survey Report. Technical report, ORCID.

Ritter, F. February 2020. Problems with ORCID, or 10 reasons why I don't have an ORCID number. *http://www.frankritter.com/problems-with-orcid.html*.

Shillum, C, Petro, J and **Demeranville, T.** May 2021. Five years of the orcid trust program: Balancing researcher control and data quality. *https://info.orcid.org/balancing-researcher-control-and-data-integrity*.

Sinha, A, Shen, Z, Song, Y, Ma, H, Eide, D, Hsu, B-J(Paul) and **Wang, K.** 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion,* pages 243–246. ISBN 9781450334730. DOI: *https://doi.org/10.1145/2740908.2742839*

Teixeira da Silva, J. November 2020. Failure of ORCID: 57 academics named "Beatriz". *Update Dental College Journal,* 10(2): 3–5, ISSN 2307-3160, 2226-8715. DOI: *https://doi.org/10.3329/updcj.v10i2.50172*

Teixeira da Silva, J. July 2021a. Abuse of ORCID's weaknesses by authors who use paper mills. *Scientometrics*, 126(7): 6119–6125. ISSN 1588-2861. DOI: *https://doi.org/10.1007/s11192-021-03996-x*

Teixeira da Silva, J. January 2021b. ORCID: Issues and concerns about its use for academic purposes and research integrity. *Annals of Library and Information Studies,* 67: 246–250.

Wang, K, Shen, Z, Huang, C, Wu, C-H, Dong, Y and **Kanakia, A.** 2020. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies,* 1(1): 396–413. DOI: *https://doi.org/10.1162/qss_a_00021*

Wang, K, Shen, Z, Huang, C, Wu, C-H, Eide, D, Dong, Y, Qian, J, Kanakia, A, Chen, A and **Rogahn, R.** dec 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data,* 2(45): ISSN 2624-909X. DOI: *https://doi.org/10.3389/fdata.2019.00045*

Warner, S. 2010. Author identifiers in scholarly repositories. *Journal of Digital Information,* 11(1): 1–10, ISSN 13687506.

Zeng, A, Shen, Z, Zhou, J, Wu, J, Fan, Y, Wang, Y and **Stanley, HE.** Nov 2017. The science of science: From the perspective of complex systems. *Physics Reports,* 714-715: 1–73. ISSN 03701573. DOI: *https://doi.org/10.1016/j.physrep.2017.10.001*

TO CITE THIS ARTICLE:
Baglioni, M, Manghi, P, Mannocci, A and Bardi, A. 2022. We Can Make a Better Use of ORCID: Five Observed Misapplications. *Data Science Journal*, 20: 38, pp. 1–12. DOI: *https://doi.org/10.5334/dsj-2021-038*

Submitted: 21 July 2021
Accepted: 17 December 2021
Published: 31 December 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.