Will open science change authorship for good? Towards a quantitative analysis

Andrea Mannocci[a], Ornella Irrera[b,a] and Paolo Manghi[a,c]

[a]CNR-ISTI – National Research Council, Institute of Information Science and Technologies "Alessandro Faedo", Pisa, Italy
[b]Department of Information Engineering, University of Padova, Italy
[c]OpenAIRE AMKE, Athens, Greece

Abstract

Authorship of scientific articles has profoundly changed from early science until now. If once upon a time a paper was authored by a handful of authors, scientific collaborations are much more prominent on average nowadays. As authorship (and citation) is essentially the primary reward mechanism according to the traditional research evaluation frameworks, it turned to be a rather hot-button topic from which a significant portion of academic disputes stems. However, the novel Open Science practices could be an opportunity to disrupt such dynamics and diversify the credit of the different scientific contributors involved in the diverse phases of the lifecycle of the same research effort. In fact, a paper and research data (or software) contextually published could exhibit different authorship to give credit to the various contributors right where it feels most appropriate. We argue that this can be computationally analysed by taking advantage of the wealth of information in model Open Science Graphs. Such a study can pave the way to understand better the dynamics and patterns of authorship in linked literature, research data and software, and how they evolved over the years.

Keywords

authorship, open science, research literature, research data, data citation, scholarly communication

1. Introduction

While, in early science, most of the papers were authored by a handful of scientists, modern science is characterised by more extensive collaborations, and the average number of authors per article has increased across many disciplines [1, 2, 3, 4, 5]. Indeed, in some fields of science (e.g., High Energy Physics), it is not infrequent to encounter hundreds or thousands of authors co-participating in the same piece of research [6]. Such intricate collaboration patterns make it particularly hard to establish a correct relationship between contributor and scientific contribution, and hence get an accurate and fair reward during research evaluation [7, 8]. Thus, as widely known, scientific authorship tends to be a rather hot-button topic in academia as roughly one-fifth of academic disputes among authors stems from this [9].

Open Science, however, has the potential to disrupt such traditional mechanisms by injecting

18th Italian Research Conference on Digital Libraries (IRCDL 2022)
andrea.mannocci@isti.cnr.it (A. Mannocci); ornella.irrera@studenti.unipd.it (O. Irrera); paolo.manghi@isti.cnr.it (P. Manghi)
https://andremann.github.io (A. Mannocci)
0000-0002-5193-7851 (A. Mannocci); 0000-0003-2284-5699 (O. Irrera); 0000-0001-7291-3210 (P. Manghi)

into the "academic market" new kinds of "currency" for credit attribution, merit and impact assessment [10, 11]. To this end, the new practices of research data (and software) deposition and citation could be perceived as an opportunity to diversify scientific attribution and eventually give credit – right where it feels most appropriate – to the different contributors involved in the diverse phases of the lifecycle within the same research endeavour [7, 12].

In this extended abstract, we outline the perspective of using modern Open Science Graphs (OSGs) to analyse whether this is the case or not and understand if the opportunity has been seized already. Offering extensive metadata descriptions of both literature and research data records and the semantic relations among them, OSGs can be conducive to computational analysis of this phenomenon and thus study the emergence of significant patterns. In particular, it will be interesting to analyse whether and how the authors' number, composition, and order varies when moving from literature to research data and software.

It would be, for example, interesting to discover that a significantly larger amount of people is involved in the development of software and the construction of datasets rather than in the editing of the related publications. This would confirm that the current reward mechanisms are obsolete and that there is a consistent, submerged workforce contributing to research that risks being underrepresented and under-evaluated if the current practices do not change for good.

Furthermore, modifications in the composition (by shuffle or by omission) of the authors participating in a publication as opposed to the ones contributing to related research data (or software) could instead reveal other interesting aspects worth investigating. While, on the one hand, such changes in the two author lists could be contingent [13], on the other, it could be interesting to relate them to the seniority of authors in order to detect patterns revealing a possible agency behind such a choice. For example, data could suggest that senior staff members are less involved or, by any mean less interested, in participating, or getting rewarded, for data production and software development, thus confirming a bias towards the *status quo* of research assessment.

2. Data and methods

In this section, we describe the dataset we intend to use to power our study, and we provide an overview of the methodology we intend to adhere to as well as the major caveats and challenges.

2.1. Data

The study here suggested is possible thanks to the ever-increasing amount of metadata about research products of the last decade. In particular, Open Science Graphs (OSGs) can be a goldmine to this extent. OSGs are Scientific Knowledge Graphs whose intent is to improve the overall FAIRness of science by enabling open access to graph representations of metadata about people, artefacts, institutions involved in the research lifecycle, as well as the relations between such entities, to support stakeholder needs, such as discovery, reuse, reproducibility, statistics, trends, monitoring, validation, and impact assessment. The represented information may span entities such as research artefacts (e.g., publications, data, software, samples, instruments) and items of their content (e.g., statistical hypothesis tests reported in publications), research organisations, researchers, services, projects, and funders. OSGs include relationships between such entities

and sometimes formalised (semantic) concepts characterising them, such as machine-readable concept descriptions for advanced discoverability, interoperability, and reuse [14].

For this analysis, we intend to adopt the OpenAIRE Research Graph[1] [15] as our dataset of reference (hereafter, the Graph). The Graph is one of the core services provided by OpenAIRE AMKE[2], a not-for-profit legal entity operating an infrastructure that offers global services in support of Open Science scholarly workflows. The Graph aggregates metadata from 96,514 scholarly sources (as of October 2021), comprising literature, research data and software repositories, publishers, and scholarly registries, such as ORCID, ROR, re3data, OpenDOAR, Crossref, and DataCite. It thus provides a longitudinal view of the global science record by delivering an extensive collection of heterogeneous research products interconnected with the relevant semantic relations. The semantic relations conducive to this study adhere to the specification drawn in the DataCite Schema documentation[3] and are both collected from DataCite[4], EMBL-EBI, and Crossref Event Data, as well as derived from the inference full-text algorithms embedded in the OpenAIRE Graph provision workflow, and the feedback from OpenAIRE portals users.

2.2. Methods

First and foremost, a strategy to select the relevant literature and research data and software records needs to be devised. To this end, let p be a publication and d a research data (or a software) contextually produced within the same research effort (e.g., p describes a research effort to conduct a measuring campaign eventually producing the dataset d released contextually to the publication). In principle, it is possible to select all the $p \leftrightarrow d$ couples by looking at the semantics of the relations linking literature records to non-literature records within the OpenAIRE Research Graph. In our case, we plan to use the semantic *IsSupplementTo* (and inverse *IsSupplementedBy*), which is the DataCite relation type indicating that a dataset d is supplementary material for a publication p.

Once the relevant $p \leftrightarrow d$ couples have been selected, we need to proceed with the analysis of the author sets and their interpretation. Let A_p be the set of authors of a publication p, and A_d the set of authors of a connected research data (or software) d, the scope of the analysis will be threefold. Firstly, the cardinality of the two sets $|A_p|$ and $|A_d|$ can be compared to understand whether there is any difference in the workforce when moving from literature items to relevant research data. Secondly, the composition of the two author sets will be considered in order to analyse the intersection $A_p \cap A_d$ and the symmetric difference $A_p \Delta A_d = (A_p \setminus A_d) \cup (A_d \setminus A_p)$. Lastly, the ordering of the two author sets will be inspected to understand whether there are significant changes in the ranks of the same authors across the two sets and, if this is the case, which are the more frequent patterns.

Indeed, if the quantity of the available data points supports it, such analysis can be put in time perspective to analyse whether and how the trend evolved throughout the years globally and across different disciplines.

[1] OpenAIRE Research Graph, https://graph.openaire.eu
[2] OpenAIRE, https://www.openaire.eu
[3] DataCite schema, https://schema.datacite.org
[4] DataCite, https://datacite.org

2.3. Caveats

This section analyses two identified major caveats and related challenges that we have to face in this analysis. The first one is related to the inherent uncertainty of semantic relations specified among literature and research data records, while the second is related to the long-standing challenge of author disambiguation. For each, we outline the strategy we intend to follow to solve, or at least mitigate, the side effects of the caveats here described.

2.3.1. Semantic relation uncertainty

Given a $p \leftrightarrow d$ couple, the semantic expressed by the relation is defined by the user (i.e., researcher, librarian, curator) taking care of the deposition process of the research product (e.g., on Zenodo). Hence, the semantic is prone to human errors as it might not be very straightforward, which is the most appropriate one. On Zenodo, for example, the choice is drawn from a dropdown menu with scarce or limited guidance on the rationale behind the choice.

In order to mitigate this aspect, we plan to run a heuristic over $p \leftrightarrow d$ couples tied by "vanilla" relations (e.g., *Cites*, *References*) and infer the unintentionally lost relations indicating supplemented material. A viable strategy could consist in retrofitting as supplemented material relations all the *Cites* (and inverse *IsCitedBy*) and *References* (and inverse *IsReferencedBy*) relations when the author sets share at least an author and the year indicates that the two records are contextual (e.g., within six months apart).

A possible generalisation of the heuristic above would rely on multiple metadata fields such as the date of publication, the title and the author list itself to create a feature vector describing research outputs. Then the distance between such vectors representing publication records and non-literature records related with the proper semantic would allow us to define a confidence interval of similarity which characterises literature and related non-literature records. New supplement semantics can be inferred relying on such confidence interval: if the similarity between two feature vectors tied by "vanilla" relations lays within the interval, then the semantics has been probably misassigned, and thus it can be retrofitted as *IsSupplementedBy* (or *IsSupplementTo*, depending on the direction).

2.3.2. Author names disambiguation

Author disambiguation is essential to make the set of authors A_d of the dataset (or software) d and the set of authors A_p of the publication p comparable.

The metadata definition of an author a who contributed both in the publication p and in the supplement dataset d, may not be the same in p and d respectively. In this case, if the intersection $A_p \cap A_d$ is computed, the author a will not belong to the intersection because there are different definitions of a in A_p and A_d. In this context, disambiguation is crucial to correctly detect that the author a is the same in A_p and A_d despite multiple definitions.

Consider, for example, the publication with the DOI:https://doi.org/10.1186/s12865-015-0113-0 (*Immune cell subsets and their gene expression profiles from human PBMC isolated by Vacutainer Cell Preparation Tube (CPT^{TM}) and standard density gradient*). One of the datasets it is supplemented by is https://doi.org/10.6084/m9.figshare.c.3600443_d4.v1 (*Additional file 4: Table S4. of*

Immune cell subsets and their gene expression profiles from human PBMC isolated by Vacutainer Cell Preparation Tube (CPT™) and standard density gradient). The lists of authors A_p and A_d are:

$$A_p = \{\text{Corkum, Christopher } \mathbf{P.}\text{; Ings, Danielle } \mathbf{P.}\text{; Burgess, Christopher; Karwowska, Sylwia; Kroll, Werner; Michalak, Tomasz } \mathbf{I.}\}$$
$$A_d = \{\text{Corkum, Christopher; Ings, Danielle; Burgess, Christopher; Karwowska, Sylwia; Kroll, Werner; Michalak, Tomasz}\}$$

If the lists of authors are analysed, there are three authors which co-occur both in A_p and in A_d (*Burgess, Christopher*, *Karwowska, Sylwia* and *Kroll, Werner*). The three remaining authors differ in how their *name* is laid out: in A_p in fact the first names are followed by another initial (highlighted in boldface), while in A_d they do not; without finer author disambiguation strategies (e.g., plain string match) they would be considered different authors.

To address the author disambiguation problem, we can rely on the deduplication framework of OpenAIRE [16, 17] and the distance metrics it provides to compute the distance between single authors and lists. It is worth noting that, in contrast to the standard deduplication task (i.e., establish the equivalence of alike research products), we are comparing lists of authors belonging to research outputs different in kind (i.e., literature with non-literature); these lists may not necessarily contain the same authors; hence, the methods provided for the deduplication need to be customised according to our needs.

Acknowledgments

This work was co-funded by the European Commission H2020 project OpenAIRE-Nexus (grant number: 101017452).

References

[1] B. Cronin, Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?, Journal of the American Society for Information Science and Technology 52 (2001) 558–569. doi:10.1002/asi.1097.

[2] J. D. Wren, K. Z. Kozak, K. R. Johnson, S. J. Deakyne, L. M. Schilling, R. P. Dellavalle, The write position: A survey of perceived contributions to papers based on byline position and number of authors, EMBO reports 8 (2007) 988–991.

[3] C. Baethge, Publish together or perish: the increasing number of authors per article in academic journals is the consequence of a changing scientific culture. some researchers define authorship quite loosely, Deutsches Arzteblatt International 105 (2008) 380.

[4] J. M. Fernandes, M. P. Monteiro, Evolution in the number of authors of computer science publications, Scientometrics 110 (2017) 529–539.

[5] T. F. Frandsen, J. Nicolaisen, What is in a name? Credit assignment practices in different disciplines, Journal of Informetrics 4 (2010) 608–617. doi:10.1016/j.joi.2010.06.010.

[6] G. Aad, et al., Atlas collaboration, PHYSICAL REVIEW D Phys Rev D 85 (2012) 012003.

[7] A. Brand, L. Allen, M. Altman, M. Hlava, J. Scott, Beyond authorship: Attribution, contribution, collaboration, and credit, Learned Publishing 28 (2015). doi:10.1087/20150211.

[8] N. A. Vasilevsky, M. Hosseini, S. Teplitzky, V. Ilik, E. Mohammadi, J. Schneider, B. Kern, J. Colomb, S. C. Edmunds, K. Gutzman, D. S. Himmelstein, M. White, B. Smith, L. O'Keefe, M. Haendel, K. L. Holmes, Is authorship sufficient for today's collaborative research? a call for contributor roles, Accountability in Research 28 (2021) 23–43. doi:10.1080/08989621.2020.1779591.

[9] A. Dance, Authorship: Who's on first?, Nature 489 (2012) 591–593. doi:10.1038/nj7417-591a.

[10] H. Mooney, M. Newton, The Anatomy of a Data Citation: Discovery, Reuse, and Credit, Journal of Librarianship and Scholarly Communication 1 (2012) eP1035. doi:10.7710/2162-3309.1035.

[11] G. Silvello, Theory and Practice of Data Citation, Journal of the Association for Information Science and Technology 69 (2018) 6–20. doi:10.1002/asi.23917. arXiv:1706.07976.

[12] B. E. Bierer, M. Crosas, H. H. Pierce, Data Authorship as an Incentive to Data Sharing, New England Journal of Medicine 376 (2017) 1684–1687. doi:10.1056/NEJMsb1616595.

[13] M. Kosmulski, The order in the lists of authors in multi-author papers revisited, Journal of Informetrics 6 (2012) 639–644. doi:10.1016/j.joi.2012.06.006.

[14] A. Aryani, M. Fenner, P. Manghi, A. Mannocci, M. Stocker, Open Science Graphs Must Interoperate!, in: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Springer, 2020, pp. 195–206.

[15] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Löhden, A. Bäcker, A. Mannocci, M. Horst, P. Jacewicz, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, N. Manola, P. Principe, Openaire research graph dump, 2020. URL: https://doi.org/10.5281/zenodo.4279381. doi:10.5281/zenodo.4279381.

[16] P. Manghi, M. Mikulicic, C. Atzori, De-duplication of aggregation authority files, International Journal of Metadata, Semantics and Ontologies 7 (2012) 114–130.

[17] P. Manghi, C. Atzori, M. De Bonis, A. Bardi, Entity deduplication in big data graphs for scholarly communication, Data Technologies and Applications (2020).