RESEARCH ARTICLE

Open bibliographic data and the Italian National Scientific Qualification: Measuring coverage of academic fields

Federica Bologna[1], Angelo Di Iorio[2], Silvio Peroni[3,4], and Francesco Poggi[5,6]

[1]Bowers College of Computing and Information Science, Cornell University, Ithaca, NY
[2]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
[3]Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
[4]Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
[5]Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia, Italy
[6]Institute of Cognitive Sciences and Technologies, Italian National Research Council (CNR), Rome, Italy

Keywords: bibliographic data, open bibliographic repositories, open data

ABSTRACT

The importance of open bibliographic repositories is widely accepted by the scientific community. For evaluation processes, however, there is still some skepticism: Even if large repositories of open access articles and free publication indexes exist and are continuously growing, assessment procedures still rely on proprietary databases, mainly due to the richness of the data available in these proprietary databases and the services provided by the companies they are offered by. This paper investigates the status of open bibliographic data of three of the most used open resources, namely Microsoft Academic Graph, Crossref, and OpenAIRE, evaluating their potentialities as substitutes of proprietary databases for academic evaluation processes. We focused on the Italian National Scientific Qualification (NSQ), the Italian process for university professor qualification, which uses data from commercial indexes, and investigated similarities and differences between research areas, disciplines, and application roles. The main conclusion is that open data sets are ready to be used for some disciplines, among them mathematics, natural sciences, economics, and statistics, even if there is still room for improvement; but there is still a large gap to fill in others—such as history, philosophy, pedagogy, and psychology—and greater effort is required from researchers and institutions.

QSS

an open access journal

Citation: Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2022). Open bibliographic data and the Italian National Scientific Qualification: Measuring coverage of academic fields. *Quantitative Science Studies*, *3*(3), 512–528. https://doi.org/10.1162/qss_a_00203

DOI:
https://doi.org/10.1162/qss_a_00203

Peer Review:
https://publons.com/publon/10.1162/qss_a_00203

Received: 5 October 2021
Accepted: 5 July 2022

Corresponding Author:
Silvio Peroni
silvio.peroni@unibo.it

Handling Editor:
Staša Milojević

The MIT Press

1. INTRODUCTION

The relevance of open bibliographic data sets is continually increasing, under the pressure of internationally coordinated efforts such as the Initiative for Open Citations (I4OC, https://i4oc.org) and the Initiative for Open Abstracts (I4OA, https://i4oa.org). These data sets have not only changed the way scholars search for literature but have also enabled the advancement of the scientometrics and bibliometrics fields. The greater availability of open data has made it possible for researchers to carry out groundbreaking studies on research practices, academic literature, and academic institutions (Bedogni, Cabri et al., 2021; Chudlarský & Dvořák, 2020;

Di Iorio, Peroni, & Poggi, 2019; Huang, Neylon et al., 2020; Martín-Martín, Thelwall et al., 2021; Peroni, Ciancarini et al., 2020; Zhu, Yan et al., 2020).

One of the issues still open in this context is whether or not the open bibliographic data sets are ready to substitute for commercial ones in the research evaluation processes. In fact, bibliometrics are being widely used, both by private and governmental agencies, to evaluate the scientific performance of institutions, journals, groups, and scholars. For example, several countries employ evaluation procedures that combine bibliometrics and peer review, such as Excellence in Research for Australia (ERA), the British Research Excellence Framework (REF), and the Valutazione della Qualità della Ricerca (VQR) and the National Scientific Qualification (NSQ) in Italy. But these processes still rely on commercial data sets, such as Scopus and Web of Science (WoS).

Therefore, investigating the availability of open data in the context of national scholarly assessments is of utmost relevance. The first step in that direction is to study the coverage of the publications in these data sets. Previous works have mainly used (the publications found in) proprietary data sets as benchmarks against which to compare (the publications found in) open data sets (Harzing, 2019; Huang et al., 2020; Martín-Martín et al., 2021; Singh, Singh et al., 2021; Visser, van Eck, & Waltman, 2021).

This paper adds a piece to the puzzle. It sheds light on the differences between the open data sets when used for evaluating research productivity in different contexts and disciplines.

Specifically, we ground this study in the Italian National Scientific Qualification (NSQ). The NSQ is a nationwide research assessment exercise that establishes whether a scholar can apply to professorial academic positions as associate professor and full professor. Applications are organized according to a governmentally defined taxonomy of 190 Recruitment Fields (RF) divided into 14 Scientific Areas (SA). The disciplines are divided into two categories, citation-based (CDs) and noncitation-based (NDs), depending on the use of citations for the evaluation. In the NSQ nomenclature, CDs and NDs are actually tagged as "bibliometric" and "nonbibliometric" disciplines, respectively, but we prefer not to use this terminology, which might be misleading here. Note also that SAs can include either CDs or NDs or a mix of them.

Our work moves away from such a distinction and digs into open data sets: Do these data sets show the same structure and behavior for CDs and NDs? Which are the most relevant differences? Where do these differences derive from? And more: Are there significant differences between the coverage for candidates as associate professor or full professor? And between the coverage for candidates in different recruitment fields?

To answer these questions we analyzed Microsoft Academic Graph, Crossref, and OpenAIRE and compared the coverage of publications for a wide range of disciplines and candidates. This work in fact is an extension of a particular aspect (i.e., the coverage of publications in open data sets) of our previous study, which was limited to some recruitment fields only (Bologna, Di Iorio et al., 2021c). Here, we include all disciplines of the NSQ and analyze the publications of all candidates in the 2016, 2017, and 2018 terms of the NSQ. Overall, we consider 2,353,872 publications for 58,335 candidates in 190 Recruitment Fields. We also investigate whether coverage improves when combining the three data sets, collecting evidence of the *effectiveness* of these data sets for CDs, much more than NDs, and some peculiarities of each data set.

Note that the term *effectiveness* is used here to indicate the capability of providing data for the evaluation process, not the fairness and efficacy of the process itself. Our goal is not to assess the NSQ or similar processes but rather to understand if these processes could be built on and improved by using open data only. Having good coverage is a necessary but not sufficient condition for such a transition. The goal here is to assess the data collection process, understanding how much information is available today, from which sources, and for which disciplines.

2. RELATED WORK

2.1. Previous Work on Coverage

Numerous studies analyzing coverage of bibliographic data sets have been published since these became widely used in the scientific community. These studies differ in the methods and measurements they use to analyze and compare coverage, in the sets of data sets they compare, and in the type of document they focus their comparison on.

Some studies focus on a small sample of documents belonging to a specific discipline or single author (Harzing, 2019). Others involve millions of documents from a wide range of disciplines (Martín-Martín et al., 2021). Some works employ a straightforward method, obtaining the complete list of documents contained in each data set, matching the documents across sources and measuring the overlap (Visser et al., 2021). Other works use alternative methods, because not all bibliographic data sets offer free access to their data, by comparing documents' citation lists (Martín-Martín, Orduna-Malea et al., 2018; Martín-Martín et al., 2021).

Of these previous analyses on coverage in bibliographic data sets, a great number focus on WoS and Scopus, and use them as benchmarks to evaluate other data sets' coverage. Some studies draw comparisons in coverage between WoS and Scopus (Mongeon & Paul-Hus, 2016). Some compare their coverage to that of Dimensions (Orduña-Malea & Delgado-López-Cózar, 2018; Singh et al., 2021; Thelwall, 2018), of Google Scholar (Delgado López-Cózar, Orduña-Malea, & Martín-Martín, 2019; Martín-Martín et al., 2018), of Microsoft Academic (Huang et al., 2020; Hug & Brändle, 2017), and of both Google Scholar and Microsoft Academic (Harzing, 2016; Harzing & Alakangas, 2017a, 2017b). Others compare multiple data sets against each other (Harzing, 2019; Martín-Martín et al., 2021; Visser et al., 2021).

2.2. Bibliographic Data Sets

For more than a decade, WoS (introduced online in 1997) (Birkle, Pendlebury et al., 2020) and Scopus (launched in 2004) (Baas, Schotten et al., 2020) have been the only available options to conduct large-scale bibliometric analyses. These two commercial subscription-based bibliographic data sources provide metadata on scientific documents and on citation links between these documents. Google Scholar (Van Noorden, 2014), a free bibliographic search engine, was launched a week after Scopus. However, it did not provide bulk access to its data.

The introduction of new open bibliographic data sources changed this trend (Martín-Martín et al., 2021). In 2013, Crossref, a nonprofit membership association between publishers, made all its metadata available to the public via a REST API, and in 2017, thanks to the Initiative for Open Citations (I4OC; https://i4oc.org), millions of citation links between documents have also been made openly available. In 2014, OpenAIRE (Manghi, Bolikowski et al., 2012), an EU-funded infrastructure to share bibliographic data across institutions, released free API access to its data. In 2016, Microsoft Academic (Wang, Shen et al., 2020) provided a scholarly search engine and bulk access to its data via an API, both without charge. This study focuses on these mentioned open-access data sets.

A number of other bibliographic data sources have not been considered in this study, for various reasons:

- Dimensions (Herzog, Hook, & Konkiel, 2020) requires payment of a fee for bulk access.
- CiteSeerX (Wu, Kim, & Giles, 2019) indexes documents in the public web, but not those behind paywalls.
- ResearchGate (https://www.researchgate.net/) does not provide a tool to extract data in bulk.
- Lens.org provides free bulk-access to noncommercial projects only for a limited time and at a limited access rate.
- Regional and subject-specific data sets do not offer multidisciplinary coverage by design; hence, they are not comparable to the other sources considered in this study.

2.3. The Italian National Scientific Qualification (NSQ)

In 2011, Italian Law of December 30, 2010 n.240 (L. 240/2010, 2011) implemented the NSQ, a nationwide research assessment exercise that attests the scientific maturity of scholars. The law made it mandatory to pass the NSQ to apply to academic positions. The NSQ consists of two distinct qualification processes, one for the academic position of full professor (FP), and one for that of associate professor (AP). Passing the NSQ does not grant a tenure position. It is each university's responsibility to create new positions and hire scholars according to financial and administrative requirements.

Moreover, the Ministerial Decree of June 14, 2012 (D. L. 2012, 2012) defined a taxonomy of 184 (extended to 190 a few years later) Recruitment Fields (RF) divided into groups and sorted into 14 different Scientific Areas (SA). SAs correspond to vast academic disciplines, whereas RFs correspond to specific scientific fields of study. Each scholar is assigned to a specific RF, which belongs to a single SA. In the taxonomy, RFs are identified by an alphanumeric code in the form AA/GF. AA is a number indicating the SA, ranging from 1 to 14. G is a single letter identifying the group of RFs. F is a digit indicating the RF. For instance, Neurology's code is 06/D5, where 06 indicates the SA Medicine and D indicates the group Specialized Clinical Medicine (D. L. 2012, 2012). When applying for the NSQ, scholars can choose to be evaluated for more RFs at a time. Because each RF has its own assessment rules, the candidate may pass the qualification in some fields but not in others.

The NSQ divides academic disciplines into two categories: citation-based disciplines (CDs) and noncitation-based disciplines (NDs). This division affects only the metrics used for assessing the candidates of that discipline in the first part of the process. Candidates applying to CDs are evaluated using:

- CD_M1: their number of journal papers;
- CD_M2: the total number of citations received; and
- CD_M3: their *h*-index.

Candidates applying to NDs are evaluated using:

- ND_M1: their number of journal papers and book chapters;

- ND_M2: their number of papers published on Class A journals[1]; and
- ND_M3: their number of published books.

To apply to the NSQ, candidates must submit a curriculum vitae (CV) with detailed information about their research accomplishments. Then, NSQ assessment is organized in two steps. In the first step of the evaluation, candidates' metrics are expected to exceed two of the three thresholds in their RF. Successively, the candidate's maturity is evaluated based on their CV. The aforementioned metrics are computed for each candidate, taking into consideration only publications that are less than 15 years old for candidates for the role of FP and 10 years old for candidates for the role of AP. This process utilizes data retrieved from Scopus and WoS and is conducted by the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR). ANVUR also sets thresholds for each metric by RF. Normalization based on the scholars' scientific age (the number of years since their first publication) is used to compute most of the metrics.

As shown in Table 1, citation-based disciplines are predominantly either STEM-based (science, technology, engineering, and mathematics) or medicine-based RFs, such as all the RFs in the first nine SAs (01–09), with the exception of the RFs 08/C1, 08/D1, 08/E1, 08/E2, 08/F1, which are considered NDs, and the four RFs in psychology (11/E), which are considered CDs. Noncitation-based disciplines are predominantly HASS-based (humanities, arts and social sciences) RFs, such as the last five SAs (10–14), with the exceptions just described. The reason for this division is that, according to ANVUR, reliable and sufficiently complete citation databases exist for CDs, but they do not for NDs.

For the purposes of this study, we take into consideration the second session of the NSQ which took place from 2016 to 2018, with one term in 2016, two terms in 2017, and two terms in 2018.

To keep the process as transparent as possible, the full CV of each candidate is publicly posted (in PDF) on the NSQ website and is accompanied by the full scripts of the judgements of the evaluation committee (also called a commission) of each RF—composed of five full professors responsible for assessing applicants for AP and FP. We use the metadata contained in these CVs to investigate and compare coverage of the candidates' publications by Microsoft Academic Graph, Crossref, and OpenAIRE across disciplines.

3. METHODS AND MATERIALS

This section introduces all the methods and materials used for our study. The data and software developed for this work are available in Bologna, Di Iorio et al. (2021b) and in the project GitHub repository: https://github.com/sosgang/coverage_asn.

3.1. Data

For the purposes of this study, we considered all candidates that participated in the 2016, 2017, and 2018 sessions of the NSQ. From each candidate's CV, we extracted all the available publications metadata as described in the following section. The specifics of the data set obtained from these CVs are available in Table 2.

[1] The top-rated journals according to official classification provided by ANVUR, available at https://www.anvur.it/attivita/classificazione-delle-riviste/classificazione-delle-riviste-ai-fini-dellabilitazione-scientifica-nazionale/elenchi-di-riviste-scientifiche-e-di-classe-a/.

Table 1. Identification of the RFs included in the various SAs that are defined either as citation-based disciplines (CDs) or noncitation-based disciplines (NDs) according to ANVUR

	Citation-based disciplines (CDs)	Noncitation-based disciplines (NDs)
SA01	all RFs	no RFs
SA02	all RFs	no RFs
SA03	all RFs	no RFs
SA04	all RFs	no RFs
SA05	all RFs	no RFs
SA06	all RFs	no RFs
SA07	all RFs	no RFs
SA08	RFs 08/A1, 08/A2, 08/A3, 08/A4, 08/B1, 08/B2, 08/B3	RFs 08/C1, 08/D1, 08/E1, 08/E2, 08/F1
SA09	all RFs	no RFs
SA10	no RFs	all RFs
SA11	RFs 11/E1, 11/E2, 11/E3, 11/E4	RFs 11/A1, 11/A2, 11/A3, 11/A4, 11/A5, 11/B1, 11/C1, 11/C2, 11/C3, 11/C4, 11/C5, 11/D1
SA12	no RFs	all RFs
SA13	no RFs	all RFs
SA14	no RFs	all RFs

Table 2. Number of applications and publications considered in the study

Unique applications considered	58,364
Unique applications with relevant publishing data, of which 41,668 applications to CDs and 16,668 applications to NDs	58,335
Missing CVs	9
CVs without relevant publishing data	19
Publications with metadata, of which 1,951,515 publications in CDs 402,357 publications in NDs	2,353,872
Publications without any metadata	17
Publications with parsing issues	18,437
Publications without enough metadata	2,384

3.2. Sources

We considered three open access sources. The first, Microsoft Academic Graph (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/) (Wang et al., 2020), referred to as MAG here, results from the efforts of the Microsoft Academic Search (MAS) project. This data set is updated biweekly and is distributed under an open data license for research and commercial applications. We use a copy of MAG created and made available by the Internet Archive in January 2020 (Microsoft Academic, 2020).

The second, OpenAIRE Graph (https://www.openaire.eu/) (Manghi et al., 2012), referred to as OA here, includes information about objects of the scholarly communication life cycle (publications, research data, research software, projects, organizations, etc.) and semantic links among them. It is created bimonthly and is accessible for scholarly communication and research analytics. We use the dump that OpenAIRE released on Zenodo in April 2021 (Manghi, Atzori et al., 2021).

The third, Crossref (https://www.crossref.org/) (Hendricks et al., 2020), referred to as CR here, was born as a nonprofit membership association among publishers to promote collaboration to speed research and innovation. The data set is fully curated and governed by the members. We use the dump released in January 2021 (Crossref, 2021).

3.3. Dumps Processing and Database Creation

To efficiently query MAG, OA, and CR for each publication of each candidate in each quick succession, we create a database containing all the bibliographic metadata present in each data set dump.

First, we download and process each dump. We take each publication in the dump, select the metadata of interest for our analysis (author, title, year, DOI, and any MAG-specific identifier) and store it in a JSON file, thus transforming the three dumps into three large JSON files.

Second, we set up the MongoDB database. MAG's, OA's and CR's JSON files are imported as separate collections into a MongoDB database. We then create indexes in each collection to improve query efficiency. We set a compound index, combining a text index on the field "title" of the publications and an ascending index on the field "year," and an ascending index on the field "doi," In MAG's collection we set two other ascending indexes, one on the field "id.mag," containing MAG's publication identifier, and one on field "authors.id.mag," containing MAG's author identifiers.

We then proceed to query the database with the candidates' publication metadata to collect coverage information.

3.4. Querying the Database and Collecting Coverage Information

To obtain coverage information on the candidates' publications, we first extract all bibliographic metadata (e.g., the title, authors and DOIs of the publications) from the CVs the candidates submitted when applying to the NSQ (in the 2016, 2017, and 2018 sessions). The CVs were available in PDF and have been converted into a pure textual format to extract structured information (such as the title, authors, and DOIs of the publications) to be stored in JSON.

We obtain a list of publications for each candidate with their relevant metadata (*title*, *year*, *doi*, *authors*). Then, for each candidate, we use this metadata to query MAG, OA, and CR collections in the database to find each publication in each collection. We query the database either by *doi*, if present among the publication's metadata, or by *year* and *title*. In OA, when we find a publication, we add a "coverage marker" to the publication's metadata to signal that said publication is present in the collection. The same is done for CR collection. In MAG, when we find a publication in the database, we collect and store the Paper Id (*pId*), which identifies the publication, and the *Author Id* of the candidate, which identifies the author. The *pId* acts as "coverage marker" for MAG collection. Because in MAG each author can be assigned multiple *Author IDs*, we retrieve one for each publication we find and keep only the unique IDs. Then, we query MAG by each *Author Id*, retrieve all publications associated with that id, and compare them to our list of publications. We do so to catch publications that are present in MAG's collection but that we were not able to find by querying the collection using the publication's metadata in the CV.

Lastly, for each candidate we calculate the coverage of their publications by MAG, OA, CR, or the combination of the three. We count the number of publications in the candidate CV for which we have either the *doi*, or *year* and *title* information. We then count the number of publications that have a *pId*, the number of publications that have OA's coverage marker, and the number of publications that have CR's coverage marker, and the number of publications that have either the *pId* or one of these markers. We also calculate the percentage of found publications in MAG, OA, CR, and the combination of the three for each candidate.

The results of this procedure are presented in the following section.

4. RESULTS

Overall, we compare the coverage of 2,353,872 publications by 58,335 candidates across 190 RFs. In the following sections we refer to Crossref, OpenAIRE, and Microsoft Academic Graph data sets as CR, OA, and MAG respectively.

4.1. Overall Coverage by Data Set

Figure 1 shows the overall coverage of candidates' publications in each of the three data sets, as well as their combination. Each data point represents the percentage of publications found in the data set of interest for a single candidate. Percentages are calculated by taking the number of found publications and dividing this number by the total number of publications in the CV that have relevant publishing metadata. In this diagram, we consider all the candidates who took part in the 2016–2018 NSQ sessions, regardless of their RF.

Overall, all three data sets have very good coverage of candidates' publications. However, CR's distribution presents a slightly different behavior: CR's minimum is distinctly lower than that of the other data sets; its first and second quartiles are more spread out and there are no outliers. This indicates that there are more data points in the lower half of CR's distribution than in the other data sets' distributions. We hypothesize that this phenomenon is due to how metadata is collected to build these data sets and what methods are used for this purpose. MAG is built using web crawling and OA by joining the metadata shared by a network of EU institutions and libraries (including Italian libraries, that collect bibliographic data about all the Italian researchers participating in the NSQ), whereas CR is built by its members (i.e., the publishers) and is predominantly DOI-based. Therefore, publications that are not assigned a DOI could be found in MAG and OA, but not in CR.

4.2. Coverage by Citation-Based and Noncitation-Based Disciplines

Figure 2 displays the coverage of candidates' publications by data sets and field category. As we expected, there is a sharp difference in coverage between candidates who applied to CDs and those who applied to NDs. CDs' distributions are more in line with the overall results shown than NDs' distributions. This phenomenon is caused by the overwhelmingly higher number of candidates in CDs than in NDs—41,668 applications to CDs and 16,668 applications to NDs—and the disproportionately higher number of publications in CDs than NDs—1,951,515 publications in CDs, making up 82% of the overall number of publications, and 402,357 publications in NDs. Indeed, the median number of publications per candidate in CDs is 35, whereas in NDs it is 20. As a result, CDs' coverage weighs more heavily in the overall results.

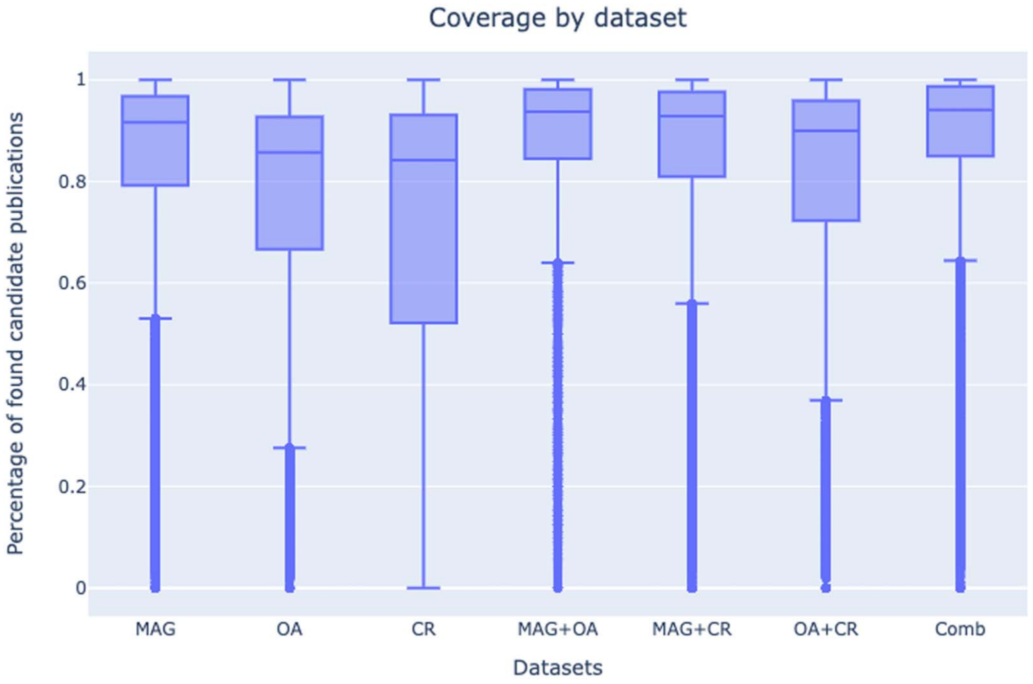

Figure 1. Overall coverage of candidates' publications in each of the three data sets—Microsoft Academic Graph (MAG), OpenAIRE (OA), and Crossref (CR)—as well as their multiple combinations.

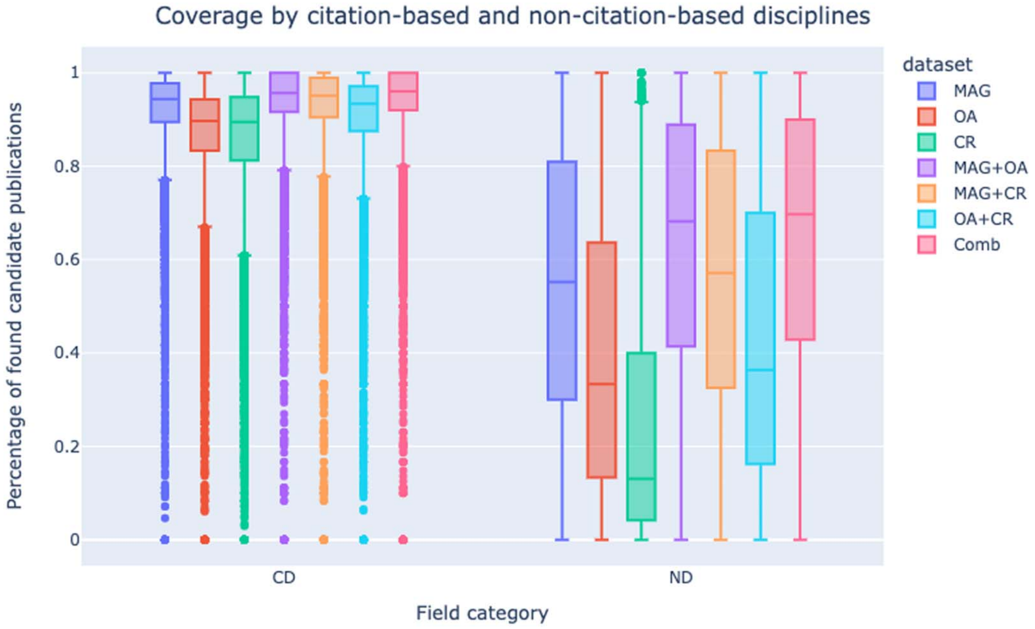

Figure 2. The coverage of candidates' publications by data set and field category.

Once again, CRs' results are worse than those of MAG and OA, both for CDs and NDs. However, CR's coverage is particularly low in NDs, as its median percentage of found candidate's publications is 13%.

Furthermore, there is little difference in coverage of CDs between MAG and the combination of the three data sets for CDs, indicating that MAG contains almost all of the open publication data for those disciplines. In addition, there is virtually no difference in the coverage of CDs between the combination of MAG and OA and the combination of all three data sets, and between MAG and the combination of MAG and CR. This shows that OA contributes almost all of the additional data not covered in MAG. We find an almost identical pattern in the coverage of NDs. Given this evidence and the demonstrated poor coverage of NDs by CR, we hypothesize that the added data comes from OA.

4.3. Coverage by Data Sets and Roles

The diagram in Figure 3 shows the coverage of candidates' publications by data set and the academic role the candidate applied to (AP and FP). There is not much difference in coverage between candidates applying for AP and those applying for FP. After all, in the NSQ, candidates are evaluated on the publications published in the last 15 years the last 10 years for FP and AP respectively. Therefore, any older publication—that could have affected the results, weighting more for FP who had published more—is not included in the CVs and not considered in this study.

4.4. Coverage by Data Sets and Scientific Areas

Figure 4 and Figure 5 present the coverage of candidates' publications by data set and SA (Scientific Area). SAs solely constituted by CDs—1 through 7 and 9—all show great coverage results, with tight quartiles and median values that are above 0.85.

SAs solely constituted by NDs—10 and 12 through 14—show the worst coverage results, with the exception of SA 13 *Economics and Statistics*. Indeed, 13 presents higher values than 10, 11, and 14, probably caused by its proximity in topic to other SAs only consisting of CDs. However, it also has wider first quartiles than SAs only consisting of CDs, indicating the greater presence of low data points in its distribution.

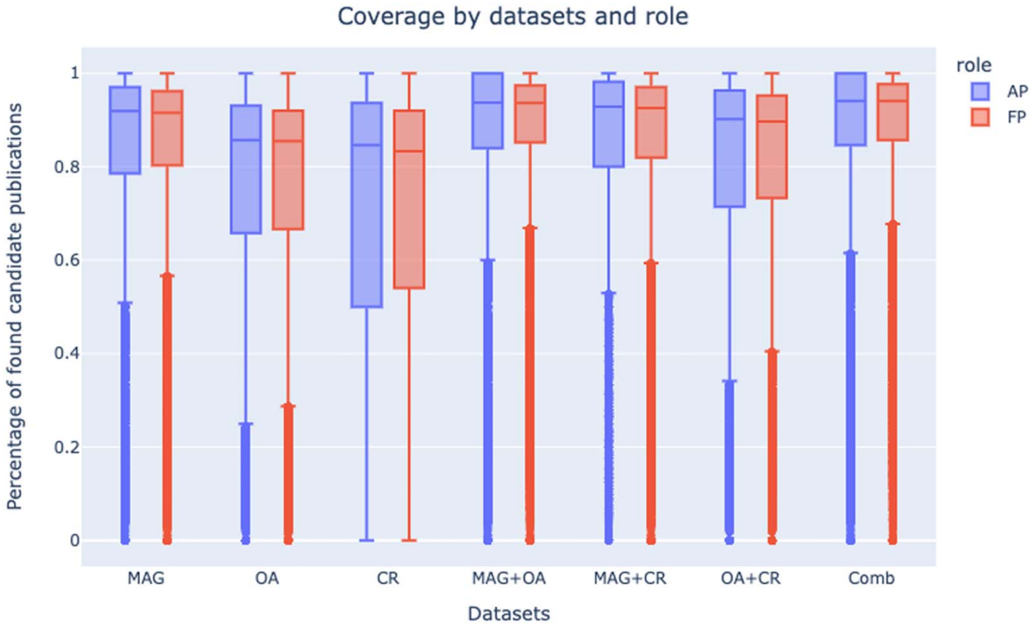

Figure 3. The coverage of candidates' publications by data set and the academic role the candidate applied to in the NSQ.

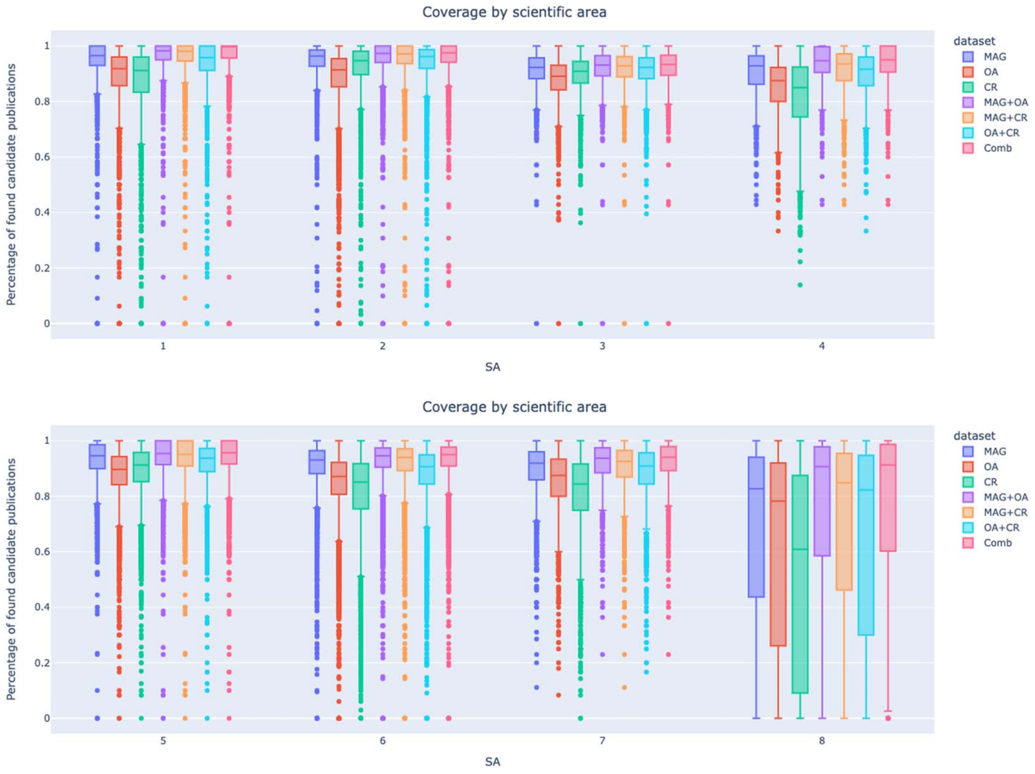

Figure 4. The coverage of candidates' publications by data set and SA 1–8.

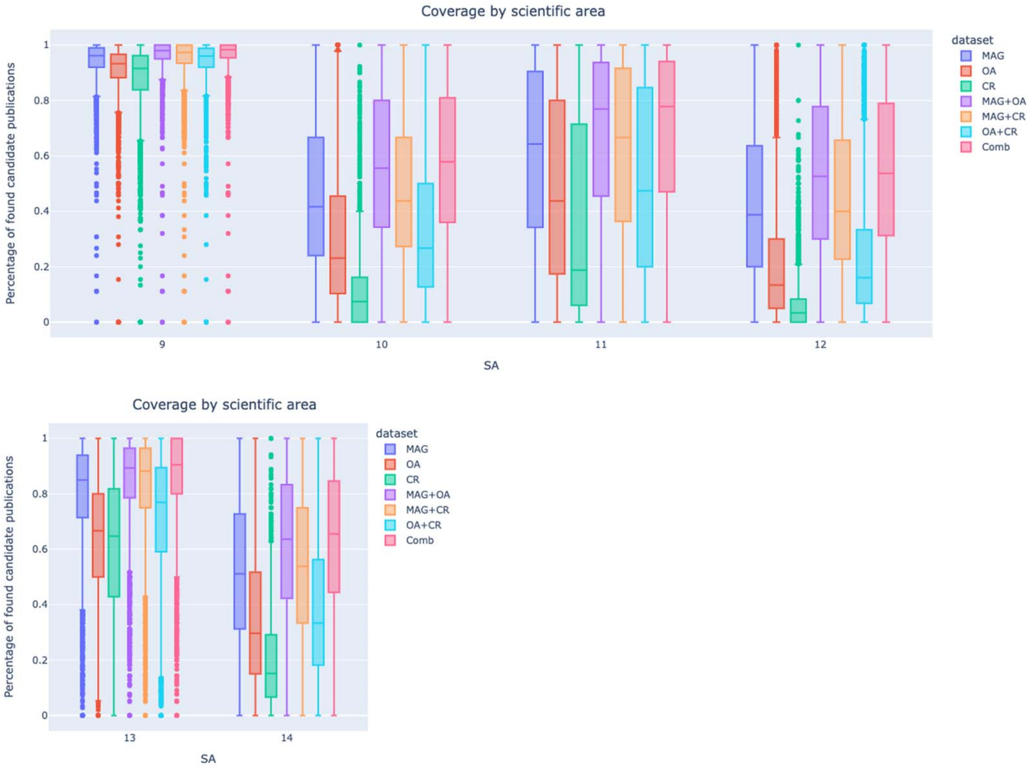

Figure 5. The coverage of candidates' publications by data set and SA 9–14.

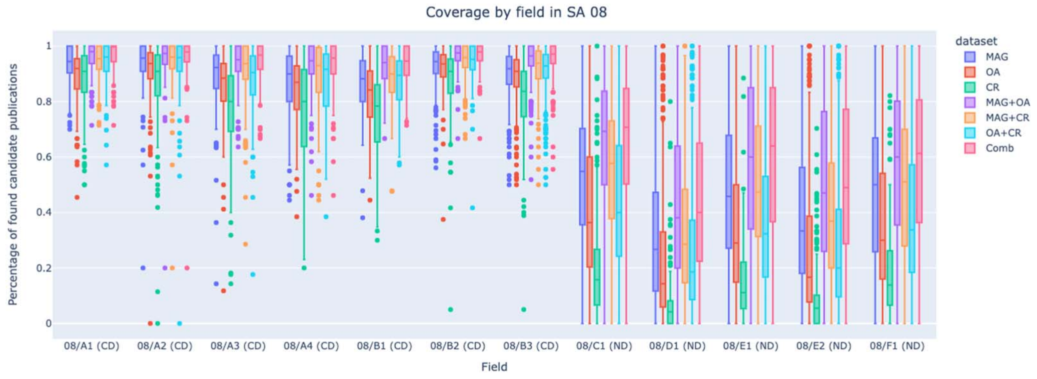

Figure 6. The coverage of candidates' publications by data set and RF in SA 8 (*Civil Engineering and Architecture*).

SAs 8 and 11 are constituted of both CDs and NDs and their distributions are characterized by wide quartiles, indicating the joint presence of high and low data points. It is also worth pointing out CR's poor coverage of 10, 11, 12, and 14, with median values below 20%.

4.5. Coverage by Field in the Scientific Area with CDs and NDs

The diagrams in Figure 6 and Figure 7 present coverage of candidates' publications by data set and RF in mixed SAs (i.e., constituted by CDs and NDs): 8, *Civil Engineering and Architecture*; and 11, *History, Philosophy, Pedagogy and Psychology*. When focusing on the individual RFs inside mixed SAs, the difference in coverage between CDs and NDs clearly emerges. CDs are the RFs with higher values, whereas NDs are the RFs with lower values. 08/D1, *Architectural Design*, presents the worst coverage percentages with median values sharply below 50% for all three data sets, as well as their combination.

SA 13, *Economics and Statistics*, is not officially considered a mixed SA by ANVUR—indeed, it is entirely composed by NDs. However, as shown in Figure 8, it presents similar behaviors and characteristics to mixed SAs: some RFs inside SA 13 have high coverage percentages—13/A, *Economics*, and 13/D, *Statistics and Mathematical Methods for Decisions*—while others have low coverage percentages—13/B, *Business Administration and Management*, and 13C, *Economic History*. This diagram casts a light on how publishing and

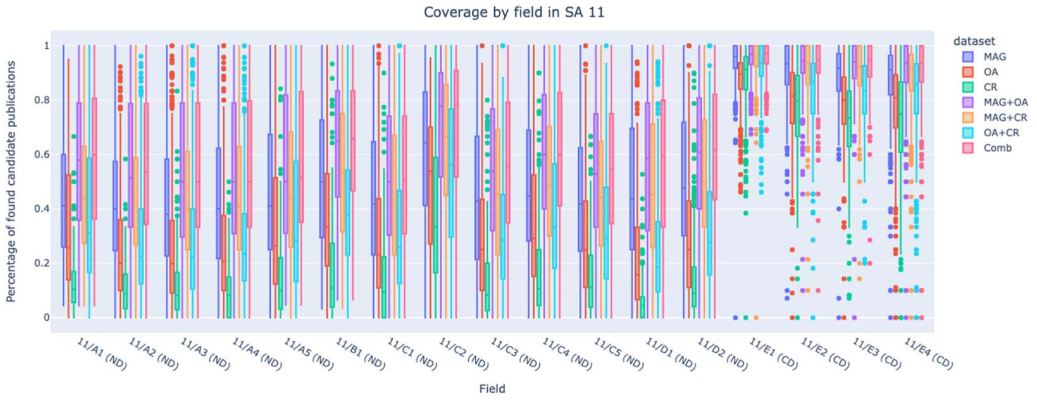

Figure 7. The coverage of candidates' publications by data set and RF in SA 11 (*History, Philosophy, Pedagogy and Psychology*).

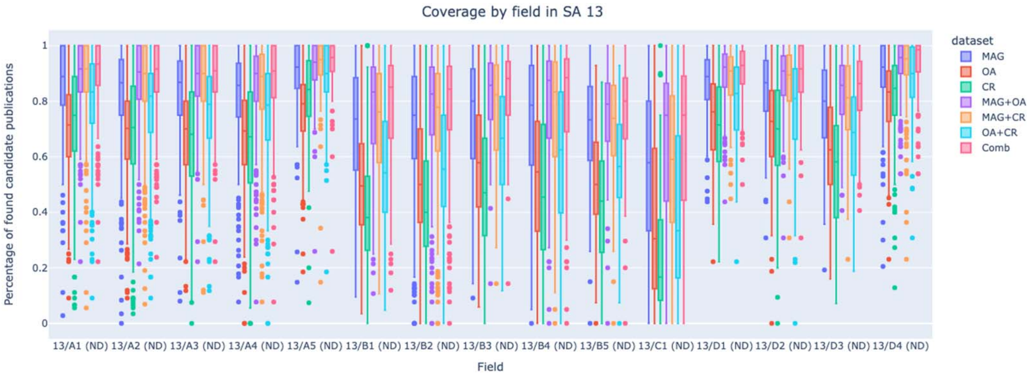

Figure 8. The coverage of candidates' publications by data set and RF in SA 13 (*Economics and Statistics*).

editorial traditions connected to specific topics and disciplines greatly influence the coverage of publications in such fields by open data sets.

5. DISCUSSION AND CONCLUSIONS

The main aim of our work was to check the coverage, in terms of bibliographic entities, of three of the most prominent and discipline-agnostic open data sets considering a collection of publications mediated by a particular use-case scenario: the Italian National Scientific Qualification (NSQ). We have shown that the coverage is pretty good, in particular, for those RFs for which citation data are considered in the evaluation of the NSQ, namely, those belonging to SAs from 1 to 7, part of SA 8, SA 9, and part of SA 11. This is reasonable, considering how the NSQ works. Indeed, candidates applying for a CD recruitment field can present their publications if and only if those are listed in Scopus or WoS. The list of publications we used matches the data available in the proprietary services. Thus, considering that the coverage for CD as gathered from open data sets is very high (~95%, as shown in Figure 2) we conclude that, in principle, these data sets are ready to be used as sources for evaluation processes in CD-based recruitment fields as an alternative for proprietary bibliographic databases. However, it is worth mentioning that, according to our analysis, they can be used only for identifying the relevant articles that a candidate can propose in the application, but we do not have any evidence about their potential use as an alternative for computing the citation-based metrics used in the NSQ, namely the *h*-index and the citation count.

It is also worth mentioning that, in this work, we have addressed only one step of the assessment process: the availability of (some) data to perform quantitative measurements. In the context of research assessment exercises, in fact, replacing closed data sources with open ones is an important step to address but it is not substantial for the scholarly community. Indeed, the community urges the design of research evaluation processes that consider multiple factors—such as sources' reputation and strength of the indicators, to cite just a few. These and other aspects might need to be reshaped in such a new context and are under discussion by the community and institutions—see, for example Directorate-General for Research and Innovation, European Commission (2021).

While we claim that, in principle, open data sets can be used in the NSQ when assessing CD-based recruitment fields, the same conclusion does not apply to ND disciplines, where the coverage was lower. Note that, in this case, the candidates certify themselves that the metadata of the publications they present in the NSQ are correct without mandatorily verifying them using external services. Thus, a comparison with these services is much more difficult to perform. Furthermore, we cannot compare the coverage of NDs directly against proprietary bibliographic databases (i.e., Scopus and WoS) as we did not have access to their full data sets, even if this could be a good input for a future study. However, by analyzing the data summarized in Figures 4–8, we conclude that there is a clear distinction between the RFs that relate to CDs and those that relate to NDs. Indeed, all the RFs of the first kind (all of those included in SAs 1–7 and 9, plus RFs 08/A1–4, 08/B1–B3, 11/E1–4) are characterized by having very high coverage in the combined data set (i.e., more than 95%) and at least two out of three data sets with coverage of more than 90%. All the other RFs, which do not comply with these rules, are indeed related to NDs. SA 13 (Figure 8) seems to be a slight exception to this rule: While being labeled by NDs only, it is more blurred due to its intrinsic nature, and it is very close to satisfying the rules mentioned above for some of its RFs.

We can also compare the results of our coverage of CDs against those obtained in other studies on MAG, Scopus, and WoS. For instance, Visser et al. (2021) show that MAG contains 81% of the publications that are listed in Scopus, computed considering the whole data sets in consideration. This value aligns well with our results, which are even better in favor of MAG, which contains more than 90% of the articles of CDs that are also included in both Scopus and WoS (as reported in Figure 3). This higher coverage compared with that of Visser et al. can result from:

- the additional matching of the articles in our collection with WoS, which may have resulted in better coverage; and

the fact that Italian authors applying to the NSQ for some CDs, being aware that only publications listed in Scopus and WoS will be considered, tend to publish more in venues that are indexed in such proprietary bibliographic databases. Another study by Huang et al. (2020) compared MAG against Scopus and WoS using a subset of articles published by authors working in 15 distinct institutions. Their study shows that MAG covers around 70% and 69% of Scopus and WoS publications, respectively. Instead, considering Scopus and WoS as a unique data set with disambiguated publications, MAG covers 67% of the publications in such proprietary services, which is much lower than the coverage (shown in Figure 2) we obtained for CDs and very close to that we have for NDs. In a future study, it could be worth investigating whether the disciplines to which the articles used in the study by Huang et al. (2020) are either CDs or NDs, to check if that low-coverage behavior could be derived from this aspect or is due to other factors.

Methods for correcting data in open and commercial data sets are also worth discussing. Currently, Crossref and MAG do not enable authors to correct wrong metadata of their own publications, while OpenAIRE allows university libraries to validate the metadata they provide to OpenAIRE and to enrich the metadata records with missing or extra information (Manghi, Artini et al., 2014). Instead, commercial services usually have curatorial units that react to authors' feedback when provided. This feature is perceived as an added value that can enable authors to have clean data before they are gathered for evaluation in the NSQ. In principle, this is possible for open data sets as well: The authors could provide similar corrections to the public, thus enabling the teams managing the open data sets to reuse and take in this information and allowing the NSQ commissions to correct possible mistakes in the original data before starting the evaluation procedure.

By analyzing the metadata gathered from the three open sources, we observe that the union of the data in MAG and OA approached that of all the data sets combined, as shown in Figures 1–8. This may mean that the support of Crossref data in our analysis did not add a lot to the other two data sets. However, MAG and OA probably took in Crossref data in advance, which would explain why Crossref did not show a big contribution to the overall coverage. In addition, the use of MAG could be perceived as a limit for the replicability of our study using updated data because of MAG's discontinuation at the end of 2021. Indeed, it has recently been replaced by OpenAlex (https://openalex.org/)—its first release has been entirely based on the last available snapshot of MAG. However, it is still not clear if future OpenAlex releases will show the same data coverage that Microsoft guaranteed with MAG.

There is an orthogonal aspect of the Italian NSQ that is not addressed in our study, and which will be explored in future work: citation coverage for CDs. Indeed, one of the parameters that is checked by the NSQ is the number of citations that all the candidates' publications have received in the past. Although we claim that the coverage of the open data sets used is good compared with that of Scopus and WoS, we cannot affirm the same for citation counts, at least in the context of the NSQ. Indeed, another study we have recently performed (Bologna, Di Iorio et al., 2021a) shows that open citation data available in the December 2020 release (OpenCitations, 2020) of OpenCitations' COCI (Heibi, Peroni, & Shotton, 2019) are not yet complete to substitute the data used in the NSQ made available by proprietary services. However, the combined use of several open citation sources—such as the new release of COCI (OpenCitations, 2022), which includes more than 1.29 billion citations, and the additional citation data from the open data sets used in this work and others, such as DataCite (Brase, 2009), and the proved fact that open citations are have been increasing dramatically in recent years (Hutchins, 2021)—are encouraging. In addition, a recent update (Martín-Martín, 2021) of a study by Martín-Martín et al. (2021) showed that the coverage of open citation data is approaching parity with those of WoS and Scopus. In a future study, we plan to analyze the coverage of citations in the context of the NSQ, and to compare the results with those available in past studies, such as Visser et al. (2021) and Martín-Martín et al. (2021).

Finally, it is important to stress one last point. As mentioned above, the NSQ evaluation only takes into account the publications indexed in WoS or Scopus for CDs. As shown in the outcomes of our study, there is a key distinction between the coverage of CDs and NDs in the open data sets: They are authoritative for CDs (at least compared with WoS and Scopus) and there are a few cases of publications that are relevant for the community but not listed there. The scenario is totally different for NDs disciplines, where a lot of relevant works (e.g., books discussing Humanities and Social Sciences research) are often omitted in commercial repositories. These publications might instead be available in open data sets. Then, it would be interesting to also investigate how much information is missing in Scopus and WoS but available in open data sets. This aspect could not be measured so far—as we start from the list of CDs publications selected by the candidates from the commercial data sets only and we could not measure the NDs publications in either WoS or Scopus because we did not have access to these commercial indexes—but we plan to explore it with a specific study in the future.

ACKNOWLEDGMENTS

This research has been supported by the University Fund for Research 2020 of the University of Modena and Reggio Emilia and by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017452.

AUTHOR CONTRIBUTIONS

Federica Bologna: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Angelo Di Iorio: Conceptualization, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing. Silvio Peroni: Conceptualization, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing. Francesco Poggi: Conceptualization, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

We gratefully acknowledge support from the University fund for Research 2020 (FAR) of the University of Modena and Reggio Emilia, and from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017452.

DATA AVAILABILITY

Data and software developed for this work are available at https://doi.org/10.5281/ZENODO.5025114 (Bologna et al., 2021b) and in the GitHub repository at https://github.com/sosgang/coverage_asn.

REFERENCES

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Bedogni, L., Cabri, G., Martoglia, R., & Poggi, F. (2021). Does the venue of scientific conferences leverage their impact? A large scale study on computer science conferences. *arXiv:2105.14838*. https://doi.org/10.48550/arXiv.2105.14838

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. https://doi.org/10.1162/qss_a_00018

Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2021a). Can we assess research using open scientific knowledge graphs? A case study within the Italian National Scientific Qualification. *arXiv:2105.08599*. https://doi.org/10.48550/arXiv.2105.08599

Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2021b). Data and code for "Open bibliographic data and the Italian National Scientific Qualification: Measuring coverage of academic fields." *Zenodo*. https://doi.org/10.5281/zenodo.5025114

Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2021c). Do open citations inform the qualitative peer-review evaluation in research assessments? An analysis of the Italian National Scientific Qualification. *arXiv:2103.07942*. https://doi.org/10.48550/arXiv.2103.07942

Brase, J. (2009). DataCite—A global registration agency for research data. *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology* (pp. 257–261). https://doi.org/10.1109/COINFO.2009.66

Chudlarský, T., & Dvořák, J. (2020). Can Crossref citations replace Web of Science for research evaluation? The share of open citations. *Journal of Data and Information Science*, *5*(4), 35–42. https://doi.org/10.2478/jdis-2020-0037

Crossref. (2021). January 2021 Public Data File from Crossref. *Academic Torrents*. https://academictorrents.com/details/e4287cb7619999709f6e9db5c359dda17e93d515

D. L. 2012. (2012). Redefinition of scientific disciplines (Rideterminazione dei settori concorsuali) (Prot. N. 159). Gazzetta Ufficiale Serie Generale n.137 del 14/06/2012—Suppl. Ordinario n.119. https://www.gazzettaufficiale.it/eli/id/2012/06/14/12A06786/sg

Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 95–127). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_4

Directorate-General for Research and Innovation, European Commission. (2021). *Towards a reform of the research assessment system: Scoping report (KI-09-21-484-EN-N)*. Publications Office. https://doi.org/10.2777/707440

Di Iorio, A., Peroni, S., & Poggi, F. (2019). Open data to evaluate academic researchers: An experiment with the Italian Scientific Habilitation. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019)* (pp. 2133–2144). Edizioni Efesto. https://arxiv.org/abs/1902.03287

Harzing, A.-W. (2016). Microsoft Academic (Search): A Phoenix arisen from the ashes? *Scientometrics*, *108*(3), 1637–1647. https://doi.org/10.1007/s11192-016-2026-y

Harzing, A.-W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1), 341–349. https://doi.org/10.1007/s11192-019-03114-y

Harzing, A.-W., & Alakangas, S. (2017a). Microsoft Academic: Is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383. https://doi.org/10.1007/s11192-016-2185-x

Harzing, A.-W., & Alakangas, S. (2017b). Microsoft Academic is one year old: The phoenix is ready to leave the nest. *Scientometrics*, *112*(3), 1887–1894. https://doi.org/10.1007/s11192-017-2454-3

Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, *121*(2), 1213–1228. https://doi.org/10.1007/s11192-019-03217-6

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

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., … Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. https://doi.org/10.1162/qss_a_00031

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551–1571. https://doi.org/10.1007/s11192-017-2535-3

Hutchins, B. I. (2021). A tipping point for open citation data. *Quantitative Science Studies*, *2*(2), 433–437. https://doi.org/10.1162/qss_c_00138, PubMed: 34505061

L. 240/2010. (2011). Rules concerning the organization of the universities, academic employees and recruitment procedures, empowering the government to foster the quality and efficiency of the university system (Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonché delega al Governo per incentivare la qualità e l’efficienza del sistema universitario). Gazzetta Ufficiale Serie Generale n.10 del 14/01/2011—Suppl. Ordinario n.11. https://www.gazzettaufficiale.it/eli/id/2011/01/14/011G0009/sg

Manghi, P., Bolikowski, L., Manold, N., Schirrwagen, J., & Smith, T. (2012). OpenAIREplus: The European scholarly communication data infrastructure. *D-Lib Magazine*, *18*(9/10). https://doi.org/10.1045/september2012-manghi

Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., … Principe, P. (2021). OpenAIRE Research Graph Dump (3.0). *Zenodo*. https://doi.org/10.5281/zenodo.4707307

Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., … Pagano, P. (2014). The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, *48*(4), 322–354. https://doi.org/10.1108/PROG-08-2013-0045

Martín-Martín, A. (2021). Coverage of open citation data approaches parity with Web of Science and Scopus [Blog]. *OpenCitations Blog*. https://opencitations.wordpress.com/2021/10/27/coverage-of-open-citation-data-approaches-parity-with-web-of-science-and-scopus/

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4, PubMed: 32981987

Microsoft Academic. (2020). Microsoft Academic Graph (2020-01-23). *Internet Archive*. https://archive.org/details/mag-2020-01-23

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

OpenCitations. (2020). COCI CSV dataset of all the citation data—December 2020 dump. *figshare*. https://doi.org/10.6084/m9.figshare.6741422

OpenCitations. (2022). COCI CSV dataset of all the citation data—March 2022 dump. *figshare*. https://doi.org/10.6084/m9.figshare.6741422.v14

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: Redescubriendo el ecosistema de la información científica. *El Profesional de La Información*, *27*(2), 420. https://doi.org/10.3145/epi.2018.mar.21

Peroni, S., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Poggi, F., & Presutti, V. (2020). The practice of self-citations: A longitudinal study. *Scientometrics*, *123*(1), 253–282. https://doi.org/10.1007/s11192-020-03397-6

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*, *12*(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006

Van Noorden, R. (2014). Google Scholar pioneer on search engine’s future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112

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

Wu, J., Kim, K., & Giles, C. L. (2019). CiteSeerX: 20 years of service to scholarly big data. *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse 2019* (pp. 1–4). https://doi.org/10.1145/3359115.3359119

Zhu, Y., Yan, E., Peroni, S., & Che, C. (2020). Nine million book items and eleven million citations: A study of book-based scholarly communication using OpenCitations. *Scientometrics*, *122*(2), 1097–1112. https://doi.org/10.1007/s11192-019-03311-9