RESEARCH ARTICLE

Open access at the national level: A comprehensive analysis of publications by Finnish researchers

Janne Pölönen[1], Mikael Laakso[2], Raf Guns[3], Emanuel Kulczycki[4], and Gunnar Sivertsen[5]

[1]Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)
[2]Hanken School of Economics, Information Systems Science, Arkadiankatu 22, 00100, Helsinki (Finland)
[3]University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (ECOOM), Middelheimlaan 1, 2020 Antwerp (Belgium)
[4]Adam Mickiewicz University, Scholarly Communication Research Group, Szamarzewskiego 89c, 60-568 Poznań (Poland)
[5]Nordic Institute for Studies in Innovation, Research and Education (NIFU), P.O. Box 2815,0608 Tøyen, Oslo (Norway)

Keywords: bibliographic data, bibliometrics, data quality, open access, scholarly communication, science policy

ABSTRACT

Open access (OA) has mostly been studied by relying on publication data from selective international databases, notably Web of Science (WoS) and Scopus. The aim of our study is to show that it is possible to achieve a national estimate of the number and share of OA based on institutional publication data providing a comprehensive coverage of the peer-reviewed outputs across fields, publication types, and languages. Our data consists of 48,177 journal, conference, and book publications from 14 Finnish universities in 2016–2017, including information about OA status, as self-reported by researchers and validated by data-collection personnel through their Current Research Information System (CRIS). We investigate the WoS, Scopus, and DOI coverage, as well as the share of OA outputs between different fields, publication types, languages, OA mechanisms (gold, hybrid, and green), and OA information sources (DOAJ, Bielefeld list, and Sherpa/Romeo). We also estimate the role of the largest international commercial publishers compared to the not-for-profit Finnish national publishers of journals and books. We conclude that institutional data, integrated at national and international level, provides one of the building blocks of a large-scale data infrastructure needed for comprehensive assessment and monitoring of OA across countries, for example at the European level.

1. INTRODUCTION

While open access (OA), free of cost and other access barriers, has been gradually emerging for over two decades, it has recently gained a lot of momentum through science policy. In 2016, the European Union member states agreed to "[…] open access to scientific publications as the default option by 2020 and to the best possible re-use of research data as a way to accelerate the transition towards an open science system." (Council of the European Union, 2016). The European Commission supports the transition with a strong open science agenda (European Commission, 2018). Recently, a group of European research funders formed cOAlition S, where funders from around the world are invited to join and make a shared commitment to make immediate OA and

an open access journal

Citation: Pölönen, J., Laakso, M., Guns, R., Kulczycki, E., & Sivertsen, G. (2020). Open access at the national level: A comprehensive analysis of publications by Finnish researchers. *Quantitative Science Studies*, *1*(4), 1396–1428. https://doi.org/10.1162/qss_a_00084

DOI: https://doi.org/10.1162/qss_a_00084

Received: 25 October 2019
Accepted: 02 August 2020

Corresponding Author: Janne Pölönen
janne.polonen@tsv.fi

Handling Editor: Ludo Waltman

The MIT Press

unrestricted use a requirement for all published research funded by the signatories by 2021. This funder initiative, referred to as *Plan S*, concerns at this stage only journal articles, while a longer transition period is planned for peer-reviewed book publications.

Despite being on a steady growth curve and despite increasing support by science policy at various levels, OA has yet to become the default way of disseminating scholarly works. The number of peer-reviewed journals allowing immediate OA to all content (gold OA) is growing, but the majority of peer-reviewed journals are still subscription based. Only a small share of articles published in subscription-based journals are bought OA individually (hybrid OA; these articles are made OA against a payment) or are actually self-archived in OA repositories (green OA), even if this is permitted. Only a small fraction of scholarly books are published free and open for everyone to download and read (Piwowar, Priem, et al., 2018; doabooks.org, 2019).

Given the increasing investment in the advancement of OA by policymakers, institutions, funders, and researchers, there is a need for monitoring the state and development of OA at the international and national levels. So, what share of research publications are available OA per year, either nationally or globally? These are some of the most fundamental questions anyone with an interest in OA could reasonably ask. However, giving a straightforward answer has so far not been easy. Our understanding of the growth and uptake of OA is conditioned by the availability of data and measurement tools of OA development. There are several well-established and emerging international data sources for publication data—such as WOS, Scopus, Google Scholar, Microsoft Academic, and Dimensions—but recent large-scale analyses have highlighted their limitations in coverage of publication outputs (Martín-Martín, Thelwall, et al., 2020; Visser, van Eck, & Waltman, 2020). An inevitable question is: How representative, accurate, and biased are the results of OA monitoring based on such sources for different countries and fields?

The aim of our study is to show that it is possible to achieve a national estimate of the number and share of OA based on institutional publication data providing the most complete source of the universities' peer-reviewed output. In this paper, we explore and use the institutional publication data from 2016–2017 stored in the VIRTA Publication Information Service, which integrates data from the different types of commercial and noncommercial CRIS solutions of 14 Finnish universities (Puuska, Guns, et al., 2018; Pölönen, 2018), to describe the landscape of OA publishing in Finland, including all publication types, languages, and fields. To identify OA publications, we also use OA status information from VIRTA, which has been self-reported by researchers and validated by data collection personnel at universities (Ilva, 2017a).

More specifically, we investigate the added value of institutional data for OA study compared with WoS, Scopus, and DOIs in terms of national publication output coverage, the OA share of outputs across different fields, publication types, and languages based on comprehensive data, the coverage and information value of international sources for gold (DOAJ and Bielefeld list) and green OA journals (Sherpa/Romeo), and dominance of the largest international commercial publishers. Although our analyses are based on data concerning Finnish universities, our findings are also relevant for an international audience with regard to the options, challenges, and advantages of using institutional publication data for OA monitoring at the national level in other countries as well as across countries, for example at the European level.

In the introduction, we provide background information on the existing data sources and methods of OA monitoring. Our literature review, presented in section 2, shows that CRIS data remains underexploited in study of OA uptake. In section 3 we present our research questions, data, and methods. The results of our empirical analysis are presented in section 4, followed by discussion in section 5 and conclusions in section 6.

Data Sources and Methods for OA Monitoring

Science policy, specifically concerning OA, would benefit from having frequently updated comprehensive metrics collected through a consistent methodology and definitions to support decision-making and monitoring. However, this is an area where there is still a lot of room for improvement. A practical example of key problems with regards to publication information availability is the European Open Science Monitor (Waltman, 2019). This is a service funded by the European Commission and intended to provide regularly updated country-level metrics on OA development (European Commission, 2019). While the use of Elsevier, the largest scholarly journal publisher with an ongoing influence and financial interest in the development of the OA landscape, as a subcontractor has been appealed to no avail (Tennant, 2018), this is a concern not just limited to the potential impact on business and market competition. What is also a concern for scholarship more widely is that the metrics used for the monitor are based on Elsevier's Scopus bibliographic database, which is an index widely used for various purposes where the scholarly publication landscape is to be represented. While Scopus is more inclusive than its closest competitor, WoS by Clarivate Analytics, both still leave out a substantial part of the scholarly record and have been found to be limited in many regards to what they include (see, e.g., Archambault et al., 2006; Chavarro, Ràfols, & Tang, 2018; Hicks, 1999; Hicks & Wang, 2011; Larivière & Macaluso, 2011; Mongeon & Paul-Hus, 2015; Nederhof, 1989, 2006; Somoza-Fernández, Rodríguez-Gairín, & Urbano, 2018). The literature review section of this article will reveal that almost all studies of national OA uptake have relied on publication data from either Scopus or WoS, which is a fundamental limitation of perspective.

In many countries, universities annually report their complete bibliographic record of peer-reviewed publications to the government as part of performance-based research funding systems (PRFS) (Giménez-Toledo, Mañana-Rodríguez, et al., 2017, 2019; Hicks, 2012; Sīle, Guns, et al., 2017; Sīle, Pölönen, et al., 2018). In Norway, Denmark, Finland, Flanders (Belgium), and Poland—countries that have in some form adapted the so-called Norwegian model of PRFS—the national bibliographic database either substitutes the universities' local Current Research Information Systems (CRIS) or integrates publication data from the local CRIS (Aagaard, 2018; Engels & Guns, 2018; Kulczycki & Korytkowski, 2018; Pölönen, 2018; Sivertsen, 2016a, 2016c, 2017, 2018a). Comparisons with the comprehensive national publication data have shown that especially in the social sciences and humanities (SSH), WoS and Scopus coverage is seriously lacking, mainly due to the importance of national language and book publishing (Aksnes & Sivertsen, 2019; Giménez-Toledo et al., 2016, 2017; Kulczycki, Engels, et al., 2018; Kulczycki, Guns, et al., 2020; Ossenblok, Engels, & Sivertsen, 2012; Sivertsen, 2016b; Sivertsen & Larsen, 2012). In many SSH disciplines, the majority of journal articles are published in national or regional outlets not indexed in WoS or Scopus (den Hertog, Jager, et al., 2014; Sivertsen, 2016b). In addition, up to half of peer-reviewed outputs in the humanities, and around one-third in the social sciences, are book publications, including chapters and monographs (Engels, Starčič, et al., 2018). The implication is that only countries in which a national bibliographic database with full coverage of the SSH publications (Sīle et al., 2018) has been developed can provide an accurate picture of publications across all fields and publication types.

In addition to coverage issues of publications in international bibliographic databases, OA monitoring is conditioned by OA definitions and methodologies for identifying what is available OA among these publications. The Directory of Open Access Journals (DOAJ) and Sherpa/Romeo are the most frequently used information sources to identify gold and green OA journals. But not all gold OA journals are included in DOAJ. Bielefeld University, for example, provides an ISSN-matching of gold OA journals based—in addition to DOAJ—also on the Directory of Open Access Scholarly Resources (ROAD), PubMed Central (PMC), and Open APC (OAPC) (Wohlgemuth,

Rimmert, & Winterhager, 2016). According to the most recent analysis, Bielefeld's list contained 7,755 gold OA journals, of which DOAJ covered 33% (Bruns, Lenke, et al., 2019). Recently, Björk (2019) identified 437 OA journals published in the Nordic countries, of which DOAJ covered 42%. There were also considerable differences between the Nordic countries, as DOAJ covered 68% of OA journals from Norway but only 23% of those published in Finland. The Federation of Finnish Learned Societies and DOAJ have started a pilot project to encourage Finnish OA journals to apply to DOAJ (DOAJ, 2019). The Sherpa/Romeo register of self-archiving policies has extensive coverage of journals, but the information value of the color codes used for classification of the policies—notably the identification of green OA journals—has been questioned, as publishers have increasingly introduced additional requirements not captured by the color codes (Gadd & Troll Covey, 2016). Sherpa/Romeo recently launched a new version of the service, in which the color codes are no longer used.

It is in the interest of the European Commission to have a comprehensive open science monitor, based on open and transparent data-infrastructure independent of private operators (Tennant, 2018; Waltman, 2019). Therefore, it is important to investigate the institutional CRIS data not only from the national OA perspective but also because it potentially contributes to the large-scale international data infrastructure needed for evaluation, assessment, and monitoring of research activities at the European level (European Commission, 2010; Lauer, 2016; Mahieu, Arnold, & Kolarz, 2014; Sivertsen, 2019). OpenAIRE and Crossref are important building blocks of such an infrastructure for open metadata (Waltman, 2019). Since 2018, data collected and made available by the service Unpaywall can be used to identify different types of OA publications based on DOIs (Piwowar et al., 2018). Recent analyses show, however, that the availability of DOIs is far from complete, and there are considerable differences in DOI availability between publication types, fields, and countries (Boudry & Chartron, 2017; Fasae & Oriogu, 2018; Gorraiz, Melero-Fuentes, et al., 2016). The added value of integrated CRIS data is that it can provide well-structured and curated metadata of all publications, whether they are included in WoS and Scopus or not, are in printed or digital format, have DOI or not, and are openly available on the internet or not. Indeed, the Finnish VIRTA Publication Information Service, a national solution for integrating publication data from diverse local CRISs, has already been tested to integrate CRIS data from four European countries (Puuska et al., 2018).

Challenges with Using CRIS-based Data

The national bibliographic databases also have their own challenges of data coverage and quality. If assessed based on included publication types and languages, types of research organizations and organizational units, seniority and job positions of authors, fields of science, intended audience of publications, and peer-review status, most CRIS-based national databases can be described as very comprehensive (Sīle et al., 2017, 2018). Several studies, referred to above, have indeed demonstrated the substantially larger coverage of national publication data compared to WoS and Scopus. Similarly, a study of a single Dutch university showed a substantially larger coverage of outputs, especially in the SSH, in the local CRIS compared to WoS (van Leeuwen, van Wijk, & Wouters, 2016). Further studies are needed, however, to investigate to what extent publications included in WoS or Scopus may be missing from the CRIS-based data (e.g., due to researchers' failure to report). Especially in the case of national databases supporting PRFS, it promotes their comprehensiveness that universities not only have a considerable financial incentive to secure as complete reporting of publications as possible but the reporting is also legally mandated (Sīle et al., 2018; Sivertsen, 2018a, 2019).

CRISs are needed, among other things, to provide comprehensive, reliable, comparable, and transparent information on research activities (Science Europe, 2016). Completeness is an

important aspect of data quality, in addition to correctness, consistency, and timeliness of data (Azeroual & Schöpfel, 2019; Sīle, Guns, et al., 2019). A major challenge is the variety and complexity of publication information (e.g., OA status of publications) and the diversity of data providers and sources (e.g., researchers, data-collection personnel, external databases). Diversity of practices between fields and publication types can increase ambiguity over definitions, such as peer-review status of publications (Kaltenbrunner & de Rijcke, 2016; Pölönen, Engels, & Guns, 2019). In national databases supporting PRFS, standardization and interoperability of data are promoted by means of national level data-collection guidelines with definitions and requirements for reported publications (Sivertsen, 2019). Nevertheless, research-performing organizations (e.g., universities) have different procedures for maintaining records about the publications that affiliated researchers have authored (van Leeuwen, van Wijk, & Wouters, 2016). The quality of the data stored in national bibliographic databases has not yet been extensively researched, but Azeroual and Schöpfel (2019) shed some light on how representatives from 17 European institutions perceive the aspect of data quality in their CRISs. The survey showed that the institutions have several ways that they support and improve the data quality stored in their CRISs, both through internal validation processes and by matching entries to external data.

The growth of OA both in terms of uptake and weight in science policy has introduced a need for new data fields and functions for publication data stored in CRISs. What makes recording of OA information in CRIS systems challenging is the versatility of ways that content can be made available OA, where mechanisms are not necessarily mutually exclusive. The OA status is also likely changing over time, with overlapping access mechanisms, and not clearly or uniformly understood by all information providers (Ilva, 2017a). In addition to journals that publish all their content OA immediately, many subscription-based journals allow individual papers to be made OA on the publisher's website for a one-time fee: so-called hybrid OA. Subscription-based journals that allow self-archiving of manuscript versions of published articles may impose embargoes for the peer-reviewed postprint and publisher version, making them not compliant for example with the Plan S requirements. It has also been observed that publishing in journals that allow self-archiving does not automatically mean that publications are actually deposited in OA repositories, highlighting a gap between potential and uptake (Björk, Laakso, et al., 2014; Laakso, 2014). OA versions of articles can also be provided on, for example, personal websites or academic social networks that do not guarantee persistent access.

2. LITERATURE REVIEW

This study focuses on the context of OA measurements at the country level. There are a number of earlier studies that have contributed to this line of research, where the goal has been to cover publication records for an individual country, or multiple individual countries but each country reported separately, and study OA from some perspective. The central methodological variation in earlier studies concerns mainly (a) the data source(s) used for the baseline publication records and (b) how the identification and classification of various OA mechanisms enabling access to these publications is implemented. No studies summarized here include content that might be retrievable from Sci-Hub, a pirate website running since 2011 containing 85% of articles published in subscription journals (Himmelstein, Romero, et al., 2018).

The written summaries, ordered chronologically, provide details on how each study has approached the two central factors of data source selection and OA identification. The summaries of research focusing on country-level OA measurement are divided into two subsections depending on whether the studies are based on WoS or Scopus data or whether they use publication data from CRIS. A third subsection is reserved for studies that do not provide country-level

OA measurement but are in other ways relevant to the study. A fourth and final subsection is dedicated to sources describing the Finnish environment for academic publishing and research.

2.1. Studies Using Web of Science or Scopus

The United Kingdom has been a pioneer in implementing science policy measures facilitating OA, which has also led to many reports concerning monitoring the development over time. The most recent report by Research Information (2017) presents an analysis of the 2016 scholarly journal output by UK-affiliated authors, utilizing Scopus as the source of baseline publication data. To identify OA publications, the study adopted various methods which are documented in a methodological annex, making use of DOAJ, information on publisher websites, and manual sampling to estimate shares. Some 36% of articles were available from publishers as either gold OA, hybrid OA, or delayed OA, with a further 16% as green OA through online postings in line with journal policies. Although the study is efficient in differentiating between various OA mechanisms, being based on only Scopus-indexed outputs limits the level of insight it can provide about the entire scholarly publishing landscape. The data is also not made openly available.

In a broad study, Martín-Martín, Costas, et al. (2018) studied the OA status of 2,269,022 journal articles (including reviews) recorded in the three central WoS citation indexes for the years 2009 and 2014. For identification of OA availability, and provision mechanisms, the authors queried Google Scholar for each article in conjunction with matching to data from the DOAJ, CrossRef, OpenDOAR, and ROAR. The study found the world average of OA provided through publishers or repositories to be 35.8% of all articles published in 2014, with an additional 20% of articles being available through other freely accessible pages on the web indexed by Google Scholar. There was considerable variation in the OA levels among countries, where each publication was assigned to a country if at least one author was affiliated with an organization in that country. Focusing on the OA share provided by either publishers or repositories, the lower end of the spectrum was represented by Iran (18.6%), Russia (20.3%), and India (23.1%). The highest end of the spectrum was populated by Scotland (56.6%), England (50.9%), and Sweden (50.2%). Finland was not included among the 25 countries in the study. While the study incorporates one of the broadest lenses yet for identifying various OA mechanisms and reporting on them separately (breakdown is provided for gold OA, hybrid OA, delayed OA, bronze OA, green OA, and other free availability) it is limited by restricting the set of publications to those indexed in WoS and by only incorporating journal articles as publication type. The categories of bronze OA and other free availability consist of content to which access might be revoked at any time and their terms and licensing for redistributed openness are often unclear.

Bosman and Kramer (2018) provide a study available in preprint form based on WoS journal publication data that includes longitudinal OA development for journal articles (+ reviews) in the period 2010–2017 for 76 individual countries. To identify OA content, the authors utilize oaDOI, which is a database that harvests information about OA versions available for articles based on DOI information from various openly available sources (including DOAJ and BASE; Impactstory, 2017). The oaDOI database and API have since been made part of Unpaywall. The results demonstrate a large discrepancy in OA levels between the countries. For European countries the spread was 20% (Romania) to 42% (Netherlands) for 2016. Finland had an OA share of 32% for the year 2016. The general longitudinal trend for all countries was of increasing OA share over time, outside of the most recent measurement year (2017), which the authors suggest to be due, for example, to certain time-bound OA mechanisms not being immediately in effect.

In a report incorporating research outputs spanning a decade, Science-Metrix (2018) presents a bibliometric study of the degree of articles being OA in WoS for the years 2006–2015, which

includes a country-level analysis covering 20 countries. Finland was not among the studied countries. The world average was 41% for 2015, and this share included all versions of articles that can be downloaded for free from the web that have been harvested into the 1science database (which provides data for the 1findr product that is sold by Science-Metrix, which is now part of Elsevier, to help organizations discover OA content). The study also presents a table differentiating between gold OA and green OA shares between countries for WoS articles in 2014, where significant differences are present showcasing the results of different science policy approaches that countries have adopted to facilitate OA.

Demonstrating the variety of ways in which OA shares can be measured for a set of publications, van Leeuwen, Tatum, and Wouters (2018) compared the use of three different bibliographical methods to assess gold OA publishing at the national level, focusing on research output from the Netherlands, Denmark, and Switzerland for 2000–2013. The three approaches differ in how they rely on either only one or multiple of the following data sources: WoS publication data, DOAJ data, and a customized WoS database hosted at the Centre for Science and Technology Studies (CWTS) at Leiden University. While the three approaches differ in how the OA status of publication records is obtained (the first on OA journal status data in WoS, the second on DOI matching of articles to DOAJ journals, and the third based on ISSN matching of journal articles to DOAJ), they are all limited to the realm of publications included in WoS. Each of the three approaches had their individual pros and cons, with no approach being a full replacement of the others. The authors conclude with a discussion about the potential for utilizing CRIS data for similar purposes in the future to get around the limitations in publication data and OA identification.

While not an academic study, the previously mentioned European Open Science Monitor provided by the European Commission is a resource for regularly updated country-level metrics on OA development (European Commission, 2019). The use of Scopus data provided by Elsevier for the underlying journal publication data comes with limitations on the index coverage as well as potential conflicts of interest, with Elsevier being the largest scholarly journal publisher with a large ongoing influence on the OA landscape (Mongeon & Paul-Hus, 2015; Tennant, 2018). The monitor presents its results split into two categories: gold OA and green OA. The methodology documentation describes that DOAJ, ROAD, PubMed Central, Crossref, and OpenAIRE are used as the main data sources for identifying OA status (European Open Science Monitor, 2018). The monitor has recently also added Unpaywall to its list of data sources and pins the global OA share of 2017 publications to 35.7%, with 13.9% of articles being available as gold OA and 24% as green OA. Shares for 36 individual countries are only given for one period (2009–2017) as a single snapshot, which makes it hard to perceive recent developments. Most countries have similar shares of gold OA publishing in this timespan and the largest differences are based on variation in the level of green OA. The top end is populated by Switzerland (52.2%), the United Kingdom (50.9%), and Denmark (47.7%), while the lower end contains Russia (21.3%), China (22.9%), and India (30.1%). Finland was measured to have an OA share of 41.6%, with 11.2% as gold OA and 31.4% as green OA.

2.2. CRIS-Based Studies

In a Swedish-language report by the National Library of Sweden, Kronman (2017) analyzed CRIS publication data for 2010–2016 concerning peer-reviewed articles (including articles in conference proceedings) of 42 research-performing organizations in Sweden. The OA status of 278,195 articles was assessed by augmenting the OA publication metadata found in SwePub with matching to oaDOI. Of the articles for 2010–2016, 39% could be matched through oaDOI, with 14% being in full OA journals, 22% green open access (uniquely available), and 3% hybrid OA. This study demonstrates that combining CRIS data with external sources for article-level OA

identification is possible. The main drawback of utilizing oaDOI for this purpose is being limited to articles that have been assigned a DOI.

The most comprehensive analysis so far concerning OA in the Finnish publication landscape is a study by Ilva (2017b) which is available in Finnish. The author provides an overview of summarized data for the publication year 2016 of all publication types based on CRIS data reported from all universities, universities of applied sciences and research centers in Finland. The data includes OA status information submitted by the organizations themselves, with each publication being published in an OA journal or as hybrid OA, and/or self-archived in a repository. An embargo is allowed for the alternative of self-archiving but not for the alternative of full OA journal which omits the alternative of OA through delayed OA journals (Ilva, 2017b). The study provides a breakdown of OA availability of peer-reviewed articles published by university-affiliated authors, with 18.7% of articles being in full OA journals, 3.4% on the publisher's website but not in a full OA journal (e.g., hybrid OA), and 18.5% self-archived to a repository. This study demonstrates the viability and challenges of basing national OA measurement on CRIS data alone, avoiding many of the limitations in scope concerning which publications are included, but with some added ambiguity in OA identification, as data is self-reported in a decentralized way and can contain inconsistencies.

Mikki (2017) studied the openness of 70,882 journal articles published by Norwegian authors for 2011–2015 by analyzing data reported from the CRIS systems of Norwegian institutions and querying Google Scholar with either the DOI or name of the article to determine openness status. The study did not discern between OA mechanisms and found that 67.6% of all articles were openly available in some full-text form through Google Scholar. The web domains providing the most articles for download were researchgate.net and academia.edu, suggesting that a notable part of the measured OA is likely not in line with publisher policies. The study also included analysis of OA shares across the 15 largest publishers, research disciplines, and the four largest universities in Norway. While the author found the organizational variance in OA shares to be fairly even, disciplinary differences were notable (high shares for natural sciences and technology, low shares for SSH). In terms of publisher proportions of OA, the variation among the 15 largest publishers ranged from over 70% of articles with only paywalled article access available (Routledge and Universitetsforlaget) to corresponding shares of 35% and 32% for Elsevier and Springer. A further study built upon a similar Norwegian CRIS-based publication data set for journal articles as in Mikki (2017) is by Mikki, Gjesdal, and Strømme (2018) that extends to study one additional year of journal publications (2011–2016). In this study a comparison is made between the capabilities of Google Scholar, oaDOI, and 1findr to retrieve OA copies of articles documented in the data set describing 87,439 journal articles. Google Scholar was found to be the best at retrieving full-text OA copies of articles queried, doing so for 70% of all queries. The corresponding figures for 1findr and oaDOI were 52% and 31%.

In a recent report for The Association of Universities in the Netherlands (VSNU), Bosman and Kramer (2019) evaluated the OA status of all publications from 2017 by Dutch universities assigned with a DOI in WoS, Scopus, or Dimensions. OA status was assessed by querying each DOI to the Unpaywall database in June/July 2018, finding that the OA share of article publications varied between 45% and 55% across publications included in the three databases. The report acknowledges that relying on DOI publications favors article publications over other publication types, which in turn also likely increases the OA share should all publication types be more comprehensively included.

Sivertsen, Guns, et al. (2019) provide a study of longitudinal CRIS data for Finland (2011–2017), Flanders (2011–2016), Norway (2011–2017), and Poland (2013–2016; only SSH publications) where identification of OA status is managed by comparing journal records to those indexed in

the DOAJ. The results are presented per country per discipline, where the share of articles in DOAJ-indexed journals ranged from 5.7% (Social sciences, Flanders) to 17.3% (Medical & health sciences, Norway). The study found that publishing in full OA journals was on the rise in all four countries. As disciplinary OA shares varied between countries, the authors suggest that uptake of OA should not be seen as exclusively steered by the availability of OA outlets within a discipline, but rather also on local and contextual factors.

Based on this review of earlier country-level studies it can be concluded that national information sources remain underexploited in analysis of OA, and previous studies have focused predominantly on journal publishing (with the exception of Ilva [2017b]). The focus on journal articles is explained partly by OA policies and research funder mandates that have, so far, mainly concerned only this publication type. The reliance on Scopus and WoS for defining which publication outputs are considered and included is also a common limitation among many studies. There seem to be very heterogeneous approaches to how OA mechanisms are defined across the studies, but a common issue is noncomprehensiveness. Publication types other than journal articles are often excluded in the initial stages of studies. Though they are increasingly common, not all journal articles have DOIs (see, e.g., Boudry and Chartron (2017). The use of CRIS data in the context of national OA measurement has so far been limited to Nordic countries, where CRIS use has long been part of practice and reporting routines, with the exception of Sivertsen et al. (2019) where CRIS data for Finland, Flanders, Norway, and Poland were explored through the lens of articles being published in DOAJ-included journals.

2.3. Other Relevant Studies

In a recent preprint, Huang, Neylon, et al. (2020) thoroughly evaluate the coverage discrepancies between WoS, Scopus, and Microsoft Academic. The study results show that each database differs a lot in coverage, which suggests that any bibliometric evaluation of organizational or country output aiming to be comprehensive should include publication data from several sources rather than relying on just one. An interesting finding was that Microsoft Academic contained most unique DOIs of the three databases, including, in particular, more book chapters and conference proceedings than the other two. Given the lack of standardized CRIS data across institutions and countries, it thus seems that Microsoft Academic is the best current solution regarding output comprehensiveness.

The findings and implications of Huang et al. (2020) share a lot of commonality with those of the report by Bosman and Kramer (2019), which focuses mainly on comparing nationally aggregated Dutch publication data with the indexing coverage of WoS, Scopus, and Dimensions. The coverage between the databases varied a lot between them, often showing strengths in indexing of specific publication types, and the study found that only 43% of publications with a DOI were identifiable in all three databases. Also comparing any of the international indexes to the nationally aggregated data, in particular, nonarticle output and even very substantial shares of Arts/Humanities journal articles were left out of the population if restricting inclusion criteria to only items with DOIs. The report also provides a brief inquiry into the comprehensiveness of LENS, BASE, NARCIS, and OpenAIRE, but further insight into the comprehensiveness of these databases is limited due to lack of reliable affiliation identification. The reports provide evidence for strong disciplinary differences in publication types, which together with the knowledge that international bibliometric indexes are limited and skewed in their comprehensiveness, should have implications for using bibliometric databases for assessing and potentially influencing publication behavior with policy interactions.

In a study looking into facilitating factors for consistent institutional use of CRIS systems and OA policies in three countries (Italy, the Netherlands, and Germany), Biesenbender, Petersohn,

and Thiedig (2019) concluded that such practices are particularly facilitated if national evaluation or quality assessment policies are in place. As the next section will describe, this is very much the case in the context of Finland. The authors highlight that the role of CRIS data is often overlooked in the context of open science development, even though such data, in conjunction with self-archiving in repositories, already play a major role, with a lot of future potential for growth.

Crawford (2019) provides an extensive analysis of all journals included in the DOAJ, including annual publication volumes for each journal and detailed breakdowns of differences between research areas and regions of the world. The study includes thorough analysis of journals and articles that are published in journals that are free for authors, and the pricing levels for journals that charge APCs. Based on Crawford (2019), there were 11,465 active journals in 2018 across all major disciplines, most of which were free for authors. There are often accessible OA journals available for researchers to publish in, but there might be other incentives rather than openness guiding their publishing preferences.

Despite not including a country analysis, the most recent and robust measurement of OA availability of journal articles provided by Piwowar et al. (2018) warrants highlighting. The study is relevant by demonstrating how a wide breadth of various OA mechanisms can be classified and studied by using the Unpaywall API for articles with DOIs. As we pointed out in the introduction, the main limitation of using Unpaywall for OA measurement is the reliance on publications having DOIs.

Important information in OA measurement studies is the breakdown of which mechanisms OA is being provided through, but arguably equally important is to look at the share and likely reasons why certain parts of the literature have not been made available. Laakso (2014) provides an analysis of the maximum potential for self-archiving journal articles among the 100 largest journal publishers indexed in Scopus. While the results of this study are already outdated, the methodological concept of calculating article-level realized and unused potential based on publisher self-archiving policies is something that the current study will carry forward.

2.4. The Context of Finland

As this study concentrates on the publication output of researchers at Finnish universities, it is beneficial to briefly describe the national science policy environment, and in particular how OA has become an important part of it over time. Like many European countries, Finland has been at the forefront of developing national strategies for advancing OA. In 2014–2017, the Ministry of Education and Culture funded a national project, the Open Science and Research Initiative, which set ambitious national targets for the share of open access research publications: 65% in 2017, 75% in 2018, and 100% in 2020 (Ilva, 2017b).

Finnish universities and universities of applied sciences receive a substantial part of their public performance-based funding on the basis of their publication activities, which is one of the reasons that CRIS data in Finland is so comprehensive compared to, for example, WoS or Scopus. Like in Norway and Denmark, a nationally constructed rating based on evaluation by panels of experts, referred to as the *Publication Forum* (in Finnish *Julkaisufoorumi*, or JUFO), is used in Finland to categorize publication channels (i.e., journals, book publishers) into four different levels, which determines the weight of individual peer-reviewed outputs for calculating public funding (Pölönen, 2018). Based on calculations from realized funding from 2016, a top-ranked article generated approximately €17,000 for institutions with an affiliated author or coauthor on such a publication, while the three lower levels were approximately €12,600, €4,200, and €420 respectively (Seuri & Vartiainen, 2018). Because of this, there is a strong motivation for the organizations to provide comprehensive data on their publications on time. Universities have

reported the OA status of their publications since 2011, but the data fields used in the collection of publication data were changed from 2016 onwards to give a more comprehensive picture of both OA journal publishing and/or self-archiving through the data (Ilva, 2017b; VIRTA Wiki, 2018). Recently, the Finnish government approved a revised funding model for allocating core funding annually to universities in 2021–2024, which incorporates an extra 20% weight for the funding contributed by each publication if it is reported as being available OA (Ministry of Education and Culture, 2019a), accepting gold, green, and hybrid OA.

In Finland, university contracts with international journal publishers are mostly handled centrally by FinELib, a consortium of Finnish universities, research institutions, and public libraries. Finland has been among the pioneers in making the costs of all publisher agreements publicly available since 2016 (Etsin, 2018). FinELib is a signatory of the OA2020 initiative and has included OA elements as part of the negotiated contracts since at least 2015, aiming to include substantial OA publishing elements into all new agreements (FinELib, 2019). Given that the five largest international commercial publishers account for more than half of the global journal output indexed in WoS (Larivière, Haustein, & Mongeon, 2015), most attention at both the international and national levels is focused on negotiating with these publishers to enable OA options.

The Academy of Finland, the major national research funder, has been mandating OA for funded research projects since 2015, accepting both green OA and gold OA as viable paths to fulfilling the requirement (Academy of Finland, 2019). The Academy of Finland became a signatory of Plan S soon after the initial plan was revealed.

For national journals, there have not been strong financial incentives to convert to OA (e.g., major funding mechanisms requiring it). Nevertheless, the Federation of Finnish Learned Societies allocates state subsidies annually to journals and book series, one of the criteria being an open access plan. In a recent study of Nordic peer-reviewed OA journals, which included a subset of journals published in Finland, Björk (2019) calculated that 97 out of 334 (30%) journals were published as full OA in the autumn of 2018. A centralized publishing platform, Journal.fi, is available for any national journals that are OA with a maximum delay of 12 months from publication.

A consortium-based funding-model for journals' transition to OA is still being sought (Ilva, 2018). Since 2018, the Federation of Finnish Learned Societies has organized national coordination for the open science agenda in Finland, which recently produced the National policy and executive plan 2020–2025 (Open Science Coordination in Finland, 2019). The agreed objective is that "no later than 2022, all new scientific articles and conference publications will be immediately openly accessible" with CC-license, and that "the research community creates a jointly funded publishing model that enables immediate open access to research articles published in Finland."

3. RESEARCH QUESTIONS, DATA AND METHODS

Our introduction and literature review show that national bibliographic databases provide potential but have remained an underexploited information source to study OA at the national level. Given that Finland has very comprehensive CRIS data that is aggregated nationally, with standardized OA status information being included since 2016, it is a unique opportunity to explore the most central questions concerning such CRIS data from an OA perspective. Our research questions concerning Finnish peer-reviewed outputs published in 2016–2017, and aggregated at the national level in the VIRTA publication information service, are the following:

RQ1: What is the added value of institutional data for the study of OA at the national level? First, we establish the number of different types of publication channels and outputs. Second, we establish the share of outputs published in WoS and Scopus-indexed journals,

and the share of outputs that do not have DOIs. Third, we estimate what difference the additional publication data from VIRTA makes with regard to OA levels, by comparing the OA share of journal articles included in WoS and Scopus with articles not included in these databases, and by comparing the OA share of journal articles that have or do not have DOIs.

RQ2: What is the share of OA outputs across all fields, publication types, languages, and OA mechanisms? First, we establish the overall OA share of peer-reviewed outputs in different fields, and how OA share differs between journal articles, conference articles, and book publications, as well as between English, Finnish, Swedish, and other publication languages? Second, we analyze what share of journals/series and book publishers are identified in VIRTA data as gold, hybrid, and green channels? (The definition of these categories is provided below.) Third, we investigate how large a share of journals/series and book publishers have all Finnish outputs OA, have only closed outputs, or have both OA and closed outputs, and how OA level differs between gold, hybrid, and green channels.

RQ3: What is the coverage of sources for gold and green OA journals? First, we establish the total number and share of gold OA journals that can be identified based on DOAJ, the Bielefeld list, and VIRTA data. Second, we investigate the OA share of outputs in gold OA journals based on DOAJ, the Bielefeld list, and VIRTA. Third, we establish the coverage of Sherpa/Romeo color codes and the OA share of outputs in journals with different types of self-archiving policies.

RQ4: How dominant are the largest international commercial publishers? First, to establish the publishers' market shares we investigate what share of journal articles, conference articles, and book publications, as well as of outputs in different languages, are published with the six largest commercial publishing companies (Elsevier, Springer Nature, Wiley-Blackwell, Taylor & Francis, Sage, and ACS). Second, we analyze what share of outputs by these and other publishers are OA in gold, hybrid, and green publication channels. Third, we investigate the role of Finnish journal and book publishers compared to the "big" publishers.

The data consist of unique peer-reviewed outputs published in 2016–2017 that the 14 Finnish universities have reported to the Ministry of Education and Culture and that are stored in the national VIRTA publication information service (Sīle et al., 2017, 2018; Pölönen, 2018). Inclusion criteria for publications are provided by the Ministry of Education and Culture in the data collection guidelines. Universities can report all single-authored or coauthored outputs by the academic and administrative staff, including doctoral students in their service or having another contractual relationship with them (Ministry of Education and Culture, 2019b).

In VIRTA, copublications of Finnish universities appear as duplicates. However, duplicates are automatically identified on the basis of publication information and indicated in the data. In this study, we use deduplicated publication counts. For each publication, the reporting university has indicated the publication type, OECD field of science, peer review status, and open access. This study includes peer-reviewed articles in journals, books, and proceedings, as well as monographs and edited works from all fields of science.

The data for publication years 2016 and 2017 was downloaded in July 2018 from the website https://wiki.eduuni.fi/display/cscvirtajtp/Vuositasoiset+Excel-tiedostot, where the data sets for each publication year used as the basis of PRFS are openly available in Excel format. CSC—IT Center for Science—exports these data sets from VIRTA after the data collection needed for the calculation of performance-based funding is complete and makes them available on the website.

For 2016 and 2017 data, the exact date of export from VIRTA is not indicated on the website but can be estimated at June 2018. For 2017, the data collection was not yet entirely complete at the time. Missing publications and metadata for 2017 could still be added by universities in the 2019 data collection. It is important to notice that while some universities can make daily updates to their publication information in VIRTA, for example by updating the open access status of publications, such updates made after June 2018 do not show in the data sets used in this study.

The years 2016 and 2017 have been selected because universities have indicated the open availability of peer-reviewed outputs according to renewed definitions starting from 2016 (Ilva, 2017a). According to these new definitions OA publications need to meet the following criteria (Ministry of Education and Culture, 2019b):

(a) The publication can be read, printed out and copied on the Internet free of charge and in an accessible way, at least for noncommercial use.
(b) The publication is publicly available in a service offered by the publisher or the research organization that enables harvesting the publication's metadata and indexing its content for other search services and supports making references to the publication and linking it to website addresses that are based on permanent identifiers (DOI, URN, Handle).
(c) The publicly available version of the publication is either the author's final self-archived version of the publication or the final version published in the publisher's service, depending on the publication contract or the publisher's policy. If the publication is refereed, the OA version must also be refereed.

Each reported output needs to be associated with information concerning the publication being openly available immediately on the publisher's website in either a gold or hybrid OA publication channel. *Publication channel* is used as an umbrella term for serials with an ISSN as well as book publishers with ISBN roots: journals, proceedings series, book series, and imprints. Further, information regarding the output being openly available in an OA repository is also included for each publication record. Embargoed outputs are allowed as long as a stable URL to the resource is provided. Detailed information on embargo length or OA licenses, however, is not available in the data. Consequently, it is possible to establish if a peer-reviewed publication is openly available in a gold OA or hybrid channel, deposited in a repository, or both. Based on the VIRTA OA information we classify outputs into five exclusive categories:

- VIRTA gold: outputs indicated as being immediately openly available in a gold OA channel where all outputs are OA
- VIRTA hybrid: outputs indicated as being immediately openly available in a hybrid OA channel, including both OA and closed outputs
- VIRTA gold and hybrid: outputs with authors from more than one Finnish university that indicated the same output differently as being immediately openly available in a gold OA or hybrid OA channel
- VIRTA green: outputs indicated as being openly available in an OA repository and are not indicated as being openly available in a gold OA or hybrid OA channel
- VIRTA closed: outputs not indicated as being openly available in a gold OA or hybrid OA channel, or in an OA repository

These categories broadly correspond to the existing OA categories as defined, for example, by Piwowar et al. (2018), with the exception that VIRTA gold includes outputs in any channel where all outputs are immediately OA, not only outputs in DOAJ indexed journals. Thus, VIRTA gold also includes bronze OA, as well as diamond/platinum OA channels that do not charge authors

article processing charges (APCs). VIRTA hybrid and green quite closely correspond to Piwowar et al.'s (2018) definitions, and VIRTA has a similar definition of closed (this includes outputs OA in Academic Social Networks and Sci-Hub). In addition to analyzing the OA share of outputs, we also use VIRTA OA information to assess the OA status of publication channels (journals/series and book publishers).

Universities take responsibility for the OA status indicated for publications they report to the ministry. The identification of OA publications takes place at the universities and involves both researchers' self-reporting and validation by the data collection personnel from the university libraries. We know from the outset that there are some discrepancies in the identification of OA categories in the VIRTA data, as two Finnish universities may have reported the same output differently as being immediately OA in a gold or hybrid channel (the category VIRTA gold or hybrid). As Ilva (2017a) has noted earlier, the nature of the self-reported data can contain some inconsistencies that would warrant future study in detail; however, in this study we use the registered data as-is in order to obtain an unmodified baseline measurement.

In VIRTA, the publication channel—journal/series or book publisher—of each peer-reviewed output has been identified by matching the publication's bibliographic metadata to the Publication Forum authority list of publication channels. The authority list covers all journals/series and book publishers actually used by researchers affiliated with the 14 Finnish universities. Journals/series include mostly journals, but also some book series with ISSNs, as well as some conferences without ISSNs. Book publishers mostly have a registered ISBN prefix. For journals/series with ISSNs, the Publication Forum channel register contains the name of the publisher retrieved from the International ISSN Centre. We have complemented the ISSN Centre data with publisher information from the Scopus journal list. It is also indicated if the channel is included in DOAJ (DOAJ.org, 2019), the Bielefeld list of OA journals (Rimmert, Bruns, et al., 2017), and what the self-archiving policy is according to Sherpa/Romeo color codes (Sherpa.ac.uk, 2019).

4. RESULTS

4.1. The Added Value of Institutional Data

In 2016–2017, the 14 Finnish universities published 48,177 unique peer-reviewed outputs in 10,342 publication channels, of which 91.9% are journals/series and 8.1% are book publishers. Of the outputs, 83.5% are associated with journals/series, and 16.5% with the book publishers (Table 1). Of all outputs, 71.6% are journal articles, 13% proceedings articles and 15.3% are book publications. Practically all journal articles, 57.9% of proceedings articles, and 28.4% of book publications are associated with journals/series. 71.6% of the book publications are associated with book publishers.

Only 62% of the 48,177 peer-reviewed outputs are published in journals indexed in Scopus and 52% in WoS journals (Figure 1). We find that VIRTA brings added value in terms of coverage compared to WoS and Scopus in all fields, but the differences are most important in SSH fields. We also looked at DOI availability. Two-thirds (67%) of the peer-reviewed outputs have a DOI reported in VIRTA; however, DOIs are available more often for articles in journals (77%) and proceedings (60%) than for articles in books and monographs (22%). We also discovered that DOI availability is much more limited in the case of Finnish and Swedish language outputs (2.2%) than outputs in English (74.4%) and other languages (15.2%). In all, 69.6% of all OA outputs in VIRTA have a DOI. Note, however, that DOI is not a mandatory field in the data collection—as not all outputs have DOIs—so some outputs may have been reported to VIRTA without a DOI even if they might have one.

According to VIRTA data, the OA share among the 24,832 journal articles published in WoS indexed journals is 33%, while among 28,366 articles in Scopus indexed journals the OA share is

Table 1. Number of journals/series and book publishers and their share of outputs by main fields of science

	Publication channels			Outputs		
		Journals/series	Book publishers		In Journals/series	In Book publishers
Field of science	***N***	**%**	**%**	***N***	**%**	**%**
Natural sciences	3,750	95.3	4.7	15,230	89.7	10.3
Engineering	1,888	91.1	8.9	6,647	81.2	18.8
Medicine and health	2,541	98.4	1.6	10,189	98.5	1.5
Agriculture and forestry	404	93.3	6.7	900	95.1	4.9
Social sciences	3,307	89.0	11.0	10,608	72.4	27.6
Arts & humanities	1,782	78.0	22.0	5,920	64.7	35.3
All fields	10,342	91.9	8.1	48,177	83.5	16.5

35%. Among 9,675 articles published in journals not indexed in WoS and 6,141 articles in journals not indexed in Scopus the OA share is 52%. This result suggests that studies based on WoS and Scopus data may underestimate the OA share of journal articles. Comparison of OA shares between the 26,705 journal articles with DOI (38%) to the share of 7,802 articles without a DOI (37%) suggests that the availability of DOIs does not seem to make a difference with regard to OA levels.

4.2. OA Levels Across Fields, Publication Types, Languages, and OA Mechanisms

Of all 48,177 peer-reviewed outputs published in 2016–2017, one-third are reported in VIRTA as being OA (33.6%) and two-thirds are reported as being closed (66.4%; Figure 2). Overall, the differences between fields are not great. Nevertheless, Natural sciences (39.2%) and Medicine (37.2%) have the largest, while Social sciences (31%), Humanities (29.8%), and especially Engineering (26%) have the smallest shares of OA outputs.

The differences between fields are at least partly explained by differences in OA levels between publication types: The share of OA outputs is larger among journal articles (38.2%) than among conference articles (28.6%) and book publications (16.5%). The differences between the two dominant publication languages of Finnish researchers also play a role: A larger share of

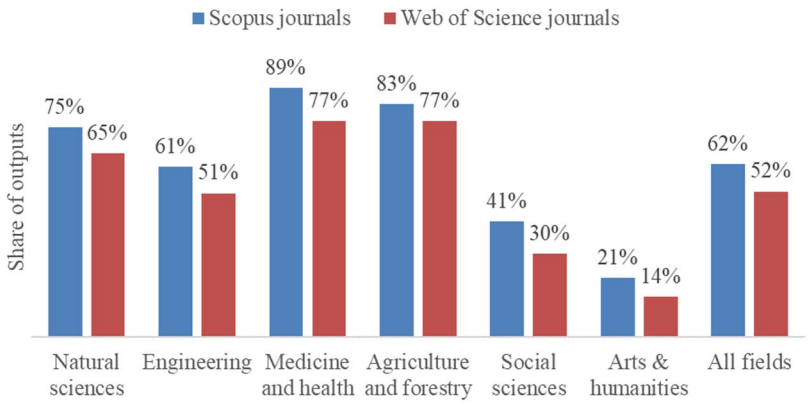

Figure 1. Scopus and WoS coverage of peer-reviewed outputs in VIRTA by field of science.

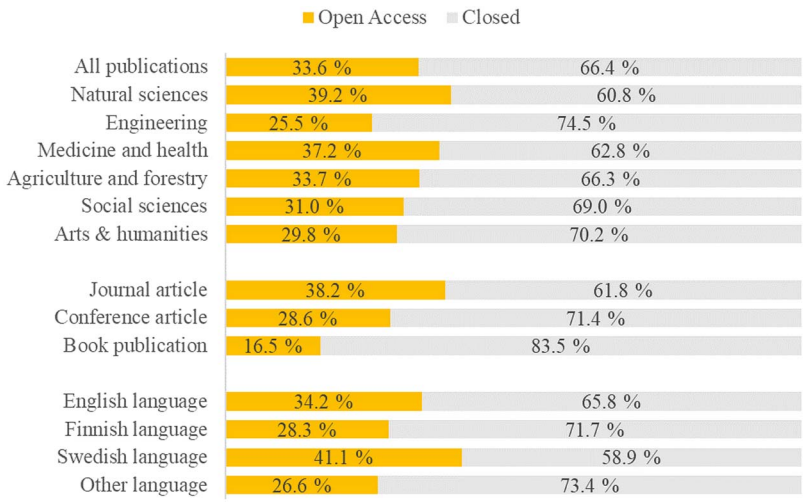

Figure 2. Open access of peer-reviewed outputs by field of science, publication type, and language.

English (34.2%) than Finnish (28.3%) language publications are OA. The numbers of publications in Swedish (OA share 41.1%), which is the other national language in Finland, and in other languages (OA share 26.6%) are much smaller. Across all fields and publication types, gold OA is the most common OA type, followed by green OA, while hybrid OA is the least common type (Table 2). There are, however, some differences in relative share of different OA types between fields, publication types, and languages.

Table 2. Type of open access of peer-reviewed outputs according to field, publication type, and language as identified in VIRTA

	Outputs	VIRTA gold	VIRTA gold or hybrid	VIRTA hybrid	VIRTA green	VIRTA closed
Field and publication type	***N***	**%**	**%**	**%**	**%**	**%**
All publications	48,177	19.3	0.2	5.2	8.9	66.4
Natural sciences	15,230	20.4	0.2	6.3	12.3	60.8
Engineering	6,647	16.3	0.1	2.7	6.3	74.5
Medicine	10,189	21.5	0.3	8.5	6.8	62.8
Agriculture	900	23.2	0.2	5.0	5.2	66.3
Social sciences	10,608	18.2	0.2	3.7	8.9	69.0
Humanities	5,920	18.8	0.1	3.2	7.7	70.2
Journal article	34,507	21.4	0.2	7.0	9.6	61.8
Conference article	6,283	17.7	0.1	1.3	9.6	71.4
Book publication	7,387	11.0	0.0	0.3	5.1	83.5
English language	42,793	19.1	0.2	5.6	9.4	65.8
Finnish language	4,280	21.6	0.0	2.1	4.6	71.7
Swedish language	411	28.5	0.5	8.0	4.1	58.9
Other language	693	17.6	0.0	3.2	5.8	73.4

Table 3. Type of open access of publication channels according to channel type and journal/series field and publisher type as identified in VIRTA

Channel type/Field/ Publisher	Publication channels	VIRTA gold channel	VIRTA gold or hybrid channel	VIRTA hybrid channel	VIRTA green channel	VIRTA closed channel
	N	%	%	%	%	%
All channels	10,342	21.9	2.8	11.0	14.6	49.8
Journals/series	9,500	21.8	3.0	12.0	14.9	48.3
- Natural sciences	3,575	20.3	4.3	15.6	17.2	42.6
- Engineering	1,720	22.1	4.7	12.2	14.7	46.3
- Medicine	2,500	21.4	4.1	20.2	12.2	42.0
- Agriculture	377	21.8	8.0	19.1	11.7	39.5
- Social sciences	2,943	23.5	4.1	10.3	18.5	43.7
- Humanities	1,390	25.6	4.6	7.3	15.9	46.5
Book publishers	842	22.1	0.7	0.5	10.6	66.2

Of the 10,342 publication channels the Finnish researchers used in 2016–2017, 21.9% are identified in VIRTA as gold OA channels, 2.8% are identified as both OA and hybrid channels (indication that their OA status is ambiguous), 11% as hybrid OA channels, and 14.6% as neither gold nor hybrid but have self-archived OA outputs (Table 3). In the case of journals/series, there are relatively small differences in the use of different OA channel types between fields: Humanities has the largest share of gold OA channels and the smallest share of hybrid OA channels.

For book publishers there is no comprehensive source on OA-status, such as DOAJ for journals, but VIRTA data can shed some light on the OA categories of 842 book publishers used by the Finnish researchers (Table 3). According to the VIRTA data, 21.1% of these publishers are gold OA channels, 0.7% have been identified as both gold and hybrid OA channels, and 0.5% have been identified as hybrid channels (0.5%). Furthermore, 10.6% of the book publishers have outputs indicated as being self-archived in an OA repository. Our analysis (below) of OA levels among books publishers identified in VIRTA with different types of OA mechanisms suggests, however, that application of OA categories—gold, hybrid, and green—is very problematic in the case of book publishers.

OA levels of publication channels and OA categories

In the VIRTA data there is some evidence of OA of outputs for about half of the 10,342 publication channels that Finnish researchers have used in 2016–2017 (Figure 3). But there is considerable variation in the share of Finnish outputs that are reported as being openly available in different channels. In roughly one-fourth of the channels (24.7%), all Finnish outputs in VIRTA are indicated as being OA; however, in one-fourth (25.5%) of the channels, the OA of the published outputs from Finland is only partial (less than 100% but more than 0% of outputs are OA). Half (49.8%) of the publication channels do not have any publications reported in VIRTA as being OA via the gold, hybrid, or green routes. This pattern is observed, more or less, in all the main fields, although the share of channels with no reported open access is somewhat larger in SSH. This is likely due to OA being more restricted in the case of book publishers than journals/series.

There is also a considerable difference in the share of openly available outputs according to the OA status of the channel based on VIRTA, as well as according to publication channel type

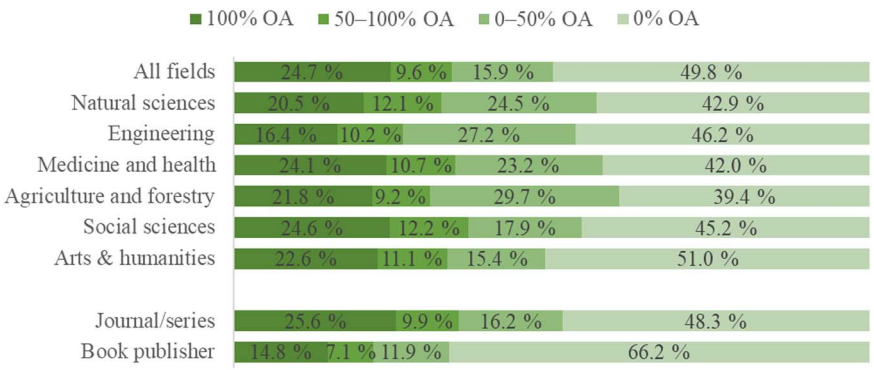

Figure 3. Share of publication channels according to the share of OA outputs by main fields of science and type of channel.

(Figure 4). The share of outputs indicated as being openly available in VIRTA is largest in the identified gold OA channels, followed by hybrid OA channels, and is smallest in green channels with only self-archived outputs. The same is observed in the case of both journal and book publishers, but the overall share of OA outputs is much smaller among book publishers.

In principle, all outputs published in gold OA channels should be immediately openly available (this is also the VIRTA definition of gold OA). The results, according to which there are gold OA channels with outputs that are not indicated in VIRTA as being openly available, suggest that some outputs have not been correctly identified as being OA, or that some of the channels have not been gold OA during the whole period of 2016–2017. The low share of OA outputs for book publishers identified as gold OA suggests that identifying OA categories is problematic for book publications (monographs and articles in books).

4.3. Coverage of Information Sources for Gold and Green Journals

DOAJ, Bielefeld, and VIRTA as sources of gold OA journals

DOAJ-indexed journals cover 12.5% of all peer-reviewed outputs, and 35.6% of outputs that are OA according to VIRTA. However, DOAJ does not cover all OA journals. Of all 9,500 journals/series used by Finnish researchers, 1,237 are gold OA journals indexed in DOAJ (Table 4). Furthermore, 372 journals/series are included in the Bielefeld list but are not indexed in DOAJ.

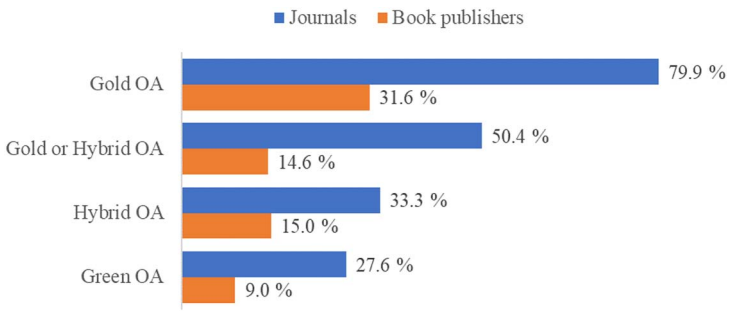

Figure 4. Share of open access outputs in journals/series and book publishers according to open access status of publication channels identified in VIRTA.

Table 4. Journal coverage of DOAJ, Bielefeld list, and VIRTA OA information, and share of open access outputs

	Publication channels	Outputs	Open access outputs	Open access outputs
	N	N	N	%
DOAJ	1,237	6,013	5,765	95.9
+Bielefeld*	372	1,249	973	77.9
VIRTA Gold-**	752	3,667	1,969	53.7
VIRTA Hybrid**	1,285	12,146	4,325	35.6
VIRTA Green**	1,388	7,300	1,995	27.3
No OA information	4,466	9,864	0	0.0
All journals/series	9,500	40,239	15,027	37.3

* Journals in Bielefeld list that are not in DOAJ.

** Journals with outputs reported in VIRTA as being openly available in gold, hybrid, or green mode and not included in DOAJ or Bielefeld list.

In addition, 752 journals/series can be identified as gold OA channels based on the VIRTA data (including gold/hybrid OA journals). Combining all three information sources it is possible to identify 2,553 potential gold OA journals, of which 48% are based on DOAJ, 15% are based on the Bielefeld list, and an additional 37% are based on VIRTA (Figure 5). This finding suggests that neither the DOAJ nor the Bielefeld list cover all gold OA journals. It is important to note, however, that it has not been possible for us to manually verify the OA status of the additional 752 journals identified as gold OA channels in VIRTA. We do not know how many of them, if any, would fulfil all the DOAJ inclusion criteria.

Analysis of VIRTA OA data suggests that the inclusion of journal/series in DOAJ is the best indicator of gold OA journals and a good predictor of OA level, as 95.9% of outputs published in DOAJ-indexed journals are actually indicated in VIRTA as being openly available (Table 4). For the Bielefeld listed journals the OA share of outputs in VIRTA is also high (77.9%), but not as high as attested in the case of DOAJ journals. The OA share of outputs published in journals/series as gold OA channels based only on VIRTA is only 54%. The OA share of outputs is considerably lower for journals identified based on VIRTA as hybrid OA (35.6%) or green OA (27.3%).

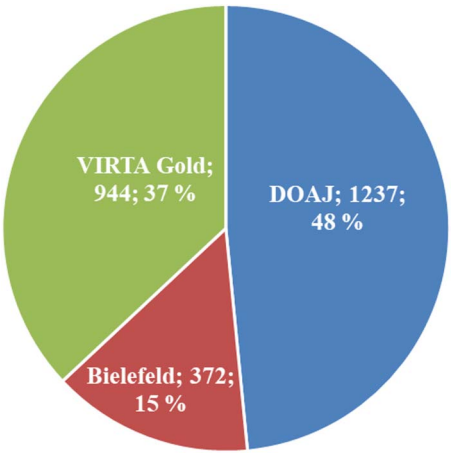

Figure 5. Share of potential gold OA journals identified based on DOAJ, Bielefeld list, and VIRTA.

Table 5. Journal coverage of Sherpa/Romeo color-codes

	Publication channels	
Sherpa/Romeo self-archiving policy	**N**	**%**
Green (publisher version)	5,034	53.0
Blue (postprint)	361	3.8
Yellow (preprint)	1,346	14.2
White (none)	267	2.8
Gray (unknown)	529	5.6
Not in Sherpa/Romeo	1,963	20.7
All journals/series	9,500	100

Sherpa/Romeo color codes

Sherpa/Romeo codes indicating self-archiving policies cover 7,537 journals/series (79% of all journals/series) used by Finnish researchers (Table 5). Sherpa/Romeo includes almost all DOAJ journals (95%), and a considerable share of Bielefeld-listed journals (43%). The self-archiving policy as indicated by the color-codes does not, however, make a great difference with regard to the OA share of outputs published in journals, especially if we look at journals/series not included in DOAJ or the Bielefeld list (Figure 6). This is because the share of OA outputs is much larger for the gold OA journals included also in DOAJ and the Bielefeld list, than for the other channels included in Sherpa/Romeo, in which OA is more dependent on self-archiving. This result is likely also valid with regard to the recently launched new version of Sherpa/Romeo, which was introduced after the analysis of this study.

4.4. Dominance of the Largest Commercial Publishers

Publication channels owned by Elsevier account for 19.4% of the 14 Finnish universities' journal outputs in all fields of science counted together (Table 6). Next come Springer Nature (13.2%), Wiley-Blackwell (9%), and Taylor & Francis (6.7%). Sage and the American Chemical Society (ACS), which are often also considered among the "big" commercial publishers, account for 2.7% and 1.9% respectively. Taken together, these publishers account for 53% of peer-reviewed journal output. In the case of peer-reviewed book publications and conference articles their dominance is weaker: 29.5% and 12% of all outputs respectively. If we take into account all publication types, the big publishers' joint share of Finnish output diminishes to less than half (44.1%).

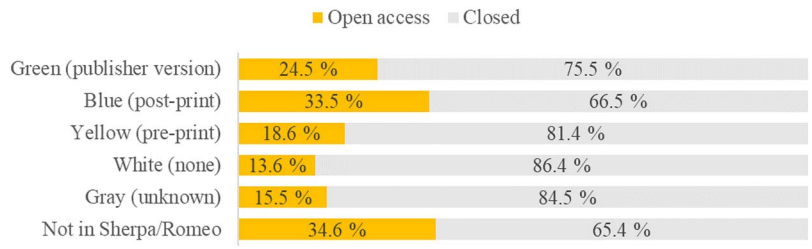

Figure 6. Sherpa/Romeo codes and share of open access outputs published in journals not included also in DOAJ and Bielefeld list.

Table 6. The six largest commercial publishers' share of outputs by field of science, publication type, and language

	Outputs	Elsevier	Springer Nature	Wiley-Blackwell	Taylor & Francis	Sage	ACS	Other
Field and publication type	***N***	**%**	**%**	**%**	**%**	**%**	**%**	**%**
All publications	48,177	14.4	12.9	6.8	6.6	2.0	1.4	55.9
Natural sciences	15,230	16.5	17.4	7.6	2.3	0.4	3.0	52.6
Engineering	6,647	22.0	9.5	4.3	2.8	1.0	2.4	58.1
Medicine and health	10,189	19.4	17.8	13.2	6.0	3.3	0.6	39.7
Agriculture and forestry	900	28.0	13.7	9.3	5.6	0.6	0.7	42.2
Social sciences	10,608	8.5	9.1	3.8	14.4	4.2	0.0	59.9
Arts & humanities	5,920	2.0	4.8	1.5	8.9	1.3	0.0	81.6
Journal article	34,507	19.4	13.2	9.0	6.7	2.7	1.9	47.0
Conference article	6,283	2.1	8.8	0.3	0.9	0.0	0.0	88.0
Book publication	7,387	1.6	15.1	1.9	10.9	0.1	0.0	70.5
English language	42,793	16.2	14.5	7.6	7.4	2.2	1.5	50.5
Finnish language	4,280	0	0	0	0	0	0	100
Swedish language	411	0	0	0	0	0	0	99.8
Other language	693	1.2	3.3	0.6	0.9	0	0	94.1

VIRTA data also suggest that the commercial publishers included in this study are most dominant in Medicine and Agriculture, and least dominant in the Social sciences and especially the Humanities. Thus, our study corroborates the findings of Larivière et al. (2015) concerning the Humanities being the field least dominated by the big publishers. In our analysis, however, Social

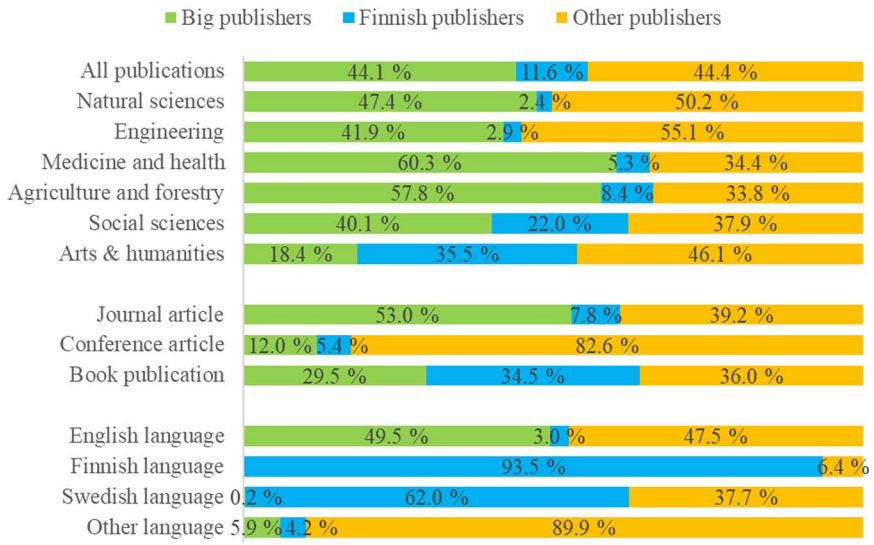

Figure 7. Share of peer-reviewed outputs published by big commercial publishers, Finnish publishers and other publishers by field of science, publication type, and language.

Table 7. Type of open access of outputs based on VIRTA by publisher

	Outputs	VIRTA gold	VIRTA gold or hybrid	VIRTA hybrid	VIRTA green	VIRTA closed
Publisher	***N***	**%**	**%**	**%**	**%**	**%**
Elsevier	6,947	6.8	0.1	6.5	11.6	75.0
Springer Nature	6,234	27.8	0.2	6.3	6.4	59.4
Wiley-Blackwell	3,259	7.1	0.1	7.4	7.9	77.5
Taylor & Francis	3,187	5.0	0.2	4.5	11.1	79.3
Sage	953	7.3	0.1	5.1	14.7	72.7
ACS	651	2.6	0.2	4.6	9.1	83.6
Finnish publishers	5,571	27.9	0.1	3.4	5.0	63.7
Other	21,380	23.8	0.2	4.8	9.3	61.9
All publishers	48,177	19.3	0.2	5.2	8.9	66.4

sciences is among the least, not the most, dominated fields (this holds true even if we limit our analysis to journal articles).

The dominance of big publishers is limited to English-language publications (Table 6), whereas in SSH research results are also communicated in languages other than English—in Finland notably in the national languages, Finnish and Swedish. VIRTA data shows the important role of Finnish journal and book publishers for scholarly communication at the national level: They account for almost 12% of Finnish universities' peer-reviewed publication output (Figure 7). They are practically the only publishers providing outlets for, and access to, research results in Finland's national languages. Their role is also particularly important in book publications (monographs, edited volumes, chapters).

The share of OA outputs is smaller for the big commercial publishers, with the exception of Springer, than the other publishers (Table 7). Outputs published with ACS and Taylor & Francis have the lowest OA levels. Among the other publishers and Springer, gold OA is the most common OA type, while hybrid and green OA are less important. In case of the other big publishers than Springer, green OA is the most common type. The Finnish publishers taken together are quite comparable to Springer Nature in terms of output size as well as OA share and type of output: They account for 11.6% and 12.9%, respectively, of the Finnish universities' peer-reviewed publication output, and 28% of their outputs are published via the gold OA route. Among the Finnish publishers, however, hybrid and green OA play a less important role. Overall, the OA share among Finnish publishers' peer-reviewed outputs (36.3%) is close to the average among all publishers (33.6%).

5. DISCUSSION

In this paper we show that it is possible to base a national estimate of the number and share of OA publications on data from institutional CRIS providing comprehensive coverage of the universities' peer-reviewed output, including all publication types and languages. In addition to the national OA estimate, it is important to investigate the institutional CRIS data also from the international perspective, because it potentially contributes to a large-scale data infrastructure needed for assessment and monitoring of OA across countries, for example at the European level.

Our data source is the VIRTA Publication Information Service, which integrates at national level publication data from the different types of commercial and noncommercial CRIS solutions of the Finnish universities. Our data set consists of 48,177 unique peer-reviewed outputs (articles in journals, proceedings, and books, as well as edited volumes and monographs) authored at the 14 Finnish universities published in 2016–2017. Based on the VIRTA data we investigated the following aspects:

(a) the added value of institutional data compared to WoS and Scopus in terms of publication output coverage
(b) the OA share of outputs across different fields, publication types and languages
(c) coverage and information value of international sources for gold and green OA journals
(d) the dominance of the largest international commercial publishers

What Is the Added Value of Institutional Data for OA Monitoring?

According to our analysis, Scopus journals cover only 62%, and WoS journals 52%, of Finnish universities' peer-reviewed outputs registered in VIRTA, ranging between 89% and 77% in Medicine to 21% and 14% in the Humanities, respectively. This demonstrates the main added value of institutional CRIS data: It is practically the only existing source that is able to provide close to complete criteria-based coverage of peer-reviewed publications of an institution across all fields, and—if integrated at the national level—of a country's higher education institutions. All the alternative information sources, notably international databases like WoS, Scopus, Google Scholar, Microsoft Academic, OpenAIRE, CrossRef, and Dimensions, have a more or less restricted or biased coverage of publications (Aksnes & Sivertsen, 2019; Martín-Martín et al., 2020; Visser et al., 2020). Notably, institutional data provides comprehensive coverage of peer-reviewed publications that are very difficult to cover in other sources: journal articles in regional and local journals, conference and book publications (chapters, edited volumes, and monographs), as well as outputs in languages other than English. Further research is needed to investigate the coverage of different information sources, including institutional data.

In addition, institutional CRIS data can provide criteria-based OA status information for all peer-reviewed publications, including outputs not included in the international databases. In the case of VIRTA data, OA status is self-reported by researchers and validated by data collection personnel at universities for all peer-reviewed outputs. Even if self-reported OA status is susceptible to inaccuracy due to individual interpretation of complicated OA mechanisms and variation in institutional data collection and validation procedures, CRIS data can offer an alternative and complementary methodology for determining OA status of publications. This is important, as only 67% of the peer-reviewed outputs in VIRTA had a reported DOI, which is a requirement for inclusion in Unpaywall and, hence, analyses based thereon. Our analysis shows that DOI availability is particularly limited in the case of book publications (22%) and those in Finland's national languages (3%). In all, DOIs cover 69.6% of all OA outputs as reported in VIRTA, and DOAJ-indexed journals only 35.6% of the OA outputs identified based on VIRTA, so these two methodologies, which are frequently used to identify OA outputs, would lead to a partial picture of OA in Finland. In our view, it is important that national scholarly publishers of journals and books operating in languages other than English also seek inclusion in the DOAJ and Sherpa/Romeo services, as well as making use of DOIs and submitting as rich metadata as possible to Crossref.

In addition to providing a more complete picture of OA at the national level, institutional CRIS data can also be used to study and understand representativeness and bias in the OA measurements based on less comprehensive international sources and different methodologies for OA

assessment and monitoring. According to VIRTA data, 33% of the Finnish articles published in WoS journals and 35% in Scopus journals are OA. This result, based on researcher self-reports validated by the data-collection personnel at universities, is fairly close to OA levels established for Finland in some previous studies: 32% in 2016 based on WoS in Bosman and Kramer (2018) and 41.6% in 2017 based on Scopus in the European Open Science Monitor. The higher OA share in the European Open Science Monitor could be due to the fact that OA shares are rapidly increasing in Finland via the hybrid and green routes (Ilva, 2020), and not all outputs that are currently OA in repositories had been self-archived or openly available at the time our data was reported to VIRTA (see section 3). Nevertheless, our analysis shows that the OA share is much higher, 52%, among peer-reviewed articles published in journals not indexed in WoS and Scopus. This finding suggests that OA monitoring based on WoS and Scopus is not only based on a limited subset of publications, but may also underestimate the OA share of journal articles at country level.

What Is the Share of OA Outputs across all Fields, Publication Types, Languages, and OA Mechanisms?

Taking all peer-reviewed outputs published in 2016–2017 into account, the share of OA at the national level in Finland based on VIRTA data is 33.6%, ranging from 39.2% in the natural sciences to 25.5% in engineering. It is difficult to compare this result directly with international studies because OA levels can change quite rapidly at the country level due to national and institutional policies, incentive structures, and services for promoting OA. According to a recent analysis by Ilva (2020), which is also based on VIRTA data, the OA share of journal, conference, and book articles has more than doubled from 28% in 2016 to 65% in 2019. Another challenge is that OA definitions may differ between studies using different methodologies, and the selection of peer-reviewed publications may also differ according to the data source used.

Our analysis also shows that OA shares differ considerably between different publication types and to a lesser extent between publication languages. Overall, the share of OA is larger among journal articles (38.2%) than among conference articles (28.6%) and book publications (16.5%). Our analysis also shows that the two dominant publication languages of the Finnish universities are English and Finnish (covering 88.8% and 8.9% of all outputs), and that a larger share of English (34.2%) than Finnish (28.3%) language publications are OA. Thus, the somewhat lower OA share in Engineering is explained at least partly by the importance of conference articles, while book publications and Finnish language publications contribute to a lower OA share in the SSH fields.

VIRTA data also contain information about the OA mechanism based on the publication channel, as it is reported for each publication if it is openly available immediately in the publisher's website in either gold or hybrid OA channel, and if it has been self-archived in an OA repository. Overall, gold OA channel is across all fields the most dominant OA route accounting for 19.3% of peer-reviewed outputs. In addition, 5.4% of outputs are OA in a hybrid OA channel, and 8.9% are OA only via repositories. It may also contribute to the lower OA share in Engineering and SSH fields that the gold and hybrid OA channels appear to be less used, perhaps because gold OA journals are considered less prestigious or perhaps due to limited resources for APCs and additional OA fees.

One of the advantages of the institutional CRIS data is that it shows the complete picture of journal and book publishing profile as well as the role of OA journals and book publishers. During 2016–2017, researchers at the Finnish universities used 10,342 different publication channels as outlets for their research, including 9,500 journals/series and 842 book publishers. In 25% of the channels used, all Finnish outputs are reported as being OA in VIRTA, in 25% of the channels

only part of the outputs are OA, and in 50% of the channels no OA outputs via gold, hybrid, or green routes were reported in VIRTA. The same pattern is observed, more or less, across all fields.

While countries strive to achieve national and international OA targets, the strategies often involve changing the publishing landscape by means of replacing currently used closed channels with gold, hybrid, and green OA channels, or by making those channels allow different OA routes. Different OA routes indeed lead to quite different OA levels, as the OA share of articles is 79.9% in journals identified in VIRTA as gold OA channels, while being only 33.3% in hybrid and 27.6% in green OA journals. The same pattern is visible also in the case of book publishers; however, the overall share of OA outputs is much smaller than among journals: only 31.6% in the case of book publishers identified in VIRTA as gold OA, and 15% and 9% respectively for hybrid and green book OA publishers.

As we pointed out above, the quality of self-reported OA status of outputs is subject to doubt with regard to correctness. This is because a large number of researchers reporting OA status may interpret and understand OA mechanisms differently, and validation of the reported OA information by data-collection personnel may work differently in different organizations and units (Azeroual & Schöpfel, 2019; van Leeuwen et al., 2016). Our analyses indeed highlight certain inconsistencies in the self-reporting and validation of the OA information to VIRTA. There is, first, uncertainty about the OA mechanisms of the publication channels, as some channels have been identified differently as supporting gold or hybrid OA routes. Second, all outputs published channels categorized as gold OA in VIRTA should be immediately openly available. Yet our analysis shows that there are channels identified as gold OA with outputs that are not indicated in VIRTA as being openly available. Further research is needed to investigate the accuracy of self-reported OA status of outputs, and to compare results for example with Unpaywall.

There are several possible explanations for the observed discrepancies in the OA status of publications and channels: Some outputs have simply been incorrectly identified as OA, some hybrid channels mistaken for gold OA channels, and some channels may not have not been gold OA during the entire period of 2016–2017. The low share of OA outputs for book publishers identified as gold OA suggests that identifying OA categories is particularly problematic for book publications (monographs and articles in books). One important aspect to consider is also that while OA policies have mostly focused on journals, OA mechanisms and definitions for book publications remain underdeveloped and there are no comprehensive international information sources that researchers and data-collection personnel could use to identify gold, hybrid, and green OA book publishers. Our findings highlight the need for an international register of academic/scholarly book publishers that would contain information—like DOAJ and Sherpa/Romeo—on their peer-review practices, as well as open access status and self-archiving policies (Giménez-Toledo, 2020).

What Is the Coverage of Information Sources for Gold and Green OA Journals?

The DOAJ and Sherpa/Romeo are information sources frequently used by researchers, libraries, and policy-makers to identify gold and green OA journals. Indeed, research funders behind the Plan S initiative also rely on DOAJ and Sherpa/Romeo as international services to identify high-quality gold and green OA channels. The Bielefeld list is also increasingly used in libraries as an information source for gold OA journals. According to our analysis, however, neither DOAJ nor the Bielefeld list provides a complete picture of gold OA publishing (see also Björk, 2019; Bruns et al., 2019). Among the journals used by the Finnish researchers we identified 2,553 potential gold OA journals, of which DOAJ covers 48%, the Bielefeld list an additional 15%, and 37% are

identified based on VIRTA data. Most gold OA journals identified based on the Bielefeld list and especially VIRTA may not, however, fulfil all the DOAJ inclusion criteria that are set to qualify journals following the best international standards and practices of gold OA publishing.

Our analysis of the VIRTA OA data suggests that inclusion of journals in DOAJ is the best indicator of gold OA journals and a good predictor of high OA level, as 95.9% of outputs in DOAJ-indexed journals are OA also in VIRTA. For the Bielefeld-listed journals the OA share of outputs is also high, 77.9%, while for the journals identified as gold OA based only on VIRTA it is only 54%. This finding indeed suggests that identification of gold OA journals based on self-reported OA information in VIRTA is less reliable than DOAJ and the Bielefeld list (probably some hybrid journals are mistakenly identified as gold OA journals), or that researchers and data-collection personnel rely mostly on DOAJ for identification of gold OA status of journals.

The majority of journals/series used by the Finnish researchers (79%) have a self-archiving policy registered in Sherpa/Romeo. Analysis of outputs published in these journals shows that only a relatively small share is indicated as OA in VIRTA, irrespective of the self-archiving policy indicated by color-coding, unless the journal also provides OA via the gold route (DOAJ-indexed or Bielefeld-listed journals). Part of this observation can likely be due to color codes having become less useful for summarizing journal self-archiving policies, as a lot of additional restrictions have been introduced by publishers (Gadd & Troll Covey, 2016). Part is likely due to unused potential for permitted self-archiving (Björk et al., 2014; Laakso, 2014). Our results confirm that there is indeed considerable potential for advancing OA via the green route. It remains to be seen if OA incentives, such as the extra weight for open access publications in the Finnish universities' core funding model, might help to increase self-archiving activity. This development can be comprehensively monitored across all higher education institutions, fields, and publication types only by using the VIRTA data (see Ilva, 2020). The new Sherpa/Romeo service, in which the color codes have been replaced with information on the availability and conditions of the different OA routes, provides an information source for the identification of gold, hybrid, and green journals.

How Dominant Are the Largest International Commercial Publishers?

Most attention in the national and international OA policies is focused on the large international commercial publishers—Elsevier, Springer Nature, Wiley-Blackwell, Taylor & Francis, Sage, and ACS—that according to recent analyses cover more than half of the international journal publishing indexed in WoS (Larivière et al., 2015). Our analysis based on the VIRTA data shows that the "big" publishers also play a dominant role in Finland, accounting for 53% of peer-reviewed journal output and 44% of all outputs, including conferences and book publications (cf. Guns, 2018). Thus, our analysis of the VIRTA data suggests that WoS data, focused on an international subset of journal articles, somewhat exaggerates the role played by the big publishers. This is seen most clearly in the case of the social sciences, which according to Larivière et al. (2015) is among the fields most dominated by the big publishers. According to VIRTA data, based of course only on outputs from Finland, the social sciences are, together with humanities, the fields least dominated by the big publishers.

In this study we were able to contrast, because of the comprehensive coverage of peer-reviewed outputs in VIRTA, the output of big publishers with that of the small-scale and not-for-profit journal and book publishers operating in Finland (Late, Korkeamäki, et al., 2020): Their combined output amounts to 11.6% of the Finnish universities' peer-reviewed publications. Thus, the Finnish publishers' share is comparable in size to some of the largest international commercial publishers, such as Elsevier (14.4%), Springer Nature (12.9%), Wiley-Blackwell (6.8%), and Taylor & Francis (6.6%). The national publishers are used in all fields, but their role is

especially important in the SSH. In the humanities, the share of outputs published with Finnish publishers (35.5%) is even larger than that of the "big" publishers (18.4%). They play a unique role in the scholarly communication of Finnish researchers by publishing peer-reviewed research in the national languages, and they also play a major role in publishing scholarly books.

Transformative "read-and-publish" agreements with the largest international publishers can significantly advance OA at the national level, including in Finland. Yet, it is important to remember that these are only a partial solution. In all fields, and especially in the SSH, the advancement of OA also requires that gold, hybrid, and green OA publishing models are also adopted by a large variety of relatively small journal and book publishers operating in international and national contexts. In Finland, most OA journals do not charge authors APCs. A new Diamond Open Access study commissioned by cOAlition S is creating a global overview of this OA publishing model. Our analysis of the publisher shares strongly suggests that the Finnish research community cannot meet the international and national OA targets if immediate open access to peer-reviewed content is not secured in a sustainable way for journals and books published in Finland (Ilva, 2018).

Contribution of Institutional CRIS Data to International Publication Infrastructure

Finally, we discuss the potential for using institutional data in cross-country comparisons, notably in monitoring publication activities and open science at the European level. As the European Commission has noted, the current conditions for constructing the European Open Science Monitor, based on the data provided by Elsevier (Tennant, 2018), are nonoptimal: "Overall, the Commission wishes to have an as comprehensive Monitor as possible. … as long as there is in the European Union no fully open and transparent data-infrastructure, we are dependent on a fragmented data infrastructure and data sources from private operators" (cited in Waltman, 2019, p. 5).

Our study confirms that Elsevier's Scopus provides only limited and biased coverage of the publications of Finnish universities. OpenAIRE and Crossref are in our view important building blocks of a comprehensive large-scale European infrastructure for publication information that is independent of private operators. In this study, we did not directly compare coverage of the VIRTA data with OpenAIRE or Crossref, but our findings strongly suggest that their coverage of peer-reviewed outputs across fields, publication types, and languages is far from comprehensive. This is because OpenAIRE mainly depends on the availability of documents in OA repositories, and Crossref on publishers using DOIs. The added value of integrated CRIS data is that it can provide well-structured and curated metadata, including OA information, of all peer-reviewed publications even if they are not included in WoS and Scopus, are not in digital format, do not have DOIs, and are not openly available on the internet.

Institutional publication data, which in many countries is already integrated at the national level in services such as VIRTA, is the only source of publication information to complement OpenAIRE and Crossref with a comprehensive picture of European research and open access development across all fields, publication types and languages. According to recent surveys, over 20 European countries already have national publication databases that go beyond WoS and Scopus (Sīle et al., 2017, 2018), and hundreds of universities and research organizations have institutional CRIS systems, from which publication information could be integrated to an international infrastructure. Ideally, an international infrastructure should also offer countries and institutions without CRIS a service for inputting their publication information (Puuska, Nikkanen, et al., 2020). In addition to being comprehensive, institutional data is independent of private operators, and governments, institutions, and researchers across Europe already invest much time, effort, and resources in producing it.

The need for a comprehensive European infrastructure for publication data has been called for during the past decade in several policy documents (European Commission, 2010; Lauer, 2016; see discussion in Sivertsen, 2019). In 2014, a report to the European Parliamentary Research Service recommended "the development of a European integrated research information system … having features of a distributed infrastructure, inter-connecting the existing national research information systems" (Mahieu et al., 2014). A proof of concept of a European publication infrastructure integrating data from six institutions across four different countries has already been carried out in the framework of EU COST-Action ENRESSH (www.enressh.eu). Nevertheless, there is still a lot of work to be done to improve the standardization and interoperability of CRIS data to build large-scale international solutions that can compete with commercial bibliometric databases (Puuska et al., 2018). It is an additional challenge to produce comprehensive and comparable OA information on all types of outputs. Self-reports by researchers and validation by data-collection personnel, such as used in VIRTA, offer one possible solution to complement other information sources, such as DOAJ, Sherpa/Romeo, Unpaywall, and OpenAIRE.

Yet there are important policy reasons for European stakeholders to further invest in the development of comprehensive publication data. "Open Science" in the title of the European Monitor entails a broad understanding of research impact. The main international responsible metrics statements endorsed by the European Open Science agenda—DORA (https://sfdora.org/), The Leiden Manifesto (Hicks, Wouters, et al., 2015), Metric Tide (Wilsdon, Allen, et al., 2015)—call for diversity of outputs to be taken into account in research evaluation (European Commission, 2018). EU policies for the Responsible Research and Innovation (RRI) promote broad access to research, interaction between science and society, and public understanding of science (Gerber, Forsberg, et al., 2020; Novitzky, Bernstein, et al., 2020). This requires that many different output types and languages are used in the dissemination of research results to all sectors of society (Sivertsen, 2018b). As the European University Association (EUA) states in support of the Helsinki Initiative on Multilingualism in Scholarly Communication (www.helsinki-initiative.org), "Multilingualism is particularly relevant for Europe, as its research is characterized by geographic, cultural and linguistic diversity and the common principle of excellence" (EUA, 2019; Kulczycki et al., 2020). We argue that the large-scale data infrastructure for monitoring Open Science at the European level should reflect its geographic, cultural, and linguistic diversity. Only institutional publication data, integrated at the national and international levels, can provide the needed comprehensiveness.

6. CONCLUSIONS

In this paper we show that institutional publication data provides an invaluable information source in terms of output coverage for assessing the number and share of OA publications at the national level—in our case, Finland. We also argue that institutional data should be used to complement other information sources—such as OpenAIRE and Crossref—in OA monitoring across countries (e.g., at the European level). This is important for two reasons: First, institutional publication data, integrated at the national and international levels, are the only source that can provide a comprehensive picture of European research and OA development across all fields, publication types, and languages. In addition, such data can also be used to analyze and test the representativeness of OA assessments based on less comprehensive international sources, such as WoS, Scopus, Google Scholar, Microsoft Academic, Dimensions, CrossRef, and OpenAIRE.

Compared to earlier studies contributing towards national-level OA measurement the methodology of this study is unique, avoiding the limitations of using only WoS or Scopus-indexed journal publications like van Leeuwen et al. (2018) and Martín-Martín et al. (2018), and at the

same time including investigations of article-level OA mechanisms through either self-reported or matching to external OA information sources that have been missing from earlier CRIS-data based studies (e.g., Mikki, 2017; Mikki et al., 2018). Beyond this there is still unused potential for future research and practice to improve the flexibility, fidelity, and reliability of the self-reported OA data as well as exploring the use of additional external data sources for OA detection, such as Unpaywall. With the publication data environment still being fragmented and under constant development, the best OA data can likely be produced by matching top-down and bottom-up approaches to identification.

We conclude that national publication data provide valuable and unique information on OA of peer-reviewed outputs. To enhance comprehensive and comparable monitoring of OA we recommend the development of well-structured and comprehensive national and international publication information sources, something which should be seen as integral to working towards open science in both policy and practice (Biesenbender et al., 2019).

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable feedback, which has helped to improve this article.

AUTHOR CONTRIBUTIONS

Janne Pölönen: Conceptualization, Methodology, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Supervision, Project administration. Mikael Laakso: Conceptualization, Methodology, Investigation, Writing—original draft, Writing—review & editing, Visualization. Raf Guns: Conceptualization, Methodology, Writing—review & editing. Emanuel Kulczycki: Conceptualization, Methodology, Writing—review & editing. Gunnar Sivertsen: Conceptualization, Methodology, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

The authors declare no funding information.

DATA AVAILABILITY

The data set used for this paper is available with a CC BY 4.0 license and can be downloaded from the following location: https://figshare.com/s/a6479fc58131fd647cdb. The original base publication data was downloaded from https://wiki.eduuni.fi/display/cscvirtajtp/Vuositasoiset+Excel-tiedostot, where also older and more recent data sets are available for download.

REFERENCES

Aagaard, K. (2018). Performance-based research funding in Denmark: The adoption and translation of the Norwegian model. *Journal of Data and Information Science*, *3*(4), 20–30. **DOI:** https://doi.org/10.2478/jdis-2018-0018

Academy of Finland. (2019). *Usein kysyttyjä kysymyksiä julkaisujen avointa saatavuutta edistävästä Plan S aloitteesta*. Accessed June 16, 2019. https://web.archive.org/web/20190616104556/https://www.aka.fi/fi/akatemia/media/Ajankohtaiset-uutiset/2019/tutkimusjulkaisujen-avoimuutta-edistavan-plan-s–suunnitelman-toimeenpano-etenee/usein-kysyttyja-kysymyksia-plan-s–aloitteesta/

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, *4*(1), 1–21. **DOI:** https://doi.org/10.2478/jdis-2019-0001

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and

humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. **DOI:** https://doi.org/10.1007/s11192-006-0115-z
Azeroual, O., & Schöpfel, J. (2019). Quality issues of CRIS data: An exploratory investigation with universities from twelve countries. *Publications*, *7*(1), 14. **DOI:** https://doi.org/10.3390/publications7010014
Biesenbender, S., Petersohn, S., & Thiedig, C. (2019). Using Current Research Information Systems (CRIS) to showcase national and institutional research (potential): Research information systems in the context of Open Science. *Procedia Computer Science*, *146*, 142–155. **DOI:** http://doi.org/10.1016/j.procs.2019.01.089
Björk, B.-C. (2019). Open access journal publishing in the Nordic countries. *Learned Publishing*, *27*(2), 227–236. **DOI:** http://doi.org/10.1002/leap.1231
Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the American Society for Information Science and Technology*, *65*(2), 237–250. **DOI:** https://doi.org/10.1002/asi.22963
Bosman, J., & Kramer, B. (2018). Open access levels: A quantitative exploration using Web of Science and oaDOI data. *PeerJ Preprints*, *6*, e3520v1. **DOI:** https://doi.org/10.7287/peerj.preprints.3520v1
Bosman, J., & Kramer, B. (2019). Publication cultures and Dutch research output: a quantitative assessment. *Zenodo*. **DOI:** http://doi.org/10.5281/zenodo.2643360
Boudry, C., & Chartron, G. (2017). Availability of digital object identifiers in publications archived by PubMed. *Scientometrics*, *110*(3), 1453–1469. **DOI:** https://doi.org/10.1007/s11192-016-2225-6
Bruns, A., Lenke, C., Schmidt, C., & Taubert, NC. (2019). *ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA) 3.0*. Bielefeld University. **DOI:** https://doi.org/10.4119/unibi/2934907
Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'? *Research Evaluation*, *27*(2), 106–118. **DOI:** https://doi.org/10.1093/reseval/rvy001
Council of the European Union. (2016). Outcome of the Council meeting, *3470th Council meeting,* Competitiveness (Internal Market, Industry, Research and Space) Brussels, 26 and 27 May 2016. Retrieved January 15, 2020 from: https://web.archive.org/web/20200115073417/https://www.consilium.europa.eu/media/22779/st09357en16.pdf
Crawford, W. (2019). *Gold Open Access 2013–2018: Articles in Journals (GOA4)*. Livermore, CA: Cites & Insights Books. Accessed January 15, 2020. https://web.archive.org/web/20190618090950/https://waltcrawford.name/goa4.pdf
den Hertog, P., Jager, C.-J., Vankan, A., te Velde, R., Veldkamp, J., … van Wijk, E. (2014). Scholarly publication patterns in the social sciences and humanities and their relationship with research assessment. *Science, Technology and Innovation Indicators*, Thematic Paper 2. Utrecht, Netherlands.
doabooks.org. (2019). *Directory of Open Access Books*. Accessed June 14, 2019. https://web.archive.org/web/20190614062351/https://www.doabooks.org/
DOAJ. (2019). *New Pilot to Encourage Finnish Open Access Journals to Apply to DOAJ*. Accessed September 21, 2019. https://web.archive.org/web/20190921021829/https://blog.doaj.org/2019/09/02/new-pilot-to-encourage-finnish-open-access-journals-to-apply-to-doaj/
DOAJ.org. (2019). *The Directory of Open Access Journals*. https://doaj.org/
Engels, T. C. E., & Guns, R. (2018). The Flemish performance-based research funding system: A unique variant of the Norwegian model. *Journal of Data and Information Science*, *3*(4), 45–60. **DOI:** https://doi.org/10.2478/jdis-2018-0020
Engels, T., Starčič, A., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management*, *70*(6), 592–607. **DOI:** https://doi.org/10.1108/AJIM-05-2018-0127
Etsin. (2018). *Academic Publisher Costs in Finland 2010–2017*. Ministry of Education and Culture of Finland and its Open Science and Research Initiative 2014–2017. https://etsin.avointiede.fi/dataset/urn-nbn-fi-csc-kata20180822134234600198
EUA. (2019). *Multilingualism in Scholarly Communication: Endorsement of Helsinki Initiative*. https://eua.eu/news/341:multilingualism-in-scholarly-communication-endorsement-of-helsinki-initiative.html
European Commission. (2010). *Assessing Europe's University-Based Research*. Accessed March 17, 2019. https://web.archive.org/web/20190317171523/http://ec.europa.eu/research/science-society/document_library/pdf_06/assessing-europe-university-based-research_en.pdf
European Commission. (2018). *OSPP-REC: Open Science Policy Platform Recommendations*. Accessed November 30, 2019. https://web.archive.org/web/20191130085613/http://ec.europa.eu/research/openscience/pdf/integrated_advice_ospp_recommendations.pdf
European Commission. (2019). *Trends for Open Access to Publications*. Accessed November 30, 2019. https://web.archive.org/web/20191130085030/https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en
European Open Science Monitor. (2018). *Updated Methodological Note*. Accessed June 29, 2019. https://web.archive.org/web/20190629064251/https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_science_monitor_methodological_note_april_2019.pdf
Fasae, J. K., & Oriogu, C. D. (2018). Digital Object Identifier and their use in accessing online scholarly materials in Africa. *Library Philosophy and Practice*. 1785.
FinElib. (2019). *FinELib—Our Goals*. Accessed June 15, 2019. https://web.archive.org/web/20190615083653/http://finelib.fi/negotiations/goals/
Gadd, E., & Troll Covey, D. (2016). What does "green" open access mean? Tracking twelve years of changes to journal publisher self-archiving policies. *Journal of Librarianship and Information Science*, *51*(1), 106–122. **DOI:** http://doi.org/10.1177/0961000616657406
Gerber, A., Forsberg, E.-M., Shelley-Egan, C., Arias, R., Daimer, S., … Steinhaus, N. (2020). Joint declaration on mainstreaming RRI across Horizon Europe. *Journal of Responsible Innovation*. **DOI:** https://doi.org/10.1080/23299460.2020.1764837
Giménez-Toledo, E. (2020). Why books are important in the scholarly communication system in social sciences and humanities. *Scholarly Assessment Reports*, *2*(1), 6. **DOI:** http://doi.org/10.29024/sar.14
Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T. C. E., Ingwersen, P., Pölönen, J., … Zuccala, A. A. (2016). Taking scholarly books into account: Current developments in five European countries. *Scientometrics*, *107*(2), 685–699. **DOI:** https://doi.org/10.1007/s11192-016-1886-5
Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T., Guns, R., Kulczycki, E., … Zuccala, A. (2019). Taking scholarly books into account, part II: A comparison of 19 European countries in evaluation and funding. *Scientometrics*, *118*(1), 233–251. **DOI:** https://doi.org/10.1007/s11192-018-2956-7
Giménez-Toledo, E., Mañana-Rodríguez, J., & Sivertsen, G. (2017). Scholarly book publishing: Its information sources for evaluation

in the social sciences and humanities. *Research Evaluation*, *26*(2), 91–101. **DOI:** https://doi.org/10.1093/reseval/rvx007
Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. **DOI:** https://doi.org/10.1016/j.joi.2015.11.008
Guns, R. (2018). Concentration of academic book publishers. In R. Costas et al. (Eds.), *Proceedings of the 23rd International Conference on Science and Technology Indicators* (pp. 518–525). https://openaccess.leidenuniv.nl/handle/1887/65268
Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215. **DOI:** https://doi.org/10.1007/BF02457380
Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. **DOI:** https://doi.org/10.1016/j.respol.2011.09.007
Hicks, D., & Wang, J. (2011). Coverage and overlap of the new social science and humanities journal lists. *Journal of the American Society for Information Science and Technology*, *62*(2), 284–294. **DOI:** https://doi.org/10.1002/asi.21458
Hicks, D., Wouters, P. F., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics: Use these 10 principles to guide research evaluation. *Nature*, *520*/7548, 429–431. **DOI:** https://doi.org/10.1038/520429a
Himmelstein, D. S., Romero, A. R., Levernier, J. G., Munro, T. A., McLaughlin, S. R., … Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *eLife*, *7*, e32822. **DOI:** https://doi.org/10.7554/eLife.32822
Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., & Hosking, R. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. **DOI:** https://doi.org/10.1162/qss_a_00031
Ilva, J. (2017a). Towards reliable data – counting the Finnish Open Access publications. *Procedia Computer Science*, *106*, 299–304. **DOI:** https://doi.org/10.1016/j.procs.2017.03.029
Ilva, J. (2017b). Suomalaisten yliopistojen avoimet julkaisut vuonna 2016 OKM:n julkaisutiedonkeruun tietojen valossa. *Informatiotutkimus*, *36*(3–4), 51–69. **DOI:** https://doi.org/10.23978/inf.68913
Ilva, J. (2018). Looking for commitment: Finnish open access journals, infrastructure and funding. *Insights*, *31*, 25. **DOI:** https://doi.org/10.1629/uksg.414
Ilva, J. (2020). Open access on the rise at Finnish universities. Think Open blog: https://blogs.helsinki.fi/thinkopen/oa-statistics-2019/
Impactstory. (2017). oaDOI FAQ. Accessed January 8, 2018. https://web.archive.org/web/20180108192808/https://oadoi.org/faq
Kaltenbrunner, W., & de Rijcke, S. (2016). Quantifying 'output' for evaluation: Administrative knowledge politics and changing epistemic cultures in Dutch law faculties. *Science and Public Policy*, *44*(2), 1–10. **DOI:** https://doi.org/10.1093/scipol/scw064
Kronman, U. (2017). *Open Access i SwePub 2010–2016*. Accessed December 16, 2017. https://web.archive.org/web/20171216091206/http://openaccess.blogg.kb.se/files/2017/12/Open_Access_i_SwePub_2010-2016_v1.pdf
Kulczycki, E., Engels, T., Pölönen, J., Bruun, K., Duskova, M., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries, *Scientometrics*, *116*(1), 463–486. **DOI:** https://doi.org/10.1007/s11192-018-2711-0
Kulczycki, E., Guns, R., Pölönen J., Engels, T., Rozkosz, E., … Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*. 1–15. **DOI:** https://doi.org/10.1002/asi.24336
Kulczycki, E., & Korytkowski, P. (2018). Redesigning the model of book evaluation in the Polish performance-based research funding system. *Journal of Data and Information Science*, *3*(4), 61–73. **DOI:** https://doi.org/10.2478/jdis-2018-0021
Laakso, M. (2014). Green open access policies of scholarly journal publishers: A study of what, when, and where self-archiving is allowed. *Scientometrics*, *99*(2), 475–494. **DOI:** https://doi.org/10.1007/s11192-013-1205-3
Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLOS One*, *10*(6), e0127502. **DOI:** https://doi.org/10.1371/journal.pone.0127502
Larivière, V., & Macaluso, B. (2011). Improving the coverage of social science and humanities researchers' output: The case of the Érudit journal platform. *Journal of the American Society for Information Science and Technology*, *62*(12), 2437–2442. **DOI:** https://doi.org/10.1002/asi.21632
Late, E., Korkeamäki, L., Pölönen, J., & Syrjämäki, S. (2020). The role of learned societies in national scholarly publishing. *Learned Publishing*, *33*(1), 5–13. **DOI:** https://doi.org/10.1002/leap.1270
Lauer, G. (2016). The ESF Scoping Project 'Towards a Bibliometric Database for the Social Sciences and Humanities'. In M. Ochsner, S. E. Hug, and H. D. Daniel (Eds.). *Research Assessment in the Humanities. Towards Criteria and Procedures* (pp. 73–77). Zürich: Springer Open. **DOI:** https://doi.org/10.1007/978-3-319-29016-4_6
Mahieu, B., Arnold, E., & Kolarz, P. (2014). *Measuring Scientific Performance for Improved Policy Making*. Brussels: European Parliamentary Research Service. Accessed May 27, 2019. https://web.archive.org/web/20190527064957/http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/527383/IPOL-JOIN_ET(2014)527383(SUM01)_EN.pdf
Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841.**DOI:** https://doi.org/10.1016/j.joi.2018.06.012
Martín-Martín, A., Thelwall, M, Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. https://arxiv.org/ftp/arxiv/papers/2004/2004.14329.pdf
Mikki, S. (2017). Scholarly publications beyond pay-walls. *Increased citation advantage for open publishing*. *Scientometrics*, *113*, 1529–1538. https://doi.org/10.1007/s11192-017-2554-0
Mikki, S., Gjesdal, Ø. L., & Strømme, T. E. (2018). Grades of openness: Open and closed articles in Norway. *Publications*, *6*(4), 46. **DOI:** https://doi.org/10.3390/publications6040046
Ministry of Education and Culture. (2019a). *Universities, Core Funding Model from 2021*. Accessed January 15, 2020. https://minedu.fi/documents/1410845/4392480/UNI_core_funding_2021.pdf/a9a65de5-bd76-e4ff-ea94-9b318af2f1bc/UNI_core_funding_2021.pdf
Ministry of Education and Culture. (2019b). *Publication Data Collection Instructions for Researchers 2019*. https://wiki.eduuni.fi/display/cscsuorat/Julkaisutiedonkeruun+tutkijaohjeistukset?preview=/39984924/112497067/Publication%20data%20collection%20instructions%20for%20researchers%202019.docx
Mongeon, P., & Paul-Hus, A. (2015). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. **DOI:** https://doi.org/10.1007/s11192-015-1765-5
Nederhof, A. J. (1989). Books and chapters are not to be neglected in measuring research productivity. *American Psychologist*, *44*(4), 734–735. **DOI:** https://doi.org/10.1037/0003-066X.44.4.734

Nederhof, T. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, *66*(1), 81–100. **DOI:** https://doi.org/10.1007/s11192-006-0007-2

Novitzky, P., Bernstein, M. J., Blok, V., Braun, R., Tung Chan, T., … Griessler, E., (2020). Improve alignment of research policy and societal values. *Science*, *369*(6499), 39–41. **DOI:** https://doi.org/10.1126/science.abb3415

Open Science Coordination in Finland. (2019). *Open Access to Scholarly Publications. National Policy and Executive Plan by the Research Community in Finland for 2020–2025 (1)* (2nd edn). Responsible Research Series 3: 2019. **DOI:** https://doi.org/10.23847/isbn.9789525995343

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science—a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation*, *21*(4), 280–290. **DOI:** https://doi.org/10.1093/reseval/rvs019

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., … Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. **DOI:** https://doi.org/10.7717/peerj.4375

Pölönen, J. (2018). Applications of, and experiences with, the Norwegian model in Finland. *Journal of Data and Information Science*, *3*(4), 31–44. https://doi.org/10.2478/jdis-2018-0019

Pölönen, J., Engels, T., & Guns, R. (2019). Ambiguity in identification of peer-reviewed publications in the Finnish and Flemish performance-based research funding systems. *Science and Public Policy*, scz041. **DOI:** https://doi.org/10.1093/scipol/scz041

Puuska, H.-M., Guns, R., Pölönen, J., Sivertsen, G., Mañana-Rodríguez, J., & Engels, T. (2018). Proof of concept of a European database for social sciences and humanities publications: Description of the VIRTA-ENRESSH pilot. *CSC & ENRESSH*. **DOI:** https://doi.org/10.6084/m9.figshare.5993506

Puuska, H.-M., Nikkanen, J., Engels, T., Guns, R., Ivanović, D., & Pölönen, J. (2020). Integration of national publication databases—towards a high-quality and comprehensive information base on scholarly publications in Europe. *Proceedings of the International Conference on ICT enhanced Social Sciences and Humanities 2020* (forthcoming).

Research Information. (2017). *Monitoring the Transition to Open Access*. December 2017. https://web.archive.org/web/20190103150623/https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2017/monitoring-transition-open-access-2017.pdf

Rimmert, C., Bruns, A., Lenke, C., & Taubert, N. C. (2017). *ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA) 2.0*. Bielefeld University. https://pub.uni-bielefeld.de/record/2913654

Science Europe. (2016). *Position Statement on Research Information Systems*. Co-ordination by Science Europe Working Group on Research Policy and Programme Evaluation. D/2016/13.324/11. Brussels: Science Europe. http://www.scienceeurope.org/wp-content/uploads/2016/11/SE_PositionStatement_RIS_WEB.pdf

Science-Metrix. (2018). *Open Access Availability of Scientific Publications*. https://web.archive.org/web/20180212165611/http://www.science-metrix.com/sites/default/files/science-metrix/publications/science-metrix_open_access_availability_scientific_publications_report.pdf

Seuri, A., & Vartiainen, H. (2018) *Yliopistojen rahoitus, kannustimet ja rakennekehitys*. Accessed June 16, 2019. https://web.archive.org/web/20190616064420/https://www.talouspolitiikanarviointineuvosto.fi/wordpress/wp-content/uploads/2018/01/Seuri_Vartiainen_2018-1.pdf

Sherpa.ac.uk. (2019). *Publisher Copyright Policies & Self-Archiving*. http://www.sherpa.ac.uk/romeo/index.php

Sīle, L., Guns, R., Ivanović, D., Pölönen, J., & Engels, T. (2019). *Creating and Maintaining a National Bibliographic Database for Research Output: Manual of Good Practices*. Antwerp: ECOOM & ENRESSH. **DOI:** https://doi.org/10.6084/m9.figshare.9989204

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). *European Databases and Repositories for Social Sciences and Humanities Research Output*. Antwerp: ECOOM & ENRESSH. **DOI:** https://doi.org/10.6084/m9.figshare.5172322.v2.

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T., … Teitelbaump, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: findings from a European survey. *Research Evaluation*, *27*(4), 310–322. **DOI:** https://doi.org/10.1093/reseval/rvy016

Sivertsen, G. (2016a). Publication-based funding: The Norwegian model. In: M. Ochsner et al. (Eds.), *Research Assessment in the Humanities: Towards Criteria and Procedures* (pp. 71–90). Springer International Publishing. **DOI:** https://doi.org/10.1007/978-3-319-29016-4_7

Sivertsen, G. (2016b). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, *107*(2), 357–368. **DOI:** https://doi.org/10.1007/s11192-016-1845-1

Sivertsen, G. (2016c). Data integration in Scandinavia. *Scientometrics*, *106*(2), 849–855. **DOI:** https://doi.org/10.1007/s11192-015-1817-x

Sivertsen, G. (2017). Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. *Palgrave Communications*, *3*, 17078. **DOI:** https://doi.org/10.1057/palcomms.2017.78

Sivertsen, G. (2018a). The Norwegian model in Norway. *Journal of Data and Information Science*, *3*(4), 3–19. **DOI:** https://doi.org/10.2478/jdis-2018-0017

Sivertsen, G. (2018b). Balanced multilingualism in science. *BiD: Textos Universitaris de Biblioteconomia i Documentació*. No. 40. **DOI:** https://doi.org/10.1344/BiD2018.40.25

Sivertsen, G. (2019). Developing Current Research Information Systems (CRIS) as data sources for studies of research. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 667–683). Cham: Springer.

Sivertsen, G., Guns, R., Kulczycki, E., & Pölönen, J. (2019). The use of Gold Open Access in four European countries: An analysis at the level of articles. In G. Catalano et al. (Eds.), *Proceedings of the 17th International Conference of the International Society for Scientometrics and Informetrics*, *Vol. II* (pp. 1600–1605). Rome: International Society for Scientometrics and Informetrics. **DOI:** https://doi.org/10.1007/978-3-030-02511-3_25

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*(2), 567–575. **DOI:** https://doi.org/10.1007/s11192-011-0615-3

Somoza-Fernández, M., Rodríguez-Gairín, J.-M., & Urbano, C. (2018). Journal coverage of the Emerging Sources Citation Index. *Learned Publishing*, *31*(3), 199–204. **DOI:** https://doi.org/10.1002/leap.1160

Tennant, J. (2018). Complaint to the European Ombudsman about Elsevier and the Open Science Monitor. *Zenodo*. **DOI:** https://doi.org/10.5281/zenodo.1305847

van Leeuwen, T. N., Tatum, C., & Wouters, P. F. (2018). Exploring possibilities to use bibliometric data to monitor gold open access publishing at the national level. *Journal of the Association for*

Information Science and Technology, *69*(9), 1161–1173. **DOI:** https://doi.org/10.1002/asi.24029

van Leeuwen, T. N., van Wijk, E., & Wouters, P. F. (2016). Bibliometric analysis of output and impact based on CRIS data: A case study on the registered output of a Dutch university. *Scientometrics*, *106*, 1–16. **DOI:** https://doi.org/10.1007/s11192-015-1788-y

VIRTA Wiki. (2018). Kerättävät tiedot. Accessed June 16, 2019. https://web.archive.org/web/20190616070432/https://wiki.eduuni.fi/pages/viewpage.action?pageId=48922139

Visser, M., van Eck, N. J., & Waltman, L. (2020). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. https://arxiv.org/ftp/arxiv/papers/2005/2005.10732.pdf

Waltman, L. (2019). *Open Metadata of Scholarly Publications: Open Science Monitor Case Study*. Brussels: European Commission. **DOI:** https://doi.org/10.2777/132318

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., … Johnson, B. (2015). *The Metric Tide. Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE. **DOI:** https://doi.org/10.13140/RG.2.1.4929.1363

Wohlgemuth, M., Rimmert, C., & Winterhager, M. (2016). *ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA)*. Bielefeld University. **DOI:** https://doi.org/10.4119/unibi/2906347