OPEN SCIENCE MONITOR

DRAFT METHODOLOGICAL NOTE

Brussels, April 30th 2018

Consortium partners:

the Lisbon council
think tank for the 21st century

CWTS
Meaningful metrics

ESADE
RAMON LLULL UNIVERSITY

Subcontractor:

NON SOLUS
ELSEVIER

1 Introduction

Open science has recently emerged as a powerful trend in research policy. To be clear, openness has always been a core value of science, but it meant publishing the results or research in a journal article. Today, there is consensus that, by ensuring the widest possible access and reuse to publications, data, code and other intermediate outputs, scientific productivity grows, scientific misconduct becomes rarer, discoveries are accelerated. Yet it is also clear that progress towards open science is slow, because it has to fit in a system that provides appropriate incentives to all parties. Of course dr. Rossi can advance his research faster by having access to dr. Svensson's data, but what is the rationale for dr Svensson to share her data if no one includes data citation metrics in the career assessment criteria?

The European Commission has recognized this challenge and moved forward with strong initiatives from the initial 2012 recommendation on scientific information (C (2012) 4890), such as the Open Science Policy Platform and the European Open Science Cloud. Open access and open data are now the default option for grantees of H2020.

The Open Science Monitor (OSM) aims to provide data and insight needed to support the implementation of these policies. It gathers the best available evidence on the evolution of Open Science, its drivers and impacts, drawing on multiple indicators as well as on a rich set of case studies. [1]

This monitoring exercise is challenging. Open science is a fast evolving, multidimensional phenomenon. According to the OECD (2015), "open science encompasses unhindered access to scientific articles, access to data from public research, and collaborative research enabled by ICT tools and incentives". This very definition confirms the relative fuzziness of the concept and the need for a clear definition of the "trends" that compose open science.

Precisely because of the fast evolution and novelty of these trends, in many cases it is not possible to find consolidated, widely recognized indicators. For more established trends, such as open access to publications, robust indicators are available through bibliometric analysis. For most others, such as open code and open hardware, there are no standardized metrics or data gathering techniques and there is the need to identify the best available indicator that allows one to capture the evolution and show the importance of the trend.

The present document illustrates the methodology behind the selected indicators for each trend. The purpose of the document is to ensure transparency and to gather feedback in order to improve the selected indicators, the data sources and overall analysis.

The initial launch of the OSM contains a limited number of indicators, mainly updating the existing indicators from the previous Monitor (2017). New trends and new indicators will be added in the course of the OSM project, also based on the feedback to the present document.

[1] The OSM has been published in 2017 as a pilot and re-launched by the European Commission in 2018 through a contract with a consortium composed by the Lisbon Council, ESADE Business School and CWTS of Leiden University (plus Elsevier as subcontractor). See https://ec.europa.eu/research/openscience/index.cfm?pg=home&section=monitor

1.1 Objectives

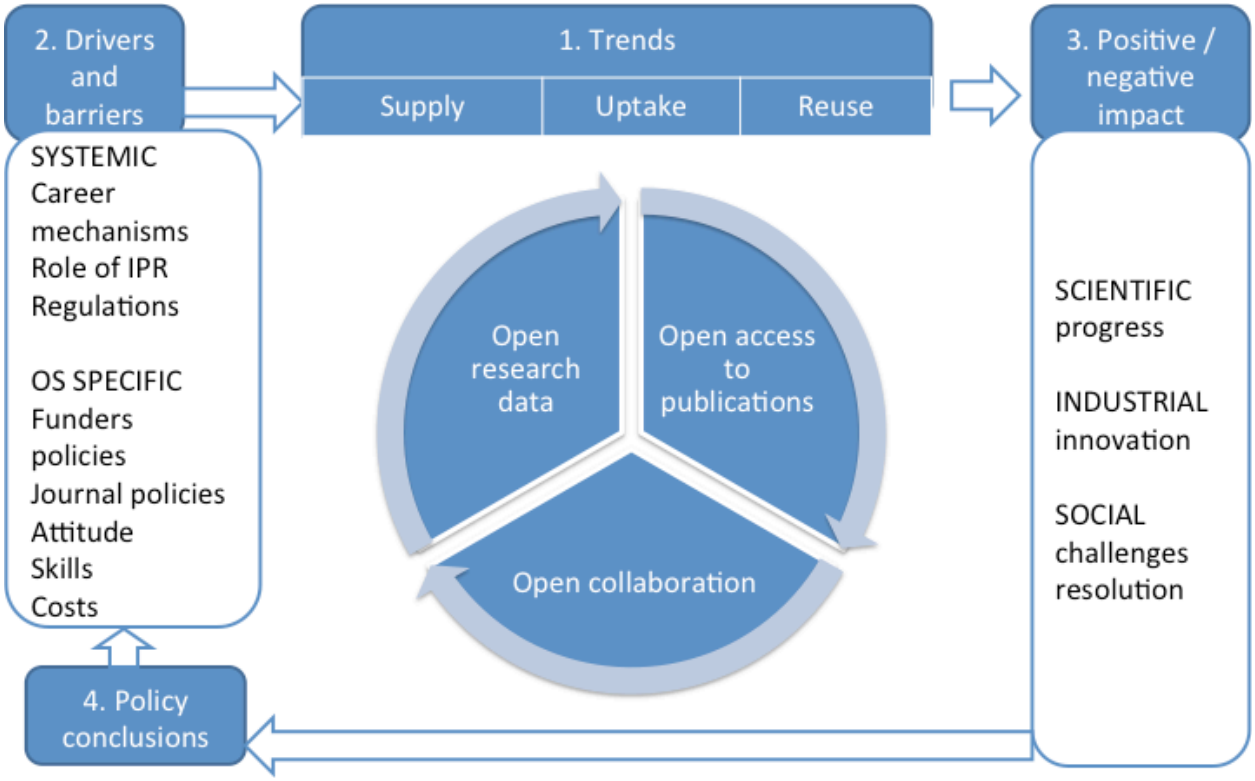
The OSM covers four tasks:

1. To provide metrics on the open science trends and their development.
2. To assess the drivers (and barriers) to open science adoption.
3. To identify the impacts (both positive and negative) of open science
4. To support evidence based policy actions.

The indicators presented here focus mainly on the first two tasks: mapping the trends, and understanding the drivers (and barriers) for open science implementation.

The chart below provides an overview of the underlying conceptual model.

Figure 1: A conceptual model: an intervention logic approach



The central aspect of the model refers to the analysis of the open science trends and is articulated alongside three dimensions: *supply*, *uptake* and *reuse* of scientific outputs.

In the OSM framework, *supply* refers to the emergence of services such as data repositories. The number of data repositories (one of the existing indicators) is a *supply* indicator of the development of Open Science. On the demand side, indicators include, for example, the amount of data stored in the repositories, the percentage of scientists sharing data. Finally, because of the nature of Open Science, the analysis will go beyond usage, since the reuse dimension is particularly important. In this case, relevant indicators include the number of scientist reusing data published by other scientists, or the number of papers using these data.

On the left side of the chart, the model identifies the key factors influencing the trends, both positively and negatively (i.e. *drivers* and *barriers*). Both drivers and barriers are particularly relevant for policy-makers as this is the area where an action can make greatest difference, and are therefore strongly related to policy recommendations. These include "policy drivers", such as funders' mandates. It is important to assess not only policy drivers dedicated to open science, but also more general policy drivers that could have an impact on the uptake of open science. For instance, the increasing reliance on performance based funding or the emphasis on market exploitation of research are general policy drivers that could actually slow down the uptake of open science.

The right side of the chart in the model, illustrates the *impacts* of open science to research or the scientific process itself; to industry or the capacity to translate research into marketable products and services; to society or the capacity to address societal challenges.

1.2 Scope

By definition, open science concerns the **entire cycle** of the scientific process, not only open access to publications (Burgelman et al., 2010). Hence the macro-trends covered by the study include: open access to publications, open research data and open collaboration.

Table 1: Articulation of the trends to be monitored

Categories	**Trends**
Open access to publications	• Open access policies (funders and journals), • Green and gold open access adoption (bibliometrics).[2]
Open research data	• Open data policies (funders and journals) • Open data repositories • Open data adoption and researchers' attitudes.
Open collaboration	• Open code, • Altmetrics, • Open hardware, • Citizen science.

New trends within the open science framework will be identified through interaction with the stakeholder's community by monitoring discussion groups, associations (such as Research Data Alliance- RDA), mailing lists, and conferences such as those organised by Force11 (www.force11.org).

[2] According to the EC, "'Gold open access' means that open access is provided immediately via the publisher when an article is published, i.e. where it is published in open access journals or in 'hybrid' journals combining subscription access and open access to individual articles. In gold open access, the payment of publication costs ('article processing charges') is shifted from readers' subscriptions to (generally one-off) payments by the author.[...] 'Green. open access' means that the published article or the final peer-reviewed manuscript is archived by the researcher (or a representative) in an online repository." (Source: H2020 Model Grant Agreement)

The study covers **all research disciplines**, and aims to identify the differences in open science adoption and dynamics between diverse disciplines. Current evidence shows diversity in open science practices in different research fields, particularly in data-intensive research domains (e.g life sciences) compared to others (e.g humanities)

The **geographic coverage** of the study is 28 Member States (MS) and G8 countries, including the main international partners, with different degrees of granularity for the different variables. As far as possible, data has to be presented at **country level**.

Finally, the analysis focuses on the factors at play for **different stakeholders** as mapped in the chart below (table 2). For each stakeholder's category, OSM will deliberately consider both traditional (e.g Thomson Reuters) and new players in research (e.g F1000).

Table 2: Stakeholders types

Researchers	Professional and citizens researchers
Research institutions	Universities, other publicly funded research institutions, and informal groups
Publishers	Traditional publishers New OA online players
Service providers	Bibliometrics and new players
Policy makers	At supranational, national and local level
Research funders	Private and public funding agencies.

2 Indicators and data sources

Because of the fast and multidimensional nature of open science, a wide variety of indicators have been used, depending on data availability:

- Bibliometrics: this is the case for open access to publications indicators, and partially for open data and altmetrics.
- Online repositories: there are many repositories dedicated to providing a wide coverage of the trends, such as policies by funders and journals, APIs and open hardware.
- Surveys: surveys of researchers shed light on usage and drivers. Preference is given to multi-year surveys.
- Ad hoc analysis in scientific articles or reports: for instance, reviews of journals policies with regard to open data and open code
- Data from specific services: open science services often offer data on their uptake, as for Sci-starter or Mendeley. In this case, data offer limited representativeness about the trend in general, but can still be useful to detect differences (e.g. by country or discipline). Where possible, in this case, we present data from multiple services.

At the time of the publication of the new OSM (May 2018), only a sub-group of the indicators listed below are already published (indicated with a "*" sign in the tables). The others will be published at regular intervals.

2.1 Open access to publications

Beside the long list of indicators below, the detailed methodology for calculating the percentage of OA publications is presented in the annex at the end of this document.

Indicator	**Source**
Number of Funders with open access policies *	Sherpa Juliet[3]
Number of Journals with open access policies *	Sherpa Romeo[4]
P - # Scopus publications that enter in the analysis*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
P(oa) - # Scopus publications that are Open Access (CWTS method for OA identification)*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
P(green oa) - # Scopus publications that are Green OA*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
P(gold oa) - # Scopus publications that are Gold OA*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
PP(oa) - Percentage OA publications of total publications*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
PP(green oa) - Percentage gold OA publications of total publications*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
PP(gold oa) - Percentage green OA publications of total publications*	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
TCS - Total Citation Score. Sum of all citations received by P in Scopus.	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
FWCI – Field Weighted Citation Score.	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire
TP1/TP10 - Top1/Top10 percentile highly cited publications	Scopus, DOAJ, ROAD, PubMedCentral, CrossRef, OpenAire

[3] http://v2.sherpa.ac.uk/juliet/
[4] http://www.sherpa.ac.uk/romeo/index.php?la=en&fIDnum=|&mode=simple

2.2 Open research data

Indicator	Source
Number of Funders with policies on data sharing*	Sherpa Juliet
Number of Journals with policies on data sharing*	Vasilevsky et al, 2017[5]
Number of open data repositories*	Re3data
% of paper published with data	Bibliometrics: Datacite
Citations of data journals	Bibliometrics: Datacite
Attitude of researchers on data sharing*	Survey by Elsevier, follow-up of the 2017 report.[6]

2.3 Open collaboration

Indicator	Source
Membership of social networks on science (Mendeley, ResearchGate, f1000)	Scientific social networks

2.3.1 Open Code

Indicator	Source
Number of code projects with DOI	Mozilla Codemeta
Number of scientific API*	Programmableweb
% of journals with open code policy*	Stodden 2013[7]
Number of scientific projects on Github	Github

2.3.2 Open scientific hardware

Indicator	Source
Number of projects on open hardware repository*	Open Hardware repository[8]

[5] Vasilevsky, Nicole A., Jessica Minnier, Melissa A. Haendel, and Robin E. Champieux. “Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark?” PeerJ 5 (April 25, 2017): e3208. doi:10.7717/peerj.3208.

[6] Berghmans, Stephane, Helena Cousijn, Gemma Deakin, Ingeborg Meijer, Adrian Mulligan, Andrew Plume, Sarah de Rijcke, et al. “Open Data : The Researcher Perspective,” 2017, 48 p. doi:10.17632/bwrnfb4bvh.1.

[7] Stodden, V., Guo, P. and Ma, Z. (2013), “Toward reproducible computational research: an empirical analysis of data and code policy adoption”, PLoS One, Vol. 8 No. 6, p. e67111. doi: 10.1371/ journal.pone.0067111.

Number of projects using open hardware license*	Open Hardware repository

2.3.3 Citizen science

Indicator	**Source**
N. Projects in Zooniverse and Scistarter*	Zooniverse and Scistarter
N. Participants in Zooniverse and Scistarter	Zooniverse and Scistarter

2.3.4 Altmetrics

Indicator	**Source**
P(tracked) - # Scopus publications that can be tracked by the different sources (e.g. typically only publications with a DOI, PMID, Scopus id, etc. can be tracked).	Scopus & Plum Analytics
P(mendeley) - # Scopus publications with readership activity in Mendeley	Scopus, Mendeley & Plum Analytics
PP(mendeley) - Proportion of publications covered on Mendeley. P(mendeley)/P(tracked)	Scopus, Mendeley & Plum Analytics
TRS - Total Readership Score of Scopus publications. Sum of all Mendeley readership received by all P(tracked)	Scopus, Mendeley & Plum Analytics
TRS(academics) - Total Readership Score of Scopus publications from Mendeley academic users (PhdS, Professors, Postdocs, researchers, etc.)	Scopus, Mendeley & Plum Analytics
TRS(students) - Total Readership Score of Scopus publications from Mendeley student users (Master and Bachelor students)	Scopus, Mendeley & Plum Analytics
TRS(professionals) - Total Readership Score of Scopus publications from Mendeley professional users (librarians, other professionals, etc.)	Scopus, Mendeley & Plum Analytics
MRS - Mean Readerships Score. TRS/P(tracked)	Scopus & Plum Analytics
MRS(academics) - TRS(academics)/P(tracked)	Scopus & Plum Analytics
MRS(students) - TRS(students)/P(tracked)	Scopus & Plum Analytics
MRS(professionals) - TRS(professionals)/P(tracked)	Scopus & Plum Analytics
P(twitter) - # Scopus publications that have been mentioned in at least one (re)tweet	Scopus & Plum Analytics
PP(twitter) - Proportion of publications mentioned on Twitter. P(twitter)/P(tracked)	Scopus & Plum Analytics

[8] https://www.ohwr.org

TTWS - Total Twitter Score. Sum of all tweets mentions received by all P(tracked)	Scopus & Plum Analytics
MTWS - Mean Twitter Score. TTWS/P(tracked)	Scopus & Plum Analytics

3 Next steps

This methodological note is published for public feedback, to be gathered until July 30th 2018. The feedback is particularly aiming at concrete, actionable suggestions for improvement:

- By identifying new trends, not selected so far
- By proposing improved indicators for the selected trends
- By selecting new, improved data sources for the selected indicators.

In October 2018 the final methodology will be released, together with new indicators and data.

Annex: Technical report on the identification of Open Access publishing

Thed van Leeuwen & Rodrigo Costas

Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

Introduction

In this document the approach for the identification and creation of the Open Access (OA) labels for the *Open Science Monitor* (hereafter referred to as OS Monitor) is presented. As stated in the Terms of reference, CWTS is following the exact same method that has been developed over the last two years, and which has been reported at the Paris 2017 STI Conference (van Leeuwen et al, 2017). In this method we strive for a high degree of reproducibility of our results based upon data carrying OA labels following from the methodology we developed. Our initial developments were based on the Web of Science, but for the OS Monitor the exact same method will be based on Elsevier's Scopus data.

The methodological approach that we propose mainly focuses on adding different OA labels to the Scopus database, using various data sources to establish this OA status of scientific publications. It is important to highlight that two basic principles for this OA label are ***sustainability*** and ***legality***. By sustainability we mean that it should, in principle, be possible to reproduce the OA labeling from the various sources used, repeatedly, in an open fashion, with a relatively limited risk of the sources used disappearing behind a pay-wall, and particularly that the reported publications as OA will change their status to closed. The second aspect (legality) relates to the usage of data sources that represent legal OA evidence for publications, excluding rogue or illegal OA publications (i.e. we do not consider OA publications made freely available in platforms such as ResearchGate or Sci-hub). While the former criterion is mainly oriented to a scientific requirement, namely that of reproducibility and perdurability over time, the latter criteria is particularly important for science policy, indicating that OA publishing aligns with policies and mandates.

Data sources used for establishing OA labels

As main data sources to identify evidence of Open Access for publications covered in the Scopus database for the years 2009 to 2016, we used:

- the DOAJ list (Directory of Open Access Journals) [https://doaj.org/],
- the ROAD list (Directory of Open Access scholarly Resources) [http://road.issn.org/],
- PMC (PubMed Central) [https://www.ncbi.nlm.nih.gov/pmc/],
- CrossRef [https://www.crossref.org/], and
- OpenAIRE [https://www.openaire.eu/]

These five sources serve to label the publications according to the terminology used in the OA development. The first two sources (DOAJ and ROAD) serve to identify and label ***Gold OA***, while the last three sources (PMC, CrossRef and OpenAIRE) serve to identify and label ***Green OA***. In cases where publications published in Gold OA journals were also identified in one of the other sources, we determine the status of the publication as Gold OA. So Gold OA goes over Green OA, as Gold is a more deliberate choice of the authors, often driven by a mandate of publishing in a journal that is fully OA.

All these five sources fulfill the above-mentioned requirements while other popular 'apparent' OA sources such as ResearchGate and SciHub fail to meet these two principle requirements. Thus, it is important to stress here that our approach has a more policy perspective than a utilitarian one (i.e. just identifying publications that are freely available). In other words, our approach aims to inform the number and share of sustainable and legal OA publications (i.e. publications that have been published in OA journals or archived in official and legal repositories), instead of the mere identification of publications whose full text can be retrieved online (regardless the source or the legal status of the access to the publication). For a broader discussion on other types of OA as well as other possibilities of identifying OA we refer the reader to our recent paper Martín-Martín et al. (2018) [https://arxiv.org/abs/1803.06161]

Sources of Open Access evidence

The sources that were mentioned above were fully downloaded (as provided by the original sources) using their public Application Programming Interfaces (API). The metadata obtained has been parsed and incorporated into an SQL environment in the form of relational databases.

DOAJ

A first source we used was the DOAJ list of OA journals. This list was linked to the Scopus database on the basis of the regular ISSN code, as well as the eISSN code available in both the DOAJ list as well as in the Scopus database. This resulted in a recall of 1,028,447 publications labeled in Scopus as being OA, via the regular ISSN code, while the eISSN code resulted in 95,162 additional publications.

ROAD

A next source used to add labels to the Scopus database is the ROAD list. ROAD has been developed with the support of the UNESCO, and is related to ISSN International Centre. The list provides access to a subset of the ISSN Register. This subset comprises bibliographic records which describe scholarly resources in OA identified by an ISSN: journals, monographic series, conference proceedings and academic repositories. The linking of the ROAD list is based upon the ISSN code, as well as the eISSN code available in both the Scopus as well as in the ROAD list. This resulted in a total of 524,082 publications being labeled as OA, while the eISSN code resulted in 938,787 additional publications.

CrossRef

A third source that was used to establish an Open Access label to Scopus publications was CrossRef, based upon the DOI's available in both systems. This led to the establishment of a total of 37,119 publications as being licensed as OA according to CrossRef.

PubMed Central

A fourth source used is the PubMed Central database. This is done in two ways; the first based upon the DOI's available in both the PMC database as well as in the Scopus database. This resulted in total in 1,974,941 publications being labeled as OA in the Scopus environment. The second approach was based upon the PMID code (where PMID stands for PubMedID) in the PMC database as well as in the Scopus database. This resulted in a total of 1,102,937 publications being labeled as OA in the Scopus database.

OpenAIRE

A fifth and final data source used to add OA labels to the Scopus database is the OpenAIRE database. OpenAIRE is a European database that aggregates metadata on OA publications from multiple institutional repositories (mostly in Europe), including also thematic repositories such as **arxiv.org**. The matching is done in two different ways: the first one based upon a matching by using the DOI's or PMIDs available in both OpenAIRE and in Scopus (resulting in 2,326,442 publications); and second, on a fuzzy matching principle of diverse bibliographic metadata both in Scopus and OpenAIRE (including articles' titles, publication years and other bibliographic characteristics) (resulting in total in 2,976,620 publications) (the methodology is similar to the methodology for citation matching employed at CWTS – Olensky et al. 2016.

In comparison with the previous studies in which our methodology of labeling OA was applied to Web of Science (WoS), the implementation of the methodology on the Scopus database offers with respect to the DOAJ and ROAD lists the advantage that Scopus also contains the eISSN codes, contrary to WoS. This results in a relative larger number of publications covered by the methodology related to DOAJ and ROAD, hence the numbers of publications as well as the share of publications in Gold OA are higher as compared to results obtained for the WoS database.

The fuzzy matching algorithms underlying the linking of OpenAIRE to Scopus have been revised, and made more accurate in comparison to the previous version of the algorithm. So this probably leads to higher recall as well. Due to the fact that this is applied first in WoS and now in Scopus, with both databases differing in coverage and also time periods, it is impossible to state what the exact difference is.

A new source of Open Access evidence: Unpaywall data

More recently, a new source for OA evidence appeared on the scene, the former OADOI, nowadays *Unpaywall* database (https://unpaywall.org/). We have not yet integrated that into the current analysis, but plan to integrate the information stored in this system in a next run of the analysis, leading to expanding the filling of the OS Monitor with (potentially) additional OA publishing information. For now we have conducted a few analyses, comparing our methodology and the numbers of publication labeled with OA tags, with the Unpaywall data (see also Martín-Martín et al, 2018). A few immediate differences worth mentioning are:

- Our methodology includes the ROAD list, a source not covered by Unpaywall;
- Our methodology includes the OpenAIRE dataset, a source not covered by Unpaywall; this implies that our methodology has a somewhat better coverage in Europe (which is the scope of OpenAIRE), while Unpaywall seems to slightly better represent OA publishing in the US and other non-European countries;
- UnPayWall discloses hybrid OA publishing, of which a sub-set consists of Bronze OA tags to publications.

In the immediate future we will start working on the possibilities to include Unpaywall data into the methodology that tags publications with OA labels. This requires conducting research to better understand what data Unpaywall actually disclose, whether all types of OA evidence actually fit into our criteria of building OA evidence, and whether there are other potential conceptual issues related to some typologies of OA provided by Unpaywall (e.g. it is not totally clear whether the Bronze OA typology disclosed by UnPayWall can really be considered a sustainable form of OA, cf. Martín-Martín et al, 2018).

References:

van Leeuwen TN, Meijer I, Yegros-Yegros, A & Costas R, Developing indicators on Open Access by combining evidence from diverse data sources , Proceedings of the 2017 STI Conference, 6-8 September, Paris, France (https://sti2017.paris/) (https://arxiv.org/abs/1802.02827)

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. SocArXiv papers. DOI: 10.17605/OSF.IO/K54UV

Olensky, M., Schmidt, M., & Van Eck, N.J. (2016). Evaluation of the Citation Matching Algorithms of CWTS and iFQ in Comparison to the Web of Science. *Journal of the Association for Information Science and Technology*, 67(10), 2550-2564. doi:10.1002/asi.23590.