Feedback on EC Open Science Monitor Methodological note

French Open science committee

The draft methodological note[1] published by the EC about the Open Science Monitor raises several questions. As it encourages for feedback, we detail below our remarks and recommendations for future work.

First, the main point of this feedback is about the choice made by the consortium (CWTS, ESADE and Lisbon Council) to choose Elsevier as the sole subcontractor for the Open Science Monitor.

1. Consequences implied by the subcontractor choice

- **Coverage bias**. The vast majority of the indicators proposed to track the Open Science progress are based on Elsevier proprietary data (namely Scopus). As stated in the methodological note itself[2] (p. 14), Scopus coverage is partial as "both databases [WoS and Scopus] [are] differing in coverage and also time periods". That means that choosing Scopus as the only referential for all publications creates a huge bias that is not even mentioned in the note.

- **Major conflict of interest.** Also, having Elsevier as the sole subcontractor for the OSM puts Elsevier in a position where they will be, at the same time, publishing and monitoring science communications on one side, and also evaluating it on the other side, which creates a huge conflict of interest. This point is even more crucial as the next step of the OSM is to have indicators not only on quantity (OA percentage) but also on "quality" of the OA publications. That means that qualifying what is "good science" will only be in Elsevier's hands.

- **Elsevier's competitors discriminated.** This bias will also discriminate against Elsevier's competitor whose product (scientific publishing) would be evaluated by Elsevier. The OSM construction would a great opportunity to build a sustainable and open alternative to Elsevier's Scopus and Thomson Reuters' Web of Science. This alternative open database could become a base reference in bibliometrics.

- **Non-reproducibility.** On top of that, monitoring open science based on not-open data makes impossible to reproduce the analysis (trends, split by discipline, country …) that will be needed at the local level to get the right insights on open science

[1] https://ec.europa.eu/info/sites/info/files/open_science_monitor_methodological_note_v2.pdf
[2] Ibid.

progress and make the right decisions.

As a consequence of the issues listed above, we first need to start from the objectives stated by the EC and from that express the conditions needed to build an effective Open Science Monitor.
The Open Science Monitor aims at providing data and insights to support the Open Science policies. It has to monitor the evolution of open science, its drivers and its impacts.
As stated in the methodological note, this monitoring exercise is a challenge, in particular because open science is a fast evolving and multi-dimensional phenomenon.
As a consequence, several prerequisites appear for the OSM.

2. Identified prerequisites for the OSM

- **Full reproducibility** which implies OSM to be based on **open data**. This holds for determining OA evidence (DOAJ, Unpaywall, OpenAIRE, DOAB …) but also for the whole database used to gather all the publications. Using Scopus or WoS would prevent the community to reproduce the analysis and being able to detect potential bias or bugs.

- **Granularity.** The OSM analysis need to be splitted across multiple dimensions to be actionable. This requires to have the most granular data (each publication entry and its meta-data) enriched as much as possible by combining multiple sources to provide informations on the authors, the publisher, the journal and the article itself (including OA evidence, hosting, license).

3. Recommandations

Based on previous comments, we can formulate some recommendations for future work on OSM.

- **Using open databases as base references** (Crossref, Unpaywall), that contains one entry per DOI. That implies in particular to focus on publications that do have a DOI.
  The coverage and data quality (especially for affiliations) of these open data bases have to be studied carefully as we propose them to be the backbone of the methodology.
  A blog post published by CWTS[3] studied the matching between Crossref, WoS and Scopus using the DOI as the join key for the period 2012-2016. According to this article, it appears that “Crossref has 19.1 millions publications for the period, which is substantially more than the 11.9 and 13.9 millions” publications respectively for WoS and Scopus. It also states that “a large share of the scholarly literature indexed in

[3] https://www.cwts.nl/blog?article=n-r2s234

WoS and Scopus is also available in Crossref. For recent years, 68% of the WoS publications and 77% of the Scopus publications can be matched with Crossref using DOIs as a crosswalking mechanism. These figures are likely to underestimate the true overlap between the data sources, since matching based on DOIs presents several difficulties, such as missing, incorrect, and duplicate DOIs."
These are pretty encouraging to choose an open database like Crossref as a backbone to build a publications dataset. It has to be mentioned that this blog post also encourages publishers to improve the data quality of the references in Crossref which would be very helpful when studying citations-based metrics.

- **Improving sources to determine OA evidence**. Unpaywall should definitely be added to the sources already proposed by the consortium. The performance of the OA evidence finding has to be studied as well, with two key metrics: precision and recall. Recall is the fraction of the actual open publications that are successfully retrieved as OA by the system. Conversely, precision is the fraction of publications classified as OA by the system that are actually OA. It is to be remarked that the precision is more important than recall in this context as a low recall will result in a conservative estimation of the Open Science penetration (in the case of not finding OA evidence for publication that are actually OA).
A recent article[4] shows that precision for the oaDOI system (now called Unpaywall) is 96.6% and the recall 77%. These figures were obtained with a manual check on a 500 random sample of DOIs from Crossref.

- **Refining OA types** :
Following the same paper[5], we propose to refine the types of OA. The "gold" and "green" OA definition used in the consortium note may hide some complexity of the reality. In particular for publisher-hosted publications (publication available on the publisher website), it is key to differentiate cases.

    - ➔ **Gold**: Published in an open-access journal (publisher-hosted), with an explicite open license (with a machine-readable format)
    - ➔ **Hybrid**: Free under a proper open license in a toll-access journal (publisher-hosted)
    - ➔ **Bronze**: Free to read on the publisher page, but without a clearly identifiable license (publisher-hosted)
    - ➔ **Green**: There is a free copy in an OA repository (repository-hosted)

Non legal alternatives like Sci-hub should not be considered as OA.

Neither should be counted as OA publications freely available only on an academic social network (like Academia.edu or ResearchGate).

- **Deeper analysis.** More data enrichment will be needed to conduct deeper analysis, especially at the national level. Identifying publications authors affiliations and

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815332/
[5] Ibid.

discipline for instance would be mandatory to give insights on Open Science evolution.
As far as it is concerned, France could work on enriching data for its publications to be able to monitor closely the Open Access evolution in France.

4. An opportunity for France

On the 4th of July 2018, France published its National Plan for Open Science[6], which - among other major commitments for open access to publications and to research data, commits to set-up an Open Science Monitor in France.

Thereby, the work initiated by the EC is a good step and national Open Science Monitor should not reinvent the wheel.

However, to be actionable at the national level, the Open Science Monitor has to follow the identified prerequisites that are **full reproducibility** and **granularity**, which are not compatible with the current setting based on Elsevier’s Scopus proprietary data.

A cooperation between the EC OSM consortium and the French Ministry of Higher Education, Research and Innovation would allow to strengthen one another for the OSM construction and set-up. Thus we would like to arrange a meeting to discuss further these topics and in particular the open data standards for the OSM.

[6] http://cache.media.enseignementsup-recherche.gouv.fr/file/Recherche/50/1/SO_A4_2018_EN_01_leger_982501.pdf