Open Science Monitoring Initiative

OSMI WG4 – May 2025

Eric Jeangirard's talk — key take-aways

- **Why the Monitor exists** – Launched **with the 2018 National Plan for Open Science**, the French Open Science Monitor (F-OSM) is meant to be a **sovereign, ever-evolving, totally open-source / open-data instrument** for measuring the impact of French open-science policy. It has to stay transparent, reproducible and *free of any proprietary data feeds*.

- **The 'metadata gap' challenge** – In 2018 proprietary indexes (WoS, Scopus) still contained richer metadata information than open sources. The team therefore decided to *create* missing open metadata themselves using ML, cloud and "a bit of common sense".

- **Pipeline for publications**

    1. **Collect** bib-records only from open aggregators (Crossref, PubMed, HAL, Unpaywall, etc.) plus HTML/PDF scraping.

    2. **Infer French affiliation** (country-level is "good enough" for policy) by combining Unpaywall, GROBID-derived strings and rule-based cleaning. An external benchmark later showed this all-open workflow *outperformed* Scopus & WoS for French coverage .

    3. **Enrich & classify –** using Unpaywall metadata for publications with a DOI and HAL for non -DOI publications, a custom discipline tagger, publisher/ platform normalisation and custom KPIs to steer public policy with a focus on diamond OA and national APC spend estimates.

- **Beyond publications** – Monitor **clinical-trial transparency** (EU & US registries) and **research data / software** using a **dual approach**: harvest DOIs/SWHIDs from repositories *and* text-mine PDFs with the open-source **Softcite/DataSet** modules (trained on ≈5 000 manually-annotated docs) to label each mention as *used / created / shared*.

- **Current picture (2024 data cut)** – OA plateauing at ~66 %; data-sharing KPI climbing (~25 %), software-sharing stagnant; <20 % of French academic clinical trials report full results .

- **Lessons learned** – A small team can build a national monitor without paywalled data; iterate yearly with human-in-the-loop fixes (e.g. **Works Magnet** crowdsourcing 50 k+ affiliation corrections); coordinate internationally to share training corpora and "sharable" full-text outputs or open source software.

Live Q & A — highlights

Topic / asker	What they asked	Eric’s reply
Skills & LLM capacity	How can India replicate F-OSM without in-house ML/LLM experts?	You don’t need “10 experts per country”; pool global effort, reuse open code; LLM work still exploratory
Coverage biases	Training data is biomed/CS-heavy; plans for SSH? https://osf.io/kgj52 results of analysis with SSH publications at VU	Agrees; exploring new partners or retraining to broaden domains; welcomes collaboration
Publication granularity & licence gaps, challenges related to the quality of open data	Does the F-OSM distinguish professional vs scientific works? How do you handle “free access, no licence”? affiliation normalisation in OpenAlex?	F-OSM limits corpus to peer-reviewed works; treats “free/no-licence” via Unpaywall status; regex + merger tables to tidy publishers & affiliations
Data/software ‘available on request’	Do you count that as sharing? does monitor feed OECD stats?	“Request” ≠ shared; only explicit links count. Results steer French policy; unsure about OECD uptake
OpenAlex quality	Wrong affiliations/duplicates in OpenAlex?	Built **Works Magnet**: community fixes pushed daily to OpenAlex while waiting for larger solutions
Global coordination / hackathon	Could we align F-OSM, Barcelona Declaration, etc.? A hackathon is planned and all interested should contact Melissa at mharrison@ebi.ac.uk	Fully supports joint hackathons and shared “gold-standard” corpora

Next-step figures	How will new KPIs roll out?	Ministry holds quarterly reviews; same infrastructure offered free to French institutions so every campus gets identical KPIs and cost-effective compute