Codebook

Analytical coding of open science monitoring initiatives

OSMI Working Group 2

Version 3

Date: 2026/11/08

Content

Content 2
1. Purpose and scope 3
1.1 Purpose 3
1.2 Scope of coding 3
2. Variables 4
2.1. Variables 5
V1. Monitoring purpose archetype 5
Values 5
Decision rules 6
V2. Compliance linkage strength 6
Values 7
Decision rules 7
V3. Data source typology 8
Values 8
Decision rules 9
V4. Data source identification 9
Controlled vocabulary 10
Decision rules 11
V5. Open science pillars monitored 12
Controlled vocabulary 12
Decision rules 14
Appendix A. Codebook change log 15

How to cite

OSMI Working Group 2: Understanding the open science monitoring landscape, Szybisty, T., & Yahia Mohamed Elkheir, L. (2026). Open Science Monitoring Initiatives Dataset v2 [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.24051502.

1. Purpose and scope

1.1 Purpose

This codebook defines the analytical variables used to code the dataset of open science (OS) monitoring initiatives compiled by OSMI Working Group 2 (WG2). Its purpose is to convert heterogeneous free-text descriptions of each initiative into a small set of structured analytical variables that support consistent, reproducible cross-case comparison.

The codebook is the central methodological artefact for the WG2 analysis phase. It defines what is being coded, by whom, with what decision rules, and against what value sets. Once finalised, it will be published alongside the WG2 report and dataset as supplementary material, consistent with OSMI Principle 2 on transparency and reproducibility.

1.2 Scope of coding

Coding covers all initiatives included in the WG2 consolidated dataset that meet the WG2 definition of an open science monitoring initiative. The unit of analysis is the individual monitoring initiative.

Variables in this codebook are derived from free-text fields in the dataset. Structured fields already present in the dataset (such as operational scope, region, country, lead organisation type, year started, operational status) are not re-coded here.

2. Variables

The codebook defines 5 variables. The table below summarises them; full definitions follow in Section 2.1.

ID	Name	Description	Source of information
V1	**Monitoring purpose archetype**	**Primary purpose of the initiative serves [Single-select + optional secondary]**	**Intended Use; Brief Description; Compliance column**
V1	v1_purpose_primary	Primary monitoring purpose archetype (numeric code 1–7, 99).	
V1	v1_purpose_secondary	Optional secondary purpose archetype (numeric code 1–7, blank if none).	
V1	v1_notes	Coder rationale.	
V2	**Compliance linkage strength**	**Degree to which monitoring results have formal consequences [Single-select]**	**Compliance column; Intended Use; Brief Description**
V2	v2_compliance_linkage	Compliance linkage strength (numeric code 1–5, 99).	
V2	v2_notes	Coder rationale.	
V3	**Data source typology**	**Dominant type of underlying data used [Single-select + open/proprietary flag]**	**Data types/sources; Monitoring Mechanism**
V3	v3_data_source_typology	Data source typology (numeric code 1–7, 99).	
V3	v3_proprietary_flag	Proprietary input as core: yes / no / unclear.	
V3	v3_notes	Coder rationale.	
V4	**Data source identification**	**Specific named data sources used (controlled vocabulary) [multi-select]**	**Data types/sources; Monitoring Mechanism**
V4	v4_data_sources	Multi-select list of canonical source labels from controlled vocabulary, semicolon-separated.	
V4	v4_notes	Distinguishes core vs. enrichment, specifies CRIS or survey instruments.	

ID	Name	Description	Source of information
V5	**Open science pillars monitored**	**Which open science pillar(s) the initiative monitors [multi-select, controlled vocabulary]**	**Open Science Focus / Pillars Monitored**
V5	v5_pillars_monitored	Multi-select list of canonical pillar labels from controlled vocabulary (codes 1-14, 99), semicolon-separated.	
V5	v5_notes	Coder rationale	

2.1. Variables

V1. Monitoring purpose archetype

Variable ID	V1
Variable name	Monitoring purpose archetype
What it captures	The dominant purpose the initiative serves. Initiatives often have multiple purposes; this variable assigns the primary purpose, with an optional secondary purpose recorded separately when clearly present.
Source columns	Primary: "Intended Use of Monitoring Results". Secondary: "Brief Description of the Initiative and its Goal or Intent"; "Use of Monitoring Results for Compliance or Enforcement".
Variable type	Single-select for primary purpose; optional single-select for secondary purpose.
Notes on application	Code the purpose as stated and operationalised by the initiative, not the purpose a reader might infer from the type of organisation alone. A funder-led initiative is not automatically compliance-tracking; some funder-led monitors are explicitly informational only.

Values

1	Policy informing	Produces evidence intended to shape policy at institutional, national, or international levels, without direct enforcement function.
2	Compliance tracking	Exists primarily to verify that researchers, institutions, or funded projects meet specific mandates.

3	Strategic negotiation	Produces evidence used in negotiations with publishers or other external actors, often by library consortia.
4	Internal improvement	Supports institutional or organisational self-improvement, typically without external accountability function.
5	Advocacy and field-building	Produces evidence to demonstrate progress, build the case for open science, or support community-level change.
6	Operational monitoring	Provides ongoing operational visibility into a specific service, programme, or process (e.g., transformative-agreement uptake).
7	Other	Does not fit categories above. Coder notes must specify the alternative purpose.
99	Unclear	Insufficient information to assign a value.

Decision rules

- Read the "Intended Use" column first. If it explicitly names a primary purpose, use that.
- If multiple purposes are described, prioritise based on the stated primary user of the results.
- A funder-led initiative tracking grantee outputs is coded as "compliance tracking" only if there is an explicit link to mandate verification. Otherwise it may be "policy informing" or "operational monitoring".
- A consortium-led initiative producing data feeding publisher negotiations is "strategic negotiation", even if it also informs policy.
- Code a secondary purpose in a separate field if a clearly different second purpose is present in the source text.
- If two purposes are described as equally primary, code the more specific one as primary and the broader one as secondary.
- Do not infer a purpose that is not articulated in the dataset entry. If unstated, code as "unclear".

V2. Compliance linkage strength

Variable ID	V2
Variable name	Compliance linkage strength
What it captures	The degree to which monitoring results have formal consequences for the entities being monitored. Captures where each initiative sits on a spectrum from purely informational to direct enforcement.

Source columns	Primary: "Use of Monitoring Results for Compliance or Enforcement". Secondary: "Intended Use"; "Brief Description".
Variable type	Single-select.
Notes on application	Distinguish between the consequences attached to the monitoring initiative itself and the consequences attached to the underlying policies it monitors. Code the strongest linkage that the initiative directly enables or contributes to.

Values

1	None	Explicitly informational; no link to consequences for monitored entities.
2	Transparency reporting	Results made public for accountability but no formal consequence attached.
3	Soft incentives	Results inform performance agreements, rankings, reputational positioning, or institutional evaluation, but without direct funding consequences.
4	Hard sanctions	Results directly linked to funding withholding, grant reduction, exclusion from programmes, or other concrete penalties.
5	Mixed / partial	Different consequences apply to different actors or outputs within the same initiative (e.g., informational at country level but enforced at funder level).
99	Unclear	Insufficient information to determine.

Decision rules

- Phrases such as "withholding grant payments", "grant reduction", "funding withheld", or "KPI influencing budget" indicate "hard sanctions".
- "Used in evaluations" without specified financial or programmatic consequences is "soft incentives".
- "No formal linkage", "purely informational", or "informational and strategic only" maps to "none".
- "Tracks compliance" without specifying consequences usually maps to "transparency reporting", not "hard sanctions". The monitoring of compliance is not the same as the enforcement of compliance.

- Be conservative: where the column says "potentially", "may inform", or "could be used to", code as "soft incentives" rather than "hard sanctions".
- Where an initiative is a technical infrastructure used by multiple actors with different enforcement regimes (e.g., CHORUS, OpenAIRE Monitor), code based on the consequences the initiative itself directly enables, not those imposed independently by downstream users. Note the distinction in the coder notes column.
- Where evidence points in both directions, code as "mixed / partial" with explanation in notes.

V3. Data source typology

Variable ID	V3
Variable name	Data source typology
What it captures	The dominant type of underlying data the initiative uses. This variable groups data sources into broad categories for cross-case comparison and feeds the infrastructure-dependency analysis.
Source columns	Primary: "Data types/sources used for Monitoring". Secondary: "Monitoring Mechanism / Approach".
Variable type	Single-select for typology; separate binary flag (yes / no / unclear) for whether any proprietary source is used as a core input.
Notes on application	Code based on what is described as a core or primary input to the monitoring, not on enrichment sources mentioned in passing. Use the mechanism column to disambiguate which sources are central.

Values

1	Open scholarly infrastructure-based	Relies primarily on open data sources such as OpenAIRE Graph, OpenAlex, Crossref, DataCite, Unpaywall, DOAJ.
2	Proprietary database-based	Relies primarily on subscription databases such as Web of Science, Scopus, Dimensions.
3	Institutional / national CRIS-based	Relies on institutional or national current research information systems and repositories as primary source.

4	Survey-based	Primary data are surveys of institutions, researchers, or stakeholders.
5	Hybrid (open + proprietary)	Uses substantial input from both open and proprietary sources.
6	Hybrid (automated + manual/survey)	Combines automated harvesting with hand-curation or surveys.
7	Other / specialised	Uses primarily a specialised source not captured above (e.g., clinical trial registries, repository platforms, in-house bibliographic databases).
99	Unclear	Insufficient information to determine.

Decision rules

- Assign value for primary and secondary data sources based on the table above.
- A monitor that uses OpenAlex and mentions Crossref/Unpaywall enrichment is "open scholarly infrastructure-based".
- A monitor explicitly listing both proprietary and open sources as central is "hybrid (open + proprietary)".
- If the description prominently includes both a survey component and automated harvesting as central, code as "hybrid (automated + manual/survey)".
- Specialised registries (Europe PMC, ClinicalTrials.gov, EUCTR, DataON) → "other / specialised" unless they are themselves clearly wrappers around open infrastructures.
- Code the proprietary-input flag as "yes" whenever any subscription database (WoS, Scopus, Dimensions, Embase, etc.) is used as a core input, even within a hybrid typology.
- Code the proprietary-input flag as "no" only when no proprietary source is mentioned as a core input.
- Initiatives using only their own internal data (e.g., a publisher monitoring its own journals) are coded "other / specialised" with a note.

V4. Data source identification

Variable ID	V4
Variable name	Data source identification
What it captures	The specific named data sources the initiative uses, recorded as a multi-select list using a controlled vocabulary of standardised short-name labels. This variable

	provides the empirical backbone of the infrastructure-dependency analysis and complements the broader typology coded in V3.
Source columns	Primary: "Data types/sources used for Monitoring". Secondary: "Monitoring Mechanism / Approach"; "Brief Description" where additional sources are mentioned.
Variable type	Multi-select. Up to ~8 sources per initiative; coders should not exceed this without justification, and should prioritise core inputs over enrichment sources.
Notes on application	The controlled vocabulary is iterated during pilot coding. New sources encountered during coding that are not yet in the vocabulary are flagged for WG2 review and added to the vocabulary before full coding continues. Naming consistency is critical: "OpenAIRE Graph", "OpenAIRE", and "OpenAIRE Research Graph" must all be coded under one canonical label.

Controlled vocabulary

OpenAIRE Graph	Open infra.	OpenAIRE, OpenAIRE Research Graph, OpenAIRE-G
OpenAlex	Open infra.	OpenAlex API
Crossref	Open infra.	Crossref API, Crossref FundRef
Unpaywall	Open infra.	Unpaywall API
DataCite	Open infra.	DataCite Commons
DOAJ	Open infra.	Directory of Open Access Journals
DOAB	Open infra.	Directory of Open Access Books
Sherpa/RoMEO	Open infra.	RoMEO, Sherpa Romeo
Lens	Open infra.	Lens.org
OpenAPC	Open infra.	Open APC
ROR	Open infra.	Research Organization Registry
ORCID	Open infra.	—
Zenodo	Open infra.	—

Web of Science	Proprietary	WoS, Clarivate WoS
Scopus	Proprietary	—
Dimensions	Proprietary	Dimensions.ai
PubMed	Open infra.	—
Embase	Proprietary	—
Europe PMC	Specialised	EuropePMC
ClinicalTrials.gov	Specialised	CT.gov
EUCTR	Specialised	EU Clinical Trials Register
CHORUS	Specialised	—
CRIS	Institutional	Pure, Converis, VIVO, Serval, ZORA, VIRTA, COBISS, HAL, CRIStin
Institutional repository	Institutional	IR, repository
Survey	Specialised	—
Internal proprietary data	Specialised	—
DataON	Specialised	—
HEAL-Link agreements data	Specialised	—
LAS:eR	Specialised	LASeR
FinELib consortium records	Specialised	—
N/A	N/A	Unknow

Decision rules

- Code only sources explicitly mentioned in the dataset entry. Do not infer the presence of sources that are commonly used but unnamed.
- Use the canonical label from vocabulary. If a source is not in the vocabulary, flag it for WG2 review and add it to the vocabulary; do not invent a non-standard label.

- Where a source is referenced via a wrapper or aggregator (e.g., Unpaywall data accessed via OpenAlex), code both the wrapper and the underlying source if both are visible in the dataset entry.
- Institutional CRIS systems are coded as "CRIS" with a note indicating which CRIS where this is specified (e.g., "VIRTA", "Pure", "Serval").
- Surveys are coded as "Survey" with a note indicating which survey instrument or population (e.g., "biannual TSV survey").
- Where an initiative uses internal proprietary data not corresponding to any external database, code as "Internal proprietary data" with a free-text descriptor in the notes column.
- Maximum of ~8 named sources per initiative. If more are listed, code the ones most central to the monitoring and note the others in the notes column.

V5. Open science pillars monitored

Variable ID	V5
Variable name	Open science pillars monitored
What it captures	Which open science pillar(s) the initiative monitors, recorded as a multi-select list using a controlled vocabulary of canonical pillar labels. "Impact" and "Cost & Funding Transparency" are treated as pillars in their own right and are not folded into other categories.
Source columns	Primary: "Open Science Focus / Pillars Monitored".
Variable type	Multi-select. An initiative may, and typically will, receive more than one pillar code.
Notes on application	The source column is a structured field in the dataset, but contributor entries were populated inconsistently: synonyms, typos, and non-atomic entries (a single item containing an internal comma, e.g. cost/funding phrasing).

Controlled vocabulary

1	Open Access	Open access to publications and related outputs. Raw variants folded in: "Open Access"; "Open Access Journals"; "DOAJ indexing"; "Preprints" (incl. typo "Prepints"); "Open licenses (Creative Commons)".

2	Open Research Data	Open research data and FAIR data practices. Raw variants folded in: "Open Research Data"; "FAIR Data / FAIR Principles".
3	Open Research Software / Code	Raw variants folded in: "Open Research Software / Code".
4	Open Methods & Protocols	Open methods, protocols, and study/trial registration. Raw variants folded in: "Open Methods / Protocols"; "Study registrations"; "Results reporting for clinical trials".
5	Open Peer Review	Raw variants folded in: "Open Peer Review".
6	Open Hardware	Raw variants folded in: "Open Hardware".
7	Open Educational Resources	Raw variants folded in: "Open Educational Resources (OER) / Open Learning".
8	Citizen Science & Public Engagement	Raw variants folded in: "Citizen Science"; "Public Engagement".
9	Impact	Coded as a standalone pillar per WG2 decision, separate from Citizen Science & Public Engagement (code 8). Raw variants folded in: "Impact"; "Societal Impact".
10	Open Infrastructure & PIDs	Technical infrastructure and persistent identifiers supporting monitoring. Raw variants folded in: "Open Infrastructure & Tools"; "PID"; "ORCID ID".
11	Research Assessment & Evaluation Reform	Raw variants folded in: "Research Assessment & Evaluation Reform"; "CRediT taxonomy"; "Gender"; "Transparency (funding statements, limitations, conflict of interests etc.)".
12	Open Collaboration	Raw variants folded in: "Open Collaboration".
13	Open Science Policy	Raw variants folded in: "Open Science Policies".
14	Cost & Funding Transparency	Coded as a standalone pillar per WG2 decision, not folded into Open Access (code 1). Raw variants folded in: "Funding"; "Open Access Cost (subscriptions, transformative agreements, and publisher payments (APC))"; "Transformative agreement(s)"; "Open Access Cost (APC)".

99	Other / unclear	Entry present in the source column but does not map to any vocabulary term above (e.g., "Thesis").

Decision rules

- Treat the raw "Open Science Focus / Pillars Monitored" field as a set of independent items, not literally as comma-delimited text: some individual items contain an internal comma (e.g. the Open Access Cost / transformative agreements / APC phrasing), so a naive split produces spurious fragments. Reconstruct the intended items from context before coding.
- Code every pillar explicitly present in the source entry. This is a multi-select variable; most initiatives will receive more than one pillar code.
- Correct obvious typos before mapping (e.g. "Prepints" maps to "Preprints", coded under Open Access, code 1).
- "Impact" and "Societal Impact" are coded under the standalone Impact pillar (code 9). Do not fold Impact into Citizen Science & Public Engagement (code 8), even though the two are conceptually related.
- "Funding", "Transformative agreement(s)", and "Open Access Cost" entries are coded under the standalone Cost & Funding Transparency pillar (code 14). Do not fold these into Open Access (code 1).
- ORCID ID and other identifiers are coded under Open Infrastructure & PIDs (code 10), not under Open Access.
- Use the canonical label from the vocabulary above. If a source entry does not map to any vocabulary term, code it "Other / unclear" (99).
- Where the source column is blank, code as 99 (unclear).

Appendix A. Codebook change log

All changes to this codebook between versions are recorded here, with date, version, summary of change, and rationale. This log is essential for reproducibility and for any future re-use of the codebook.

Version	Date	Description	Notes	Changes made by
1	5/26/2026	Initial draft of codebook with five variables: V1 purpose archetype, V2 compliance linkage, V3 data source typology, V4 data source identification	Initial release	Tereza Szybisty
2	8/19/2026	Added V5 "Open science pillars monitored": controlled-vocabulary crosswalk (14 canonical pillars + 99 unclear) for the free-text "Open Science Focus / Pillars Monitored" field. Updated Section 1.2 to remove this field from the list of structured fields excluded from coding.	Raw field contained synonyms, typos, and non-atomic (comma-containing) entries that a straight tally misreads as separate categories - this will be fixed in the data set.	Tereza Szybisty
3	11/8/2026	Added “How to cite” and OSMI logo.	Preparation for upload to Zenodo.	Tereza Szybisty