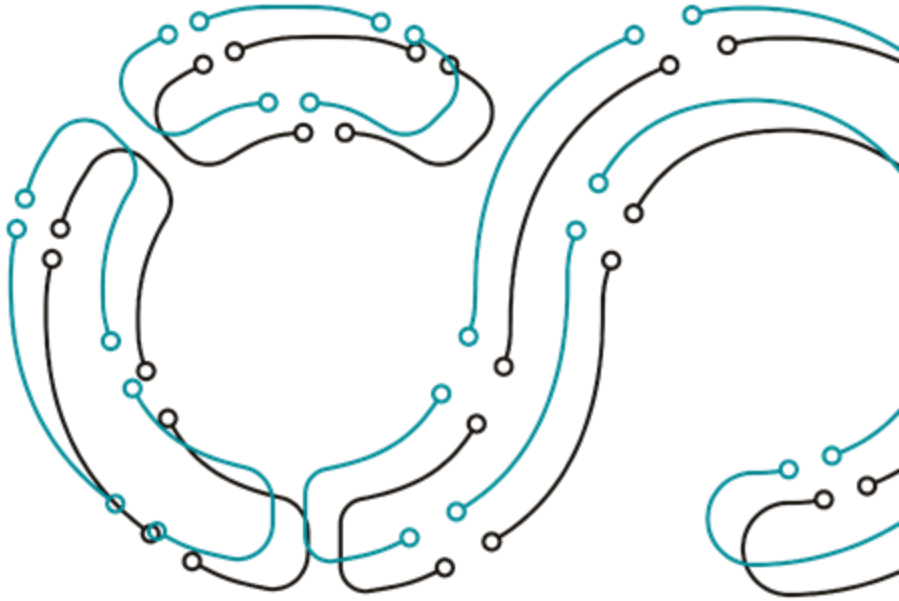


Rich inside, compliant outside

Onboarding the catalogues of a thematic node: lessons from the EOSC Node Data Terra

EOSC Academy · Course 5 · 9 November 2026

Julien Homo · technical lead of the EOSC Node Data Terra catalogues (Foxcub)



A thematic node for Earth system data

WHO BRINGS CONTENT

Data Terra: French research infrastructure for Earth system data, with its data and service hubs: AERIS (atmosphere), ODATIS (ocean), THEIA (land surfaces), FormaTerre (solid Earth), PNDB (biodiversity).

NFDI4Earth: German NFDI consortium for Earth system sciences, harvested as a partner source (Knowledge Hub, including GFZ Data Services and PANGAEA records).

DISCIPLINES

Atmosphere | Ocean | Solid Earth | Land surfaces | Biodiversity

EACH BRINGS ITS OWN STANDARD

Resource	Native standard
Data	ISO 19115, GeoDCAT-AP
Software	CodeMeta
Publications	DataCite, HAL
Services	EOSC service profile

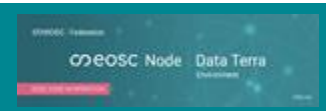
PRODUCTION ON 1 NOVEMBER 2026

14,338
records described in the node knowledge graph

8
harvested sources

360
records live in EOSC: 257 datasets, 82 publications, 21 services

Our job as a thematic node: bring to EOSC the domain resources that nobody else describes as well, without losing what the communities say about them.



Eight node capabilities, three operational today

OPERATIONAL
Resource Catalogue
REST (DCAT) and OAI-PMH, version 1.0.0

OPERATIONAL
Service Catalogue
REST (DCAT) and EOSC service catalogue API, version 1.0.0

OPERATIONAL
AAI
OIDC, Gaia-Data single sign-on

PLANNED
Monitoring

PLANNED
Helpdesk

PLANNED
Accounting

PLANNED
Order Management

PLANNED
Application Workflow Management

- ■ **Operational**
- □ **Planned (dotted outline)**

Today's focus: the two catalogues, where every onboarded resource and service becomes visible to the Federation.

The descriptor itself is a capability: a peer node discovers our endpoints, formats and status without exchanging emails.

Two catalogues, one knowledge graph

SOURCES

8 harvested sources

MTEP · PNDB · AERIS · NFDI4Earth · Research software catalogue · HAL · GEODES · Service register

Protocols

SPARQL, REST, STAC, curated register

NODE KNOWLEDGE GRAPH

Node knowledge graph

DCAT 3 pivot

all types, all sources

TWO CHANNELS

Service Catalogue

`GET /services`, EOSC service catalogue API, harvested by the EOSC EU Node.

Resource Catalogue

OAI-PMH with `oai_openaire`, registered in OpenAIRE PROVIDE, aggregated into the EOSC knowledge graph; plus `oai_dc` and `dcat` formats and a DCAT 3 JSON-LD feed for peer nodes.

One catalogue, one type The Node Registration Policy asks each catalogue to contain only its own type of item. In version 1.0.0 (17 October 2026) we removed our 21 services from the research product feed: they now reach EOSC only through the service catalogue.

Key lesson **Read the Node Registration Policy as a design rule: one type per catalogue, one channel per type.**

What makes the catalogue rich

1 One typed DCAT 3 resource

- Every record is a `dcat:Resource` typed by a COAR concept, not forced into nine subclasses.
- Common core for all types: identity, dates, creators, licence and access rights, themes, spatial and temporal coverage, distributions, typed identifiers, provenance.
- Thin per-type extension borrowed from the community standard.

2 Values are concepts, not strings

- Controlled vocabularies referenced by IRI: COAR resource types and access rights, EuroSciVoc (3 levels), EOSC service vocabularies, SPDX, ROR, ORCID, ISO 639, OGC EPSG, the Data Terra thesaurus.
- A concept carries its label, definition and alignments: a receiving node can filter, translate and reason on it.

3 Identity and history of every record

- A stable node identifier for each record on w3id.org; native PIDs kept and typed (2,026 DOIs, 179 software heritage IDs, 58 HAL IDs).
- Provenance in PROV-O: primary source, attribution, every processing step.
- Deduplication that destroys nothing: 4,076 records grouped in cross-source clusters, every member kept.

4 Open by default

- OAI-PMH in 3 formats with sets and deletion tombstones.
- DCAT 3 JSON-LD feed with filters by period and bounding box.
- Node descriptors under `/.well-known`, published vocabulary, public validation rules (slide 9).

14,338 records described richly

360 projected to EOSC today

EOSC compliance is a projection of a richer model. We never lose what the source says to fit a profile.

Two journeys to green light

May 2026	Jun	Jul	Aug	Sep	Oct 2026

RESEARCH PRODUCTS (RESOURCE CATALOGUE)

3 May start of development

30 Aug first production release (0.2.0)

9 Oct OAI-PMH interface confirmed by the OpenAIRE PROVIDE helpdesk

15 Oct first OpenAIRE aggregation, 338 records collected, 336 transformed without error

17 Oct release 1.0.0 in production

20 Oct **Resource Catalogue shown "Onboarded, Active" on the EOSC portal**

SERVICES (SERVICE CATALOGUE)

Late July service catalogue work starts on the same knowledge graph

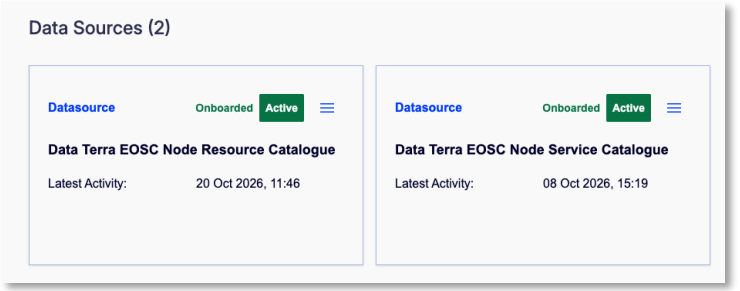
29 Sep `GET /services` channel in production

2 Oct service catalogue declared compliant and approved by the EOSC EU Node onboarding team

8 Oct **Service Catalogue shown "Onboarded, Active" on the EOSC portal**

How we kept the deadline

- **Small, frequent releases:** 103 versions in production between May and October.
- **Automated checks before every release:** 527 tests at 1.0.0, continuous integration that refuses a broken release.
- **Regular work with the node's catalogue lead and coordinator:** decisions taken fast, close to the data.
- **Direct line with the EU Node onboarding team and the OpenAIRE helpdesk:** questions asked early, answers turned into code.



Why research products took longer

Services were faster because

- One curated register, owned by the node (21 services).
- One contract: the EOSC EU Node service profile, 10 mandatory fields.
- One channel, harvested by the EU Node, validated directly in production.

RESEARCH PRODUCTS WERE SLOWER BECAUSE

1 Heterogeneous sources.
8 sources, several harvesting protocols (SPARQL, REST, STAC), native standards from ISO 19115 to DataCite and CodeMeta. One mapping per source, property by property.

2 Two references that disagree.
The EU Node registration guide and the OpenAIRE XML schema did not always say the same thing; we documented the contradictions and fed them back.

3 Moving specifications.
The research product profile changed several times during the project (v6.1, v7, v7.1, v7.2), and the terms "resource", "research product" and "service" vary from one document to another.

4 Metadata missing at the source.
13,876 records have no access rights, 8,334 no creator, 6,014 no publication date. A catalogue cannot publish what the source does not say.

Key lesson **Services are a register you control; research products are a federation of other people's metadata. Plan the time accordingly.**

From 14,338 records to 360 in EOSC

8 — 1 Sources selected by the node (curator decision)

14,338 — 2 Records harvested and described in the knowledge graph

4,076 — 3 Records grouped as cross-source duplicates, never deleted

360 — 4 Compliant with the EU Node profile and published: 339 research products, 21 services

ROLES AND APPROVAL

Node catalogue lead (curator): chooses which sources and providers enter the node; services are kept in a curated register.

The rules: every record is checked automatically against the EU Node profile at each harvest; only compliant records are published, the others stay visible internally with the reason.

Providers: stay owners of their metadata; the node records what is missing, field by field, so that it can be fixed at the source.

Key lesson **A thematic node can keep a light, curator-led process, as long as the record-level decision is made by public, testable rules.**

Quality today: the rules are code, and they are public

THE EEN 1.0.0 PROFILE IN ONE TABLE

Applies to	Required	Severity
Every research product	title; publication or creation date; resource type; identifier; recognised access rights (on the resource or a distribution)	**blocking**
Datasets, in addition	creator or originator	**blocking**
Every research product	recognised persistent identifier (DOI, Handle)	warning only

Published at: resources.earth-data.eu/shapes/eosc-een/1.0.0.ttl

TWO VIEWS, ONE DECISION

Strict view
(public, sent to EOSC): only compliant records.

Complete view
(internal back-office): every record, with the reason for each exclusion, field by field.

Never invent

When a value is missing, the record is flagged, not filled. No default licence, no guessed author, no guessed access right.

Open to providers

The shapes are public: providers, partner projects and other nodes can check their own records against our profile before connecting. A partner project already did so in October 2026.

Key lesson **Write your quality criteria as testable rules and publish them: providers can check themselves before you do.**

Quality tomorrow: a dashboard to grow what we publish

MEASURED ON 1 NOV 2026, ON THE 13,978 NON-COMPLIANT RECORDS

4,454

records are one field away from compliance

of which **4,428** miss only their access rights

4,556 miss two fields, 4,968 miss three

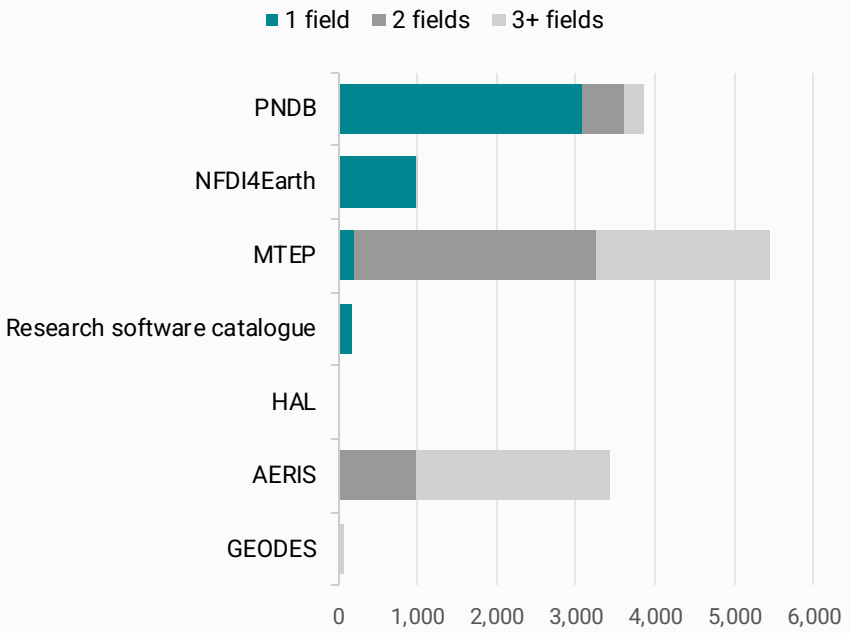
If providers declare access rights, the node could go from 360 to about 4,800 records in EOSC, a factor of 13, without inventing anything.

Why DCAT-AP tomorrow: the EU Node profile is the common floor. The Federation discussion goes further: DCAT-AP mappings and testing ranked second among the priorities of the WG Catalogues survey (October 2026), and a rich DCAT contract between thematic nodes is the next layer. The dashboard measures both.

PLANNED · MOCK-UP, NOT BUILT YET

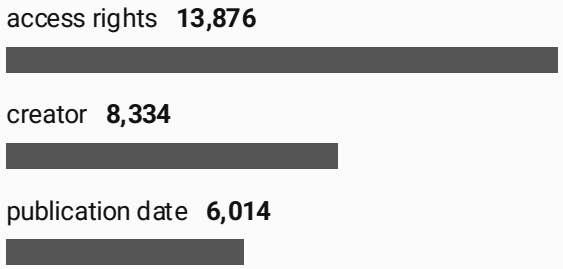
1 By provider

non-compliant records, by number of missing blocking fields



2 By field

which field blocks most records



3 By profile

the same record scored against two profiles

EU Node profile today

DCAT-AP tomorrow

Key lesson **Measure distance to compliance, not just pass or fail: it tells each provider exactly what to fix first.**

Five lessons for nodes onboarding their catalogues

KEY LESSON

1

Start with services.

A curated register and one contract give you an early, visible success.

KEY LESSON

2

Rich inside, compliant outside.

Keep your full domain description and project the EOSC profile from it; never cut your model down to the profile.

KEY LESSON

3

Write criteria as testable rules, and publish them.

Providers, partners and other nodes can check themselves.

KEY LESSON

4

Never invent metadata.

Flag what is missing and go back to the source; trust is the real onboarding criterion.

KEY LESSON

5

Talk early to the EU Node and OpenAIRE.

Most delays came from ambiguities that one email could clear.

WHAT COMES NEXT planned

- The five planned capabilities: monitoring, helpdesk, accounting, order management, workflows.
- Node-to-node exchange in DCAT 3, discovered through `/.well-known`, already tested with a peer node.
- A hackathon between thematic nodes on real research product feeds, proposed to the WG Catalogues for late November or early December 2026.

resources.earth-data.eu · services.earth-data.eu · julien.homo@foxcub.fr