

High-value datasets on data.europa.eu: what a machine sees, and what Member States say

EasyxLab · EasyByte Hub S. Coop. Mad. · Study S17

Working draft Published 2026-10-03

Abstract

The rules on high-value datasets (HVD) in Commission Implementing Regulation (EU) 2023/138 have applied since 9 June 2024. Twenty-eight months later, on 2026-10-03, we read every record the European data portal flags as HVD and checked it against the regulation and the DCAT-AP HVD 3.0.0 profile. We then compared the result with each Member State's answers to the Open Data Maturity (ODM) 2025 questionnaire. For 7 of the 27 Member States the portal does not identify any dataset as HVD. Three of those seven (Cyprus, Slovenia, Slovakia) told ODM 2025 that they had implemented the DCAT-AP HVD tag. The portal flags 27,972 records, 27,526 distinct datasets. Record counts reflect how finely each country catalogues: 10,032 come from one German municipal geoportal. Another 13,948 records carry an HVD category without the regulation's ELI. All distributions carry a licence that meets Art. 4(3) in 17,104 of 27,972 records (61.1%). Most of the rest fail only on a German catch-all licence value. 5,146 of 27,972 (18.4%) have the Data Service that DCAT-AP HVD uses to express the API obligation. In a stratified sample, 1,387 of 1,510 requested distribution URLs (91.9%) answered with a 2xx status. Poland flags HVDs in its own fields, but its DCAT-AP export carries no HVD property and writes licences under an undefined property, so the European portal receives neither.

1. Introduction

The Open Data Directive (EU) 2019/1024 created the category of high-value datasets. Implementing Regulation 2023/138 lists them in six themes: geospatial, earth observation and environment, meteorological, statistics, companies, and mobility. For these datasets the regulation sets requirements that a machine can partly check: how they are licensed, whether an API serves them, and whether their metadata say they are HVD. That last requirement exists so that re-users, and the Commission, can find the datasets.

Member States are the obligated parties. They report to the Commission (Art. 5) and answer a yearly self-assessment, the ODM questionnaire. The European portal, data.europa.eu, harvests the national catalogues every day. It counts a record as HVD when the record carries the regulation's European Legislation Identifier (ELI) as applicable legislation.

We ask what that portal shows on one date and how it compares with what Member States say. We take the portal's view literally. A count of 0 means that the European portal does not identify any dataset of that country as HVD. It does not mean that the country publishes none. Where we could, we traced where records are lost between a national portal and the European one.

2. Prior work

We searched on 2026-10-03. We ran three Crossref queries (`query.bibliographic` , from 2023-01-01): "high-value datasets implementing regulation 2023/138", "high-value datasets metadata DCAT-AP" and "high value datasets open data directive API licence". OpenAlex returned HTTP 429 to all our queries. During the independent review, OpenAlex and Semantic Scholar also returned 429, and arXiv's export host disallows crawling in its robots.txt. So two of the indexes we wanted could not be searched. We also read the data.europa.eu publications on HVD and the ODM 2025 resources. Zenodo's API is disallowed by its robots.txt, so we read the record page and the PDF only.

- **ODM 2025** (data.europa.eu, December 2025) is self-reported. Table 8 (p. 36): *"25 Member States (93 %) report that their public bodies with HVDs have denoted this in the datasets' metadata."* On the DCAT-AP HVD tag the report gives two figures: *"17 Member States (63 %)"* (p. 11) and *"Twenty-one Member States (78 %)"* (Table 41, p. 91). The per-country answers to Q5 give 21. Technical items score lowest in HVD maturity: *"machine-readable formats via APIs (73 %) and bulk downloading (73 %)"* (p. 37). The ODM does not count records.
- **ETC DI Report 2025/5** (European Topic Centre on Data integration and digitalisation, Zenodo record 21377126, 2026-06-19) analyses the environmental HVDs that Member States submitted in the noise, marine and water domains; it "expresses the views of the authors". A portal search on noise returned *"151 datasets"* on 2025-04-04 and *"215 datasets"* on 2025-10-09 (p. 18). The report finds that *"A significant percentage of datasets (24%) do not provide the HVD category at all"* (p. 33).
- **"High-value datasets: Cross-country findability and comparability of metadata"** (data.europa.eu, 27 September 2024) assesses 24 HVDs from four countries. It reports *"absence of regulatory references"*.

- **Papers found through Crossref.** These cover company HVDs ("The STIRData Approach to Interoperability of European Company High-Value Datasets", SN Computer Science 2024, doi:10.1007/s42979-024-02721-8), HVD opening in Taiwan (dg.o 2024, doi:10.1145/3657054.3657089) and HVDs for smart-city services (ICECET 2025, doi:10.1109/icecet63943.2025.11472439). Others cover DCAT-AP quality assessment (LNCS 2023, doi:10.1007/978-3-031-41138-0_17), HVD availability before the date of application (LNBIP 2023, doi:10.1007/978-3-031-43126-5_21), DCAT-AP national profiles and harvesting (EGOV-CeDEM-ePart 2026, doi:10.59490/6a91adf70fd2bd55cd254b0f), and identifying candidate HVDs in US cities (Telematics and Informatics 2026).

What is new. In the indexes we could reach, no work does any of the following:

- (a) checks every HVD record on the European portal against Arts. 3(5), 4(3) and 3(1), with a written licence-equivalence table and an RDF confirmation of "no licence";
- (b) counts the half-tagged and malformed records per catalogue, with duplicates measured;
- (c) compares the census with each country's ODM 2025 answers;
- (d) traces where HVD flags and licences are lost in harvesting.

3. Legal and technical baseline

We checked the text in force on 2026-10-03 in Cellar (CELEX 32023R0138). The regulation has not been amended. Its three corrigenda (2024-07-12, 2024-07-26, 2024-10-22) concern only the Greek, Spanish and Slovenian versions. Cellar records no act amending Directive 2019/1024 or the Data Act (EU) 2023/2854. The Digital Omnibus proposal of 19 November 2025, COM(2025) 837, would repeal Directive 2019/1024, but on the measurement date it was still a proposal. The HVD obligations were unchanged.

- Art. 3(1): *"Public sector bodies holding high-value datasets listed in the Annex shall ensure that the datasets described or referenced in the Annex are made available in machine-readable formats via APIs corresponding to the reasonable needs of re-users. Where indicated in the Annex, the datasets shall also be made available as a bulk download."*
- Art. 3(5): *"Public sector bodies holding high-value datasets listed in the Annex shall ensure that the datasets are denoted as high-value datasets in their metadata description."*
- Art. 4(3): *"High-value datasets shall be made available for re-use under the conditions of the Creative Commons Public Domain Dedication (CCO) or, alternatively, the Creative Commons BY 4.0 licence, or any equivalent or less restrictive open licence, as set out in the Annex, allowing for unrestricted re-use. A requirement of attribution, giving the credit to the licensor, can additionally be required by the licensor."*
- Art. 5(1): *"By 2 years after entry into force of this Implementing Regulation Member States shall provide the Commission with a report on the measures they have carried out to implement this Implementing Regulation."*
- Art. 6: *"It shall apply from 16 months after entry into force of this Regulation."* That date is 9 June 2024. The 16 months run from entry into force to application. From application to our census, 28 months passed.
- **Annex and APIs.** APIs are required in all six categories. There are two relaxations. For earth observation and environment, bulk download applies *"(for historical versions of datasets: APIs or bulk download, as feasible and appropriate)"*. For meteorology, *"NWP model data' ... shall be made available through APIs only"*, which relaxes the bulk download, not the API.

DCAT-AP HVD 3.0.0 (a SEMIC Recommendation published 2024-10-25) gives Dataset, Distribution and Data Service `applicable legislation 1..`, with *"For HVD the value must include the ELI http://data.europa.eu/eli/reg_impl/2023/138/oj"*. Dataset and Data Service have `HVD category 1..` *"A Dataset MUST have an associated Data Service, besides the exceptions listed in the HVD IR."* For licences, a licence is preferred, and *"As alternative rights can be used."*

4. Data

- **Census.** We read the data.europa.eu search API (`filter=dataset` , `facet is_hvd=true`) by scroll from 11:11:39 to 11:12:41 UTC. It returned 27,972 records with 152,679 distributions. A facet count taken right after gave the same number, and an independent re-count at 11:36:56 UTC found no drift.
- **SPARQL** (<https://data.europa.eu/sparql>). We queried the half-tagged records and their IDs, the ELI variants and the Data Services, and confirmed the "no licence" cases. Every query and its timestamp are in `data/queries.json`.
- **ODM 2025.** We used the per-country answers to P11, P12, Q5 and Q6, and the report. URLs and sha256 are in `data/sources.json`.
- **Poland.** We used the `api.dane.gov.pl` flags and 11 pages of its DCAT-AP endpoint. Six of those pages contain EU-list HVDs.
- **Reachability.** A stratified sample of 1,699 distribution URLs from 435 hosts in 20 Member States.

We kept no contact points; the published files hold dataset IDs, flags and aggregates.

5. Method

We wrote the rules, the licence table and the sample into `METHOD.md` and hashed it before the main collection. The hash was recorded, but the frozen text was not kept separately, so the freeze is self-attested. Corrections and deviations D1-D6 are listed in `METHOD.md` §9.

- **Tagging (Art. 3(5)).** A record passes if it has the exact ELI and at least one HVD category. Any other IRI naming the regulation counts as "malformed".
- **Licence (Art. 4(3)).** Each distribution's licence value goes into one class:
 - A: CCO or CC BY 4.0;
 - B: equivalent or less restrictive, from a written list (DL-DE-BY/Zero 2.0, Etalab, the Flemish and Croatian open licences, ODC-BY/PDDL, older CC BY...);
 - C: share-alike;
 - D: non-commercial, no-derivatives or closed;
 - E: a value that names no specific terms;
 - N: no value. A record passes if every distribution is A or B. Every record with no licence in the index was re-checked in the RDF (`dct:license` and `dct:rights`), and 50 were re-read from the repository API.
- **API (Art. 3(1)).** "Modelled" means `dcat:accessService` on a distribution, or a `dcat:DataService` that `dcat:servesDataset` the record. "Inferable" means only that a format or URL looks like a service.
- **Reachability.** Up to 100 URLs per Member State (seed 17), at most 10 per host. Every request goes through one helper that reads `robots.txt` first and allows 1 request per second per host. Data URLs get a `GET` with `Range: bytes=0-2047`, and OGC services a `GetCapabilities`. Timeout 20 s, no retries.
- **Duplicates and sources.** The portal adds `~~n` to an ID when the same source ID arrives twice. We merge records by base ID, and we read the source host embedded in each ID.

6. Results

6.1 Census

The portal flags **27,972** records as HVD: 27,909 from Member States and 63 from EU institutions, Norway, Iceland and Serbia. Merged by base ID they are 27,526 distinct datasets. The 446 surplus records are almost all Spanish (421 of Spain's 1,879, where the same dataset arrives through `codsi`, `datos-gob-es` and `idee`); Latvia has 17, Italy 6 and Germany 2. Earth observation and environment is the most frequent category (21,621 records; a record can carry several), then geospatial (2,275), statistics (1,689), mobility (864), meteorological (375) and companies (119).

For Bulgaria, Cyprus, Hungary, Poland, Romania, Slovenia and Slovakia, the European portal does not identify any dataset as HVD. On the portal those countries have 11,850, 1,948, 2,283, 25,884, 5,491, 9,423 and 2,095 datasets of any kind.

Record counts are not comparable across countries. They measure how finely a portal splits its HVDs into records. Germany has 19,162 records, but 10,032 of them come from one Rhineland-Palatinate municipal geoportal (52.4% of Germany, 35.9% of the whole census). At least 12,798 (66.8% of Germany) come from that Land's geoportals, identified by the host embedded in the record ID (`data/source_concentration.csv`). One Annex dataset, such as land-use plans, can appear as one record per municipality.

MS	PORTAL DATASETS	HVD RECORDS	DISTINCT	CATEGORY, NO ELI	ELI, NO CATEGORY	LICENCE PASS	NO LICENCE (RDF)	API MODELLED	ODM P12	ODM Q5
AT	60,114	32	32	1,996	0	32	0	31	yes	yes
BE	19,020	759	759	7	61	705	32	565	yes	yes
BG	11,850	0	0	0	0	0	0	0	no	no
CY	1,948	0	0	0	0	0	0	0	yes	yes
CZ	30,476	282	282	0	0	0	0	200	yes	yes
DE	850,554	19,162	19,160	5,994	9	11,213	2	40	yes	yes
DK	3,727	137	137	0	0	132	3	124	yes	yes
EE	8,551	87	87	10	13	0	0	48	yes	no
ES	130,645	1,879	1,458	11	6	1,441	40	1,506	yes	yes
FI	2,950	49	49	9	0	28	21	9	yes	yes
FR	183,229	766	766	0	60	425	23	573	yes	yes
GR	22,793	1,281	1,281	0	11	1,168	86	725	no	no
HR	4,802	78	78	0	7	77	0	0	yes	yes
HU	2,283	0	0	0	0	0	0	0	yes	no
IE	22,897	345	345	4,247	42	293	2	0	yes	yes
IT	88,966	186	180	30	9	178	0	138	yes	yes
LT	2,570	562	562	0	329	351	60	121	yes	yes
LU	3,695	410	410	5	1	393	1	409	yes	yes
LV	2,017	622	605	2	3	0	607	0	yes	yes
MT	840	158	158	0	2	155	0	3	yes	yes
NL	30,090	451	451	0	422	23	421	291	yes	yes
PL	25,884	0	0	0	0	0	0	0	yes	no
PT	21,000	305	305	0	52	243	0	267	yes	yes
RO	5,491	0	0	0	0	0	0	0	yes	no
SE	25,152	358	358	146	6	194	156	47	yes	yes
SI	9,423	0	0	0	0	0	0	0	yes	yes
SK	2,095	0	0	0	0	0	0	0	yes	yes

Table 1. Per Member State on 2026-10-03, in alphabetical order ([data/by_member_state.csv](#)). Record counts reflect how finely each portal splits datasets into records and are not comparable between countries. "Category, no ELI" records are not counted as HVD by the portal; 1,499 more are not linked to a catalogue. ODM answers are from 2025.

6.2 Half-tagged and malformed records

13,948 records carry an HVD category but not the ELI. They are 13,054 distinct base IDs, and 992 of the records share a base ID with a record already counted as HVD. Among those 992, 255 are twins of German govdata records that carry the ELI (183 of them in the geodata catalogue gdi-de, 72 not linked to a catalogue). The rest are 12,062 distinct datasets not counted as HVD in any form. 12,345 of the 13,948 records come from five catalogues: gdi-de (5,958), data-gov-ie (4,247), bev-at (1,099), bmlfuw-at (897) and geodata-se (144). Another 1,499 are not linked to a catalogue in the SPARQL store.

The portal counts 345 Irish HVD records, while 4,247 Irish records carry a category without the ELI. Ireland's own ODM answer to Q5 says *"it is unknown [sic] if HVDs through the harvesters are appearing as HVDs in the portal"*.

1,033 of the 27,972 HVD records (3.7%) carry the ELI without a category, most of them in the Netherlands (422) and Lithuania (329). The store holds 25 triples with a malformed ELI, in five variants; no dataset carries only a malformed one. 133,332 of 152,679 distributions (87.3%) carry the ELI at distribution level. Croatia's 199 carry none.

6.3 Licences (Art. 4(3))

Under the frozen rules, **17,104 of 27,972 records (61.1%)** have every distribution in class A or B. 7,483 (26.8%) have a distribution with a share-alike, non-commercial, or closed/catch-all licence value, and 7,304 of them are German.

Sensitivity. 5,855 of the 7,483 fail only because of `http://dcat-ap.de/def/licenses/other-closed`, which DCAT-AP.de labels "other closed licence". It is a catch-all: the portal itself types the node `adms:licencetype/UnknownIPR`. The frozen table is inconsistent here. It puts `other-open` in class E (unspecific) but `other-closed` in D. With `other-closed` in E, 1,628 records (5.8%) fail and 7,432 are unclear. Neither reading tells what the publisher's terms are. In Rhineland-Palatinate alone, 6,963 records fail under the frozen rule.

`data/licences_by_dataset.csv` lists every licence value per record, so the verdicts can be recomputed under any classification.

Of the 1,808 records with no licence in the index, 127 have one in the RDF (an inline licence document), and 225 have only `dct:rights`. **1,456 (5.2% of the census) have neither.** All 50 re-read from the repository API agreed. These cases are concentrated in Latvia (607 of 622), the Netherlands (421 of 451) and Sweden (156 of 358). The residual uncertainty is that a licence may exist on a page the metadata do not link. The index also understates conformance for Czechia. All 282 Czech records are "unclear" because the index shows a terms-of-use node. In the RDF, 839 of those 844 nodes point to CC BY 4.0 or CC0, and 5 to CC BY-SA 4.0. Deviation D1 (missed spellings) would move 709 records from unclear to pass.

6.4 API signal (Art. 3(1))

5,146 of 27,972 records (18.4%) have a modelled Data Service. Another 18,145 have an API-like distribution, and 4,681 have neither. The Annex requires APIs in every category (only historical earth-observation versions may be bulk-only), so exemptions explain little of the gap. But the count is per record, not per Annex dataset. A municipality's land-use plan may be one record of a dataset whose API is described elsewhere.

6.5 Reachability

Of 1,699 sampled URLs, 188 were not requested: robots.txt disallowed 97, and for 91 the robots.txt could not be read. One URL was malformed. Of the 1,510 requested, **1,387 (91.9%)** answered 2xx. The rest split as follows:

- 45 not found (404/410);
- 35 authentication required or forbidden (401/403);
- 22 other 4xx, 19 of them 416 Range Not Satisfiable ;
- 5 server errors;
- 16 network errors.

Weighted by each Member State's share of URLs, the rate is 95.5%. By country it ranges from 58.1% (Malta, 25 of 43) to 100% (Austria, Estonia, Luxembourg). A 2xx is not proof of content: 301 of the 998 successful ranged GETs returned HTML. After the run we found that our helper did not follow redirects on robots.txt (deviation D5). Without the 7 hosts affected, the rate is 1,361 of 1,484 (91.7%).

6.6 Self-report against census

ODM 2025 ANSWER	PORTAL ≥ 1 HVD	PORTAL 0
P12 "denoted in metadata": yes (25)	19	6 (CY, HU, PL, RO, SI, SK)
P12: no (2)	1 (GR)	1 (BG)

ODM 2025 ANSWER	PORTAL ≥ 1 HVD	PORTAL 0
Q5 "DCAT-AP HVD tag": yes (21)	18	3 (CY, SI, SK)
Q5: no (6)	2 (EE 87, GR 1,281)	4 (BG, HU, PL, RO)

P12 is the looser question. The six "yes, but 0" countries describe national marks in their own ODM text. Poland has a national flag. Romania tags HVDs *"through the keywords metadata"*. Cyprus uses a bilingual tag, Hungary a separate HVD catalogue, Slovenia a national attribute and Slovakia local catalogues. None of these reaches the European portal as the ELI it counts. On Q5, which asks about the tag the portal reads, three states said yes and have 0 records (CY, SI, SK). Two said no and do have records (EE, GR). Each row compares a 2025 answer with a 2026 count.

Poland, traced. dane.gov.pl flags 120 datasets as HVD from the EU list (82 CC BY 4.0, 38 CCO) and 18,297 with its national flag. We matched 96 of the 120 to the portal's dane-gov-pl catalogue (9,989 records), and none has the ELI, a category or a distribution licence. We read 11 feed pages. The six that contain EU-list HVDs (IDs 137, 313, 733, 925, 1,499, 2,059 and 2,060) have no `applicableLegislation`, no `hvdCategory` and no mention of 2023/138. Every licence there is written as `dcat:license`, a property DCAT does not define, and none as `dct:license`. For these records, both the HVD flag and the licence are lost at the export, before harvesting.

There is a second, hypothetical loss point. The 24 EU-list HVDs not found all have IDs of 18,216 or more, and all 96 found have IDs of 9,156 or less. The feed advertises `hydra:lastPage` = page 500, at 20 per page (10,000 items, ending at ID 12,585), while `hydra:totalItems` is 28,132. Page 501 still serves data. This is consistent with the harvest stopping at the feed's advertised last page. We have not confirmed it with the operators (we contacted no one).

7. Limitations

- **One date, one portal.** All counts are for data.europa.eu on 2026-10-03, and the portal re-harvests daily.
- **Records are not datasets.** Counts depend on how finely each portal splits datasets into records, so they cannot be compared between countries. Duplicates are merged only by the portal's own `~n` suffix. Twins with unrelated IDs are not detected.
- **ODM time gap.** The answers describe 2025 and the census October 2026.
- **Metadata, not legal compliance.** We test only records already flagged. We cannot see datasets that should be flagged, nor which Annex item a record implements.
- **Licence classes are our criterion**, and the frozen table is inconsistent on `other-open` / `other-closed`. We report both readings and publish every value.
- **Reachability is a sample** (1,510 requests, one each, no retry; Wilson intervals in `data/reach_by_member_state.csv`). Robots-excluded URLs are not in the denominator. Deviation D5 affected 7 hosts.
- **The freeze is self-attested.**
- **Poland.** The paging explanation is a hypothesis, and we read 11 of more than 1,400 feed pages.

8. A per-record checker

`scripts/hvd_check.py <dataset-id | DCAT-AP file>` applies these rules to one record. It reads the record from the portal's repository API or from a local DCAT-AP file. It returns a JSON verdict that cites Art. 3(5), 4(3) or 3(1) for each check. By default it treats `other-closed` as a catch-all and warns, and `--strict` applies the frozen rules. A publisher can run it on its own DCAT-AP feed before harvesting. EasyxLab may release it as a free tool (`hvd-lint`).

9. Data, code and licences

Code is Apache-2.0, and data and text are CC BY 4.0. The portal metadata are CCO, and the ODM answers come from data.europa.eu. `scripts/run.sh` regenerates every figure offline from `data/`. Each network step lists its timestamp in `data/queries.json`, and the source documents have sha256 hashes in `data/sources.json`.

10. Competing interests

None. EasyByte has no open-data product or service. EasyxLab may release the checker in \$8 as a free tool.

AI agents (Claude, Anthropic) collected the data through the portal APIs and SPARQL endpoints, wrote and ran the analysis code and tests, and drafted the text. An independent AI reviewer then checked the work. It recomputed every figure from the published data with its own code, verified the legal, specification and ODM quotes against their sources, and audited 48 records against their RDF. It found one factual error (the elapsed time), weaknesses in how three headline figures were framed, and a robots.txt redirect defect. The corrections and added analyses are listed in `METHOD.md` §9, and `scripts/check_headline.py` asserts the headline numbers against the data.

References

- Commission Implementing Regulation (EU) 2023/138 (OJ L 19, 20.1.2023), CELEX 32023R0138, Cellar; corrigenda 32023R0138R(01)–(03).
- Directive (EU) 2019/1024; Proposal COM(2025) 837 (Digital Omnibus), 19.11.2025.
- SEMIC, DCAT-AP HVD 3.0.0, 2024-10-25, <https://semiceu.github.io/DCAT-AP/releases/3.0.0-hvd/>
- data.europa.eu, *2025 Open Data Maturity Report* and *2025_odm_questionnaire_data.xlsx* (URLs and sha256 in `data/sources.json`).
- ETC DI, *Ongoing initiatives for once-only publication of High-Value Datasets*, ETC DI Report 2025/5, Zenodo 21377126, 2026.
- data.europa.eu, *High-value datasets: Cross-country findability and comparability of metadata*, 2024.
- STIRData, SN Computer Science, 2024, doi:10.1007/s42979-024-02721-8; dg.o 2024, doi:10.1145/3657054.3657089; ICECET 2025, doi:10.1109/icecet63943.2025.11472439; LNCS 2023, doi:10.1007/978-3-031-41138-0_17; LNBIP 2023, doi:10.1007/978-3-031-43126-5_21; EGOV-CeDEM-ePart 2026, doi:10.59490/6a91adf70fd2bd55cd254b0f.
- RFC 9309, Robots Exclusion Protocol.

Cite this study

CITATION

EasyxLab (2026). High-value datasets on data.europa.eu: what a machine sees, and what Member States say. Study S17. EasyByte Hub S. Coop. Mad. <https://github.com/easybytehub/easyxlab/tree/main/studies/s17-high-value-datasets>

BIBTEX

```
@techreport{easyxlab_s17,
  title      = {High-value datasets on data.europa.eu: what a machine sees, and what Member States say},
  author     = {{EasyxLab}},
  institution = {EasyByte Hub S. Coop. Mad.},
  number     = {S17},
  year       = {2026},
  url        = {https://github.com/easybytehub/easyxlab/tree/main/studies/s17-high-value-datasets}
}
```