*Cambridge Prisms: Water*

**www.cambridge.org/wat**

# Open but not operable: Evaluating machine-ready metadata quality across EU water data portals for policy-relevant management tasks

Angelos Alamanos[1], Georgios Feretzakis[2] and Phoebe Koundouri[3,4,5,6]

[1]Independent Researcher, Germany; [2]ReSEES.AE4RIA Research Laboratory, Athens University of Economics and Business, Athens, Greece; [3]School of Economics, Department IEES, and ReSEES.AE4RIA Research Laboratory, Athens University of Economics and Business, Greece; [4]Peterhouse, University of Cambridge, UK; [5]SDU.AE4RIA, Athena Research Centre, Greece and [6]Global Climate Hub, Sustainable Development Solutions Network (SDSN), Greece

## Research Article

**Cite this article:** Alamanos A, Feretzakis G and Koundouri P (2026). Open but not operable: Evaluating machine-ready metadata quality across EU water data portals for policy-relevant management tasks. *Cambridge Prisms: Water*, **4**, e20, 1–15
https://doi.org/10.1017/wat.2026.10030

Received: 31 March 2026 UTC
Revised: 24 August 2026 UTC
Accepted: 02 September 2026 UTC

**Keywords:**
Water management; FAIR principles; operational readiness; water framework directive; early warning systems

**Corresponding author:**
Phoebe Koundouri;
Email: pkoundouri@aueb.gr

## Abstract

This study assesses the operational fitness and machine-readiness of EU water datasets for contemporary, data-intensive water governance. We develop the EU Water Dataset Observatory, a reproducible framework that evaluates metadata fitness for three management tasks aligned with the Water Framework, Floods and INSPIRE Directives, while assessing machine-readiness for early warning, compliance reporting and cross-border coordination. The framework combines metadata sufficiency scoring with a five-component impact proxy to prioritize improvements. We apply it to 2,176 water-related datasets retrieved through SPARQL from data.europa.eu, while accounting for retrieval failures that make the corpus a non-census subset. Mean sufficiency scores range from 0.16 to 0.24 (0–1), and none of the harvested records is operationally ready for any assessed task. Manual verification of 25 datasets corroborates these deficiencies: 52% of access URLs were broken, and none of 13 accessible source pages provided standardized machine-readable license information. Major federated-catalog gaps include absent machine-readable format declarations (100%), missing open-license metadata (99.95%), uncontrolled keywords (99.4%) and absent multilingual metadata (99.4%). Sensitivity analysis and analytical bounds confirm robustness to weighting choices. Country-level patterns suggest institutional rather than capacity-driven deficiencies. These findings reveal a substantial gap between dataset availability and operational machine-readiness, highlighting targeted metadata remediation and open-source monitoring as priorities for strengthening EU water-data infrastructure and data-intensive governance.

## Impact statement

This research addresses a critical gap between data availability and operational usability in EU water governance. While the European Union has invested heavily in water monitoring infrastructure, with WFD compliance costs alone estimated at €89 billion for the 2022–2027 cycle, water managers face persistent challenges in actually using this data for operational purposes. Our task-specific assessment framework, applied to 2,176 real datasets harvested from data.europa.eu, reveals that mean sufficiency scores at the federated discovery level are critically low across all management tasks (0.16–0.24), with 0% of datasets achieving operational readiness as represented in the federated endpoint. Combined with the finding that 52% of sampled dataset URLs are broken or retired, and that the federated query endpoint itself failed to answer 37 of 43 retrieval queries on re-execution, these results reframe the policy problem: the challenge is less "more data" and more "better metadata and resilient data infrastructure." These findings arrive at a critical juncture: the European Commission has launched a targeted revision of the Water Framework Directive in 2026, which offers an opportunity to consider whether metadata quality requirements should form part of the EU's water governance architecture. The framework and prioritized improvement list we provide can guide data publishers, portal managers and policymakers in maximizing the return on existing EU water data investments through targeted, low-cost metadata interventions.

## Introduction

Water policymaking in Europe rests on an extensive regulatory architecture. The Water Framework Directive (*Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy*, 2000), the Floods Directive (*Directive 2007/60/EC of the European Parliament and of the Council on the assessment and management of flood risks*, 2007b), the Nitrates Directive (*Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources*,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1991) and the INSPIRE Directive (*Directive 2007/2/EC of the European Parliament and of the Council establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*, 2007a) collectively mandate that Member States monitor water bodies, report on their status, coordinate across transboundary basins and make spatially referenced environmental data discoverable and interoperable. Implementation of this architecture is expensive: the cost of WFD compliance alone is estimated at €89 billion for the 2022–2027 cycle (European Commission, 2024b), while annual expenditure on water supply and sanitation across all EU countries averages approximately €100 billion (OECD, 2020). A growing share of this investment flows into digital monitoring infrastructure (sensor networks, earth observation programs and centralized reporting platforms) which produce ever-larger volumes of environmental data. The European Environment Agency's WISE system, the Joint Research Centre's flood and drought observatories, Copernicus climate and land services and the federated data.europa.eu portal together host tens of thousands of water-related datasets (European Environment Agency, 2023; European Commission, 2024a). The question of whether this data infrastructure actually delivers operationally usable information is particularly timely: the European Commission has launched a targeted revision of the WFD in 2026, with a call for evidence aimed at addressing regulatory bottlenecks while maintaining high environmental standards. The evidence base for setting meaningful targets and facilitating compliance depends directly on the quality and accessibility of the underlying data.

In practice, water data are increasingly consumed in automated or semi-automated workflows rather than by manual inspection alone. Early warning systems, regulatory reporting pipelines and cross-border integrations depend on metadata that can be parsed, mapped and acted on by machines. For that reason, metadata quality is no longer only a question of human discoverability; it is also a question of machine readiness. This study is designed around that distinction.

Yet the availability of data does not guarantee its usability. The distinction matters. A dataset that exists on a portal but lacks a machine-readable format, provides no spatial reference, carries an ambiguous license (not in a standardized form) or links to a retired URL cannot be ingested into an early warning model, cannot be reliably cited in a compliance report and cannot be integrated with neighboring countries' data for transboundary planning. In practice, metadata (the structured description of a dataset's content, provenance, format, coverage and access conditions) determines whether data can be found, evaluated and reused (Wilkinson et al., 2016). The FAIR principles (findable, accessible, interoperable, reusable) have provided a conceptual framework for evaluating data stewardship since their introduction, and the Research Data Alliance's FAIR Data Maturity Model has operationalized these principles into measurable indicators (Bahim et al., 2020). These frameworks, however, were designed as general-purpose instruments (Alamanos et al., 2019; Alamanos and Linnane, 2021). They assess whether a dataset meets generic quality thresholds – for instance, whether a persistent identifier is assigned or whether a license is stated – without asking whether the metadata are sufficient for a specific management task (Alamanos et al., 2021; Alamanos and Koundouri, 2024). A dataset that scores well on general FAIRness may still be inoperable for flood early warning if it lacks the temporal resolution and spatial precision that real-time forecasting demands.

Existing assessments of European water data have focused on either domain-specific compilations or repository-level evaluations, but not on metadata fitness for operational use across domains. Pekel et al. (2016) produced a landmark global mapping of surface water extent from Landsat imagery, demonstrating the power of systematically curated geospatial data, but their contribution concerned the data themselves rather than the metadata infrastructure that enables discovery and reuse. Hall et al. (2015) compiled a pan-European flood discharge database, offering an invaluable harmonized time series for hydrological research, but without assessing whether the associated metadata would support automated ingestion into operational systems. The European Commission's annual Open Data Maturity assessments (Publications Office of the European Union, 2024) evaluate national portal capabilities across four dimensions (policy, portal, impact and quality) and have tracked steady improvements since 2015, with the EU-27 average reaching 83% in 2024. These assessments are, however, designed to benchmark countries, not individual datasets and they do not evaluate whether specific metadata fields meet the requirements of distinct water management tasks. Similarly, tools such as FAIR EVA (Aguilar Gómez and Bernal, 2023) and the FAIRsFAIR automated assessment services provide valuable repository-level scoring against FAIR indicators, but their outputs are generic: they do not differentiate between the metadata needs of, say, a flood forecaster and a WFD compliance officer. A substantial literature already establishes that metadata quality on open data portals can be measured systematically and at scale. Kubler et al. (2018) compare metadata quality across open data portals using the Analytic Hierarchy Process to derive dimension weights from structured pairwise comparison, and Nogueras-Iso et al. (2021) provide a comprehensive assessment of metadata quality across open data portals, distinguishing the presence of a metadata element from its usefulness for reuse. We do not claim novelty for automated metadata quality assessment as such, which these studies establish. Our contribution lies elsewhere: in deriving assessment profiles from specific statutory water-management obligations, and in converting the resulting diagnosis into a ranked remediation priority. Existing assessments rarely ask whether metadata are sufficient for machine-based reuse, even though this is becoming central as the WFD is revised and AI-assisted data processing becomes more common in environmental governance.

This gap between general data quality assessment and task-specific operational readiness is consequential. The three core management tasks that EU water legislation mandates place materially different demands on metadata. Early warning and flood forecasting, as operationalized through the European Flood Awareness System (Thielen et al., 2009; Arnal et al., 2018), require near-real-time or high-frequency temporal updates, precise spatial coordinates and machine-readable formats amenable to automated ingestion. Compliance reporting under the WFD demands documented provenance, temporal coverage spanning full 6-year reporting cycles and alignment with standardized vocabularies to enable cross-referencing against regulatory thresholds. Cross-border coordination under the INSPIRE framework and in the context of international river basin management plans requires multilingual metadata, controlled vocabularies and spatial coverage that extends across national boundaries. To our knowledge, no existing study systematically evaluates European water datasets against these differentiated requirements.

Recent advances in computational methods are making it feasible to conduct such assessments at scale for the first time. Neumaier et al. (2016) demonstrated that automated harvesting and quality scoring of metadata across hundreds of open data portals could be performed systematically using SPARQL queries and rule-based metrics, establishing the technical foundation for cross-

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

portal metadata evaluation. Wu et al. (2023) showed that machine learning can automate metadata annotation tasks such as keyword extraction and subject classification, achieving accuracy levels sufficient for large-scale screening when complemented by human validation. More broadly, a recent systematic review of natural language processing applications in environmental science (Raeissi and Knapen, 2025) documents a rapidly expanding set of use cases – from information extraction and classification to knowledge synthesis – that demonstrate the capacity of semi-automated workflows to process the heterogeneous, multilingual and technically complex documentation that characterizes EU environmental data portals. These developments create an opportunity to move beyond manual, one-off audits toward reproducible assessments of metadata fitness that can be periodically updated as data landscapes evolve.

In this paper, we present the EU Water Dataset Observatory, a reproducible framework for task-specific metadata quality assessment across European water data portals. Our approach departs from existing FAIR assessments in three ways. First, we define three management-task profiles: early warning, compliance reporting and cross-border coordination; each with a distinct set of weighted metrics derived from operational guidance documents, including the EFAS technical documentation, WFD reporting guidance (European Environment Agency, 2016) and the FAIR Data Maturity Model (Bahim et al., 2020). Second, we compute per-dataset sufficiency scores that quantify fitness for each task, paired with a five-component impact proxy that enables prioritization of metadata improvements where the return on effort is greatest. Third, we provide a ranked list of actionable fixes, together with reproducible SPARQL harvesting queries and Python scoring tools that allow data publishers and portal managers to replicate and extend the assessment. Relative to prior portal-level metadata assessment (Neumaier et al., 2016; Kubler et al., 2018; Nogueras-Iso et al., 2021), the specific contributions of this paper are therefore twofold: the derivation of task profiles from water-management obligations under the WFD, the Floods Directive and INSPIRE, so that sufficiency is assessed against defined operational uses rather than against generic quality criteria; and the remediation-priority measure, which combines sufficiency with an impact proxy to rank datasets by expected return on metadata repair. Our contribution is thus not to assess metadata completeness as such, but to define and test operational criteria for machine-ready water data in the EU regulatory context.

## Materials and methods

The EU Water Dataset Observatory comprises seven methodological components: a SPARQL-based dataset harvest from data.europa.eu, three task-specific scoring profiles, a per-dataset sufficiency score, a composite impact proxy that feeds into a priority ranking, a validation protocol grounded in manual verification of real portal records, an exploratory country-level equity analysis and a suite of reproducible visualization and analysis tools. Each component is described below; all code and configuration files are available in the project repository.

### *Dataset harvesting*

We harvested water-related dataset metadata from data.europa.eu using a two-phase SPARQL approach designed to work within the execution limits of the federated endpoint at the time of querying. Phase 1 queries apply a substring filter over dataset titles, which the endpoint cannot satisfy from an index; retrieval is therefore sensitive to endpoint load, and we document below the consequences this had for our results. Phase 1 collected distinct dataset URIs matching domain-specific keywords (up to 500 URIs per keyword query), while Phase 2 retrieved full metadata for batches of 50 URIs using VALUES clauses. The exact SPARQL queries and predicate mappings used for each harvesting phase are provided in the project's repository. This approach yielded 2,176 unique water-related datasets with valid metadata records. Every Phase 1 query that completed did so at the 500-URI ceiling rather than exhausting its matches, so the corpus is a truncated sample of the matching records rather than a census, and 2,176 should be read as a lower bound. Because the queries carry no `ORDER BY` clause, the records returned under that ceiling are an arbitrary rather than a systematic subset. All prevalence figures reported below are accordingly properties of the retrieved corpus, and we do not present them as estimates of catalog-wide rates. The harvesting strategy prioritizes reproducibility and scalability over completeness. While alternative query formulations or multilingual expansion may retrieve additional datasets, the observed patterns – particularly the absence of key metadata fields – are unlikely to be sensitive to marginal increases in dataset count. The framework is intentionally designed to assess machine-operational readiness, meaning whether a dataset can be discovered, interpreted and reused by automated workflows with minimal manual intervention.

The harvest queried all seven water-management domains defined in our framework through domain-specific keyword sets: floods, water quality, groundwater, WFD metrics, utilities, agricultural runoff and cross-cutting. These domains correspond to the principal data categories through which EU water directives are operationalized. However, only two domains contributed records through the federated endpoint: floods (1,477 datasets, 67.9%) and groundwater (699 datasets, 32.1%). The remaining five domains (water quality, WFD metrics, utilities, agricultural runoff and cross-cutting) contributed no records. Post hoc audit establishes that this was a consequence of query failure rather than of empty result sets, and we set out the evidence in the following paragraphs, since the distinction materially changes what the result licenses us to conclude.

The instrumentation used in the original harvest did not record per-domain query outcomes: a failed query and a query returning zero rows both produced an empty URI list and were logged identically. We therefore audited the outcome after the fact, by two independent routes.

First, we reconstructed which keywords demonstrably contributed to the deposited corpus. Because Phase 1 matched a substring against the dataset title, any record matching exactly one keyword can only have entered through that keyword's query. Six queries are established as having completed on this criterion: *flood*, *hochwasser* and *inondation* for floods, and *groundwater*, *grundwasser* and *aquifer* for groundwater. None of the 30 keywords belonging to the five non-returning domains produced a uniquely attributable record.

Second, the assignment of records to domains is itself diagnostic. Domains were iterated in a fixed order, and each URI was assigned to the first domain that retrieved it. Water quality precedes groundwater in that order, yet all 38 records in the corpus whose titles contain *water quality* (24), *water monitoring* (10) or *chemical status* (4) are labeled groundwater. Had the Water quality queries been completed, they would have claimed these records first. The records were therefore present in the catalog and retrievable, and the queries intended to find them did not return them.

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

We re-executed all 43 Phase 1 queries against the same endpoint on 22 August 2026 with per-query outcome logging. Six completed, each returning a maximum of 500 URIs; 18 returned HTTP 504 Gateway Timeout after 55–114 s; and 19 returned no response within a 120 s client timeout. *No query in the re-run returned an empty result set.* All 30 keywords of the five non-returning domains failed. Outcomes were not stable between runs: *inundación* contributed nothing in March but completed in August, while *aquifer* completed in March but timed out in August, indicating that retrieval success depends on endpoint load at the moment of querying rather than on catalog content.

We therefore report the following, and no more. The five domains did not return records, and the cause was query failure at the federated endpoint, not an observation that matching records are absent. Our earlier formulation, that these domains "returned zero matches," was not supported by the evidence available and is withdrawn. Three further mechanisms may compound the effect and cannot be separated from it by our design: keyword formulation, since several of our terms are multi-word phrases (*nitrate water*, *remote sensing water*) that are unlikely to occur as contiguous substrings in real titles; language coverage, since the two contributing domains were the only ones whose keyword sets included non-English terms; and title translation, since data.europa.eu serves machine-translated title variants and 32 records in our corpus match none of our keywords under the title finally stored, which is consistent with Phase 1 and Phase 2 having seen different language variants of the same record. Establishing the true extent of each domain in the catalog would require an ingestion-tolerant harvesting strategy with paginated, index-friendly queries, and we identify this in the "Conclusions" section as a specific priority.

An important methodological consideration is that our SPARQL harvest captures metadata as exposed through the federated data.europa.eu endpoint, not as it may exist on the original source portals. The data.europa.eu portal aggregates metadata from national catalogs using automated harvesting and mapping to DCAT-AP, a process during which metadata fields — particularly format declarations, license information and controlled vocabulary keywords – may be lost, truncated or stored under non-standard predicates (Neumaier et al., 2016). Our assessment therefore measures metadata quality *as experienced by a cross-border user* querying the European federated discovery layer, which is precisely the use case that cross-border coordination, EU-level compliance oversight and pan-European early warning systems depend upon. Where national portals maintain richer metadata that does not propagate to the federated level, this represents a failure of the federation infrastructure itself: the investment in metadata quality at the national level is effectively wasted for European-level applications. This framing (i.e., assessing the end-user discovery experience rather than the full metadata landscape) is both a deliberate methodological choice and a substantive finding about the metadata supply chain. Each record contains metadata dimensions aligned with the DCAT-AP standard (European Commission, 2020), including title, publisher, issued and modified dates, spatial coverage, license, language and keyword fields.

### Task-specific scoring profiles

We defined three management-task profiles reflecting distinct operational requirements mandated by EU water legislation. The concept of task-specific "fitness for purpose" assessment, in which data quality is evaluated against the specific requirements of an intended use rather than against a universal standard, follows the principle articulated in the data quality literature (Wang and Strong, 1996) and operationalized more recently for real-world data scoring (Kahn et al., 2016).

**Early warning** represents the operational requirements of near-real-time flood and pollution alert systems, as exemplified by the European Flood Awareness System (Thielen et al., 2009). The defining metadata needs are temporal recency (whether the dataset has been updated recently enough to inform real-time decisions), update frequency (whether refresh cycles are compatible with operational forecasting) and spatial precision (whether coordinates are sufficiently precise for local-scale alerts). Machine-readable formats and open licenses are also weighted, but less heavily, since automated ingestion pipelines can accommodate some format variation.

**Compliance reporting** represents the requirements of the WFD's 6-year reporting cycle, in which Member States must document the ecological and chemical status of water bodies with traceable provenance. Here, temporal coverage (whether the dataset spans an entire reporting period) and description quality (whether provenance and methodology are documented) receive the highest weights, followed by vocabulary standardization (whether keywords align with EuroVoc or other controlled vocabularies used in cross-referencing against regulatory thresholds). Temporal recency receives a low weight because historical data are acceptable, indeed essential, for trend analysis.

**Cross-border coordination** represents the requirements of transboundary river basin management and the interoperability mandates of the INSPIRE Directive. Vocabulary standardization receives the highest weight, followed by multilingual metadata availability and spatial coverage extending beyond national boundaries. These priorities reflect the practical barriers encountered when integrating datasets from different Member States into shared basin-level assessments.

The framework evaluates nine metadata dimensions in total. Each task profile activates seven of these nine, assigning a weight vector $\mathbf{w}^t = \left[w_1^t, w_2^t, \ldots, w_9^t\right]$ with task-specific weights summing to 1.0; the remaining two dimensions receive a weight of 0.00 for that task. This design reflects the fact that different tasks impose different metadata requirements: for instance, update frequency is critical for early warning but irrelevant for compliance reporting, while multilingual metadata matters for cross-border coordination but not for early warning. Weights were derived by structured reading of operational guidance documents: the EFAS technical documentation for early warning, the EEA's WFD Reporting Guidance 2016 for compliance reporting and the INSPIRE Metadata Regulation (2008/1205/EC) together with the FAIR Data Maturity Model (Bahim et al., 2020) for cross-border coordination. The procedure had two stages. First, for each task we extracted every statement in the governing document that specified a metadata property as necessary for the activity concerned, and used these statements to establish an ordinal ranking of the nine dimensions by importance to that task. Where documents imposed conflicting or competing requirements, we resolved the conflict by giving precedence to the requirement expressed as mandatory in the legally binding instrument over requirements expressed as guidance or good practice. Second, we assigned cardinal weights consistent with that ordinal ranking, constrained to sum to unity and to a resolution of 0.05.

We wish to be explicit that this procedure does not eliminate subjective judgment, and we no longer claim that it does. Extracting requirements from policy documents and converting an ordinal ranking into cardinal weights are both interpretive acts, and the specific numerical values in Table 1 represent the authors' considered

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

**Table 1.** Task-specific dimension weights derived from operational requirements literature

| Dimension | Early warning | Compliance | Cross-border |
|---|---|---|---|
| Temporal recency | 0.25 | 0.05 | 0.05 |
| Update frequency | 0.20 | 0.00 | 0.00 |
| Temporal coverage | 0.00 | 0.25 | 0.00 |
| Spatial precision | 0.20 | 0.15 | 0.20 |
| Format readability | 0.15 | 0.10 | 0.15 |
| License openness | 0.10 | 0.10 | 0.05 |
| Description quality | 0.05 | 0.20 | 0.10 |
| Vocabulary standard | 0.05 | 0.15 | 0.25 |
| Multilingual | 0.00 | 0.00 | 0.20 |
| *Sum* | *1.00* | *1.00* | *1.00* |

*Notes*: Each task activates seven of nine dimensions; weights sum to 1.00 per task.

judgment rather than a validated elicitation. Formal elicitation methods, in particular the Analytic Hierarchy Process applied to metadata quality by Kubler et al. (2018), or a Delphi procedure with operational practitioners, would place the weights on firmer empirical footing, and we identify this in the "Conclusions" section as the principal methodological priority for further development of the framework. The sensitivity analysis shows that the central result of the paper does not depend on the weight specification. The full weight sets are reported in Table 1 and sensitivity tested in the "Sensitivity Analysis" sub-section.

### Sufficiency score calculation

Each metadata dimension is scored on a continuous [0, 1] scale using rule-based functions that map raw metadata values to normalized quality indicators. The nine dimensions are as follows: (i) temporal recency, scored as a decay function of time since the last recorded modification date; (ii) update frequency, scored by mapping stated refresh intervals to a tiered scale (real-time or daily = 1.0, monthly = 0.7, annual = 0.4, unspecified = 0.1); (iii) temporal coverage, scored by whether the dataset spans the relevant reporting period; (iv) spatial precision, scored by the presence and resolution of geographic coordinates or bounding boxes; (v) format readability, scored by classifying file formats into machine-readable (API, CSV, JSON, NetCDF = 0.85–1.0), structured geospatial (Shapefile, GeoPackage = 0.7–0.8) and non-machine-readable (PDF, HTML report = 0.1–0.3); we note below that this ordering is not appropriate for all tasks; (vi) license openness, scored from 1.0 for clearly stated open licenses (CC BY, CC0) through 0.5 for restricted-use licenses to 0.0 for absent license statements; (vii) description quality, scored by description length, presence of methodology statements and keyword richness; (viii) vocabulary standardization, scored by whether keywords map to EuroVoc or other controlled vocabularies; and (ix) multilingual availability, scored by whether metadata are provided in more than one EU language. These scoring functions follow the weighted-metric approach to metadata quality assessment established by Ochoa and Duval (2009) and applied at scale to open data portals by Neumaier et al. (2016) and Nogueras-Iso et al. (2021). Here "readiness" refers to machine-ready metadata for operational workflows, not to the intrinsic scientific validity of the underlying measurements.

Two of these scoring rules require qualification. The license dimension as specified conflates two distinct properties, legal clarity and practical openness, along a single axis. A clearly worded but highly restrictive license, which explicitly forbids redistribution, would score 0.5 under this rule and therefore rank above a dataset carrying no license statement at all, even though it may permit less reuse in practice. Separating clarity and permissiveness into two dimensions would be the more defensible design, and we identify it as a needed refinement. We note that the rule has almost no bearing on the results reported here: license metadata were absent for 2,175 of the 2,176 harvested records (99.95%), and for every accessible source page we inspected. A single record carries a CC BY 4.0 declaration and scores 1.0 on this dimension; every other record scores 0.0, so the treatment of intermediate cases is never invoked.

The format ordering is likewise task-dependent rather than general. GeoPackage is a structured spatial database format, and Shapefile is machine-readable; placing both below CSV is inconsistent with geospatial practice, since CSV frequently lacks a schema, a coordinate reference system and topology, all of which are prerequisites for spatial analysis. A ranking appropriate to tabular compliance reporting is therefore not appropriate to spatially explicit early warning, and format suitability should properly be scored separately for each task. As with license openness, this does not affect the present results, because no harvested record carried a mapped format declaration and all records score 0.0 on this dimension irrespective of the ranking applied. We treat task-specific format scoring as an extension in the "Conclusions" section.

The sufficiency score for dataset $d$ under task $t$ is then computed as the weighted average:

$$S(d,t) = \sum_{i=1}^{9} w_i^t \cdot m_i(d), \tag{1}$$

where $m_i(d)$ is the metric score for dimension $i$ and $w_i^t$ is the corresponding task-specific weight (0.00 for dimensions not active in that task). The score ranges from 0 (no metadata present or completely unfit) to 1 (all metadata dimensions fully satisfy the task requirements). We classify the datasets into three readiness categories: ready ($S \geq 0.7$), partial ($0.5 \leq S < 0.7$) and insufficient ($S < 0.5$). These thresholds were chosen to align with commonly used quality grades in open data assessments (Publications Office of the European Union, 2024), where 0.7 conventionally marks the boundary above which a resource is treated as fit for routine use without remediation. We make no claim that 0.7 is uniquely correct, and we report in Section 4.4 how the results change when the threshold is varied.

The additive form of Equation 1 warrants comment, since it embeds a substantive assumption. A weighted sum permits compensation between dimensions: a high score on temporal recency can offset a low score on license openness. The fit-for-purpose literature, by contrast, generally treats suitability as context-dependent and frequently non-compensatory, with some requirements functioning as gating conditions rather than as contributions to an aggregate (Wang and Strong, 1996; Kahn et al., 2016). We adopt the additive form because it yields a continuous ranking suitable for prioritization, which a purely gated model does not, and because it allows partial progress to be represented. We do not intend it as a claim that dimensions are genuinely substitutable: a dataset without a license declaration cannot lawfully be reused whatever its temporal resolution, and a dataset without a machine-readable format cannot be ingested automatically whatever its spatial precision. In Section 4.4, we show that in this corpus the

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

two framings coincide, because the dimensions that would act as gates are absent for essentially all records. The development of an explicitly non-compensatory variant, with gating conditions on license and format, is discussed in the "Conclusions" section.

To test the robustness of the framework to the weight specification, we performed a sensitivity analysis in which the highest-weighted metric for each task was perturbed by ±25%, with the remaining weights renormalized to maintain a unit sum. This approach isolates the influence of the dominant metric (i.e., the parameter most likely to be contested in a review of the framework) while preserving the relative ordering of all others. We complement this with two further checks reported in Section 4.4: variation of the readiness threshold, and an analytical bound on the maximum attainable score given the observed distribution of metadata presence. We note explicitly that we did not perturb multiple weights jointly, test alternative functional forms for the individual scoring rules or compare alternative treatments of missing values; these are recorded as limitations in the "Conclusions" section.

### Impact proxy and priority score

Prioritizing metadata improvements requires an estimate of each dataset's operational importance, yet download counts and usage statistics are unavailable for most datasets on EU portals. We therefore constructed a five-component composite impact proxy, scored on [0, 1], drawing on observable metadata attributes as indicators of likely operational relevance:

- **Publisher type** (weight: 0.30) – scored by the institutional credibility and operational role of the publisher (national hydrological service = 1.0, EU agency = 0.9, ministry = 0.7, regional authority = 0.5, other = 0.3);
- **Operational keywords** (weight: 0.25) – scored by the presence of terms such as "real-time," "operational," "forecast" or "monitoring" in the title or description (present = 1.0, absent = 0.2);
- **Format accessibility** (weight: 0.20) – scored using the same format classification as the sufficiency metric;
- **Temporal extent** (weight: 0.15) – scored by the length of the time series as a proxy for data richness ( > 20 years = 1.0, 10–20 years = 0.8, 5–10 years = 0.6, < 5 years = 0.4);
- **Spatial scope** (weight: 0.10) – scored by geographic coverage (EU-wide = 1.0, multi-country = 0.8, national = 0.6, regional = 0.4).

The priority score combines impact with the unmet potential for improvement:

$$P(d,t) = I(d) \times (1 - S(d,t)), \tag{2}$$

where $I(d)$ is the composite impact score. A high priority score identifies a dataset that is both operationally important and currently underserved by its metadata – the combination that yields the greatest return on a targeted metadata improvement effort. This formulation follows the general principle of opportunity-cost prioritization used in infrastructure investment planning, where resources are directed to the highest-value gaps (OECD, 2020).

### Validation

To ground the framework in empirical evidence, we conducted a manual verification of 25 datasets drawn from the harvested corpus between March and April 2026, following the harvest of 18 March. The sample was stratified into five domains (floods, water quality, WFD metrics, groundwater and utilities), with five datasets per domain, drawn from four major EU publishers (EEA, JRC, Eurostat and Copernicus). Agricultural Runoff and Cross-cutting domains were excluded from the validation sample because their datasets overlap substantially with those in the core five domains.

Within each stratum, selection was purposive rather than random. We deliberately chose prominent, high-visibility datasets that a water manager or compliance officer would plausibly reach for first, such as the JRC river flood hazard maps, the EEA Waterbase collections, the WISE WFD database and the Eurostat water statistics tables. This design has a clear implication for how the results should be read. It is a best-case selection, and the direction of the resulting bias is conservative with respect to our conclusions: if the most visible and institutionally supported datasets from the EEA, JRC, Eurostat and Copernicus exhibit broken access links and absent license declarations, less prominent records are unlikely to perform better. It does not, however, support extrapolation of the observed rates to the corpus as a population estimate, and we do not treat them as such. A probability sample would be required for that purpose and is identified as future work in the "Conclusions" section. The dataset names given in Appendix are the descriptive labels used internally by the author team; the corresponding record identifiers and access URLs, as they appear in the harvested corpus, are given in the project repository. We also note that three of the 25 selected records resolve to the same EEA landing page, so the sample comprises 23 unique URLs and, of these, 11 unique accessible pages; we report results on both denominators below. The full dataset-level record, including URLs, publisher, per-dimension coding and free-text notes, is provided in the Appendix.

The sample size of $n = 25$ was designed as a targeted diagnostic check rather than a statistically powered representativeness study. The stratified design ensures coverage of the major EU publishers and water domains, and the sample is sufficient to reveal whether systematic infrastructure problems are present in the ecosystem at all. We note that prior automated metadata quality studies have used comparable or smaller manual validation samples (Neumaier et al., 2016), and we identify expansion to a larger probability sample as a priority for future work. Throughout, we describe the validation as a diagnostic exercise, and we do not present the observed URL breakage rate as an estimate of system-wide prevalence.

For each dataset, we manually visited the source portal and evaluated five verification dimensions: (i) description accuracy: whether the portal description matched the cataloged metadata; (ii) temporal accuracy: whether the stated temporal coverage was correct; (iii) format accuracy: whether the claimed formats were actually available for download; (iv) license presence: whether a license was explicitly stated on the portal page; and (v) URL accessibility: whether the cataloged access URL resolved to a working resource. Each dimension was scored as Yes (1.0), Partial (0.5) or No (0.0), and the overall agreement rate was computed as:

$$A = \frac{\sum(\text{Yes} + 0.5 \times \text{Partial})}{\text{Total assessments}}. \tag{3}$$

Where a URL did not resolve, the description, temporal, format and license dimensions could not be observed at all. These are recorded as *Not assessable* rather than as failures, and are excluded from the agreement calculation while being counted in full toward the assessment of infrastructure stability. This distinction matters: a page that cannot be opened provides no evidence about whether its cataloged metadata are accurate, and coding it as a disagreement would conflate link persistence with extraction fidelity. Agreement is therefore computed over the four content dimensions for the

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

13 datasets with accessible URLs, giving a denominator of $4 \times 13 = 52$ assessments; the numerator and denominator are stated explicitly alongside the result in the sub-section "Country-Level Variation in Metadata Quality".

We also report the agreement statistic in decomposed form. The description, temporal and format checks measure *extraction fidelity*, that is, whether the cataloged metadata correspond to what the source portal actually shows. The license check measures *presence*, that is, whether any license is stated at all. Averaging a presence measure that is identically zero into an accuracy statistic conflates two different properties and depresses the resulting figure for reasons unrelated to extraction quality. We therefore report both the combined four-dimension figure, for continuity with the originally submitted version, and the three-dimension extraction-fidelity figure alongside license presence reported separately. Scripts to reproduce the validation checklist are included in the project repository, and the dataset-level coding is reproduced in Appendix.

### Country-level exploratory analysis

To examine whether metadata quality varies systematically across Member States and whether socioeconomic factors explain the variation, we conducted an exploratory country-level analysis. Country of origin was extracted from the harvested metadata using a five-strategy cascade: (i) publisher URI top-level domain, (ii) spatial field parsing, (iii) dataset URI domain patterns, (iv) country-name matching in titles and keywords across five languages and (v) language inference for monolingual records. This cascade identified a country for 1,430 of 2,176 datasets (65.7%) across 20 EU Member States; the remaining 746 records could not be assigned due to blank-node publishers, absent spatial metadata and English-only descriptions with no geographic identifiers.

To ensure reliable country-level means, we restricted the analysis to countries contributing $n \geq 10$ datasets, yielding nine countries covering 1,400 datasets (97.9% of identified records). We compiled five socioeconomic indicators from Eurostat and European Commission sources: GDP per capita in purchasing power standard (2022), the Digital Economy and Society Index (DESI, 2023), Water Exploitation Index (latest available), Open Data Maturity score (2023) and e-government usage as a percentage of citizens (2023). We computed Spearman rank correlations ($\rho$) between each indicator and the mean composite sufficiency score for each country. Spearman was chosen over Pearson because with $n = 9$ countries, distributional assumptions cannot be verified. Multivariate regression was not performed because preliminary variance inflation factor (VIF) analysis revealed severe collinearity among the digital governance indicators (VIF $> 10$ for GDP, DESI and e-government usage), making coefficient estimation unreliable.

### Visualization and analysis tools

The framework generates four principal outputs designed for both research and policy audiences: (i) a country-by-task sufficiency heatmap showing mean readiness scores by Member State and management task; (ii) a failure-mode bar chart identifying the most common metadata gaps across the corpus (e.g., percentage of datasets with no spatial reference, absent license in the federated catalog); (iii) a priority-fix table listing the top-ranked datasets by priority score together with the specific metadata improvement that would yield the largest sufficiency gain; and (iv) a sensitivity analysis chart showing the stability of readiness category assignments under $\pm 25\%$ weight perturbation. All visualizations are implemented as self-contained interactive HTML files using the Plotly library and are distributed with the project repository alongside the SPARQL harvesting scripts and Python scoring code.

**Table 2.** Overall sufficiency scores by management task ($N$ = 2,176)

| Task | Mean | Ready ($\geq 0.7$) (%) | Partial (0.5–0.7) (%) | Insufficient ($< 0.5$) (%) |
|---|---|---|---|---|
| Early warning | 0.202 | 0.0 | 3.6 | 96.4 |
| Compliance reporting | 0.155 | 0.0 | 0.0 | 100.0 |
| Cross-border | 0.241 | 0.0 | 0.5 | 99.5 |

*Note*: Scores reflect metadata as exposed through the federated data.europa.eu endpoint at the time of harvest (18 March 2026).

## Results

### Overall metadata sufficiency

Mean sufficiency scores were critically low across all three management tasks, ranging from 0.16 to 0.24 on a 0–1 scale (Table 2). Early warning showed a mean sufficiency of 0.202, reflecting the stringent requirements for temporal recency and update frequency that virtually no datasets currently meet. Compliance reporting showed the lowest mean (0.155), driven by the near-complete absence of temporal coverage metadata and description quality. Cross-border coordination achieved the highest mean (0.241), though still far below operational thresholds.

The most striking finding is that 0% of datasets achieved "Ready" status ($S \geq 0.70$) for any management task, as represented in the federated endpoint at the time of harvest. Even "Partial" readiness ($0.5 \leq S < 0.7$) was rare: only 3.6% for early warning, 0.5% for cross-border coordination and 0.0% for compliance reporting. The overwhelming majority of datasets (over 96% for all tasks) fell into the "Insufficient" category, indicating fundamental metadata gaps that preclude operational use without substantial manual intervention.

### Variation between retrieval groups

Sufficiency scores varied between the two groups of records retrieved through the federated endpoint (Table 3). Floods datasets showed marginally higher early warning sufficiency (0.21) compared to groundwater (0.19), consistent with the operational requirements of EFAS-linked data products. However, both groups showed uniformly low scores across all tasks, with neither achieving mean sufficiency above 0.25 for any task. This comparison should be read with care. The labels record which keyword set first retrieved a record, not a thematic classification, and as the "Overall metadata sufficiency" section sets out, which keyword sets succeeded was determined in part by endpoint behavior. Thirty-eight records whose titles denote water quality are labeled groundwater

**Table 3.** Mean sufficiency scores by retrieval group and task

| Retrieval group | Early warning | Compliance | Cross-border |
|---|---|---|---|
| Floods ($n$ = 1,477) | 0.21 | 0.16 | 0.25 |
| Groundwater ($n$ = 699) | 0.19 | 0.15 | 0.23 |

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

for this reason. The two groups are therefore best understood as two retrieval populations rather than as two water-management domains, and the difference between them is in any case small relative to the dispersion within each.

### Common metadata failures

The most frequent metadata gaps identified across the corpus are summarized in Table 4. The results reveal systematic and near-universal metadata deficiencies at the federated catalog level. Most critically, 100% of harvested records lacked mapped machine-readable format declarations, and 99.95% (2,175 of 2,176) lacked mapped open license metadata – fundamental barriers to any form of automated reuse. These rates refer to the license and format fields as exposed through the federated SPARQL endpoint; in some cases, source portals may contain richer metadata that do not propagate through the federation pipeline (see "Discussion" section). The absence of license metadata in the federated catalog does not necessarily imply the absence of licensing terms at the source portal, but rather the absence of machine-readable, standardized license information required for automated reuse decisions. The absence of EuroVoc or other controlled vocabulary keywords affected 99.4% of records, severely limiting cross-border discoverability and semantic interoperability. Similarly, 99.4% of datasets had monolingual metadata only, compounding barriers to transboundary use.

Spatial metadata gaps (missing coordinates or bounding boxes) affected 22.7% of records, namely lower than other failure types but still a critical barrier for early warning applications that require precise geographic targeting. Temporal coverage was absent for 46.9% of datasets, limiting their utility for compliance reporting that requires documentation of measurement periods.

### Sensitivity analysis

Table 5 presents the sensitivity analysis results under ±25% perturbation of the primary (highest) weight for each task. The "Ready" percentage remained at exactly 0.0% across all perturbation scenarios for all tasks. We state the implication narrowly: the results are robust to the weight perturbations tested here. This test varies a single dominant weight at a time and does not vary the readiness threshold, perturb several weights jointly or test alternative scoring rules.

Two further checks bear more directly on whether the finding depends on how the framework was specified.

**Threshold variation**. Relaxing the readiness threshold from $S \geq 0.70$ to $S \geq 0.50$ raises the proportion of datasets classified as Ready only to 3.6% for early warning, 0.0% for compliance reporting and 0.5% for cross-border coordination. The conclusion is therefore not an artifact of a strict threshold: even a substantially more permissive criterion leaves the overwhelming majority of records below it.

**Analytical bound**. A stronger statement is available without re-estimation. Because machine-readable format declarations are absent for 100% of harvested records and license metadata for 99.95%, and controlled-vocabulary keywords and multilingual metadata for 99.4%, the weight assigned to these dimensions cannot contribute to the score of any record lacking them. The maximum attainable sufficiency score is therefore bounded above by unity minus the weight placed on the universally absent dimensions (Table 6). For compliance reporting and cross-border coordination, the ceiling lies below the readiness threshold, so no dataset can be classified as Ready under this weighting regardless of its performance on every other dimension. For early warning, the ceiling coincides exactly with the threshold, meaning a dataset would have to score perfectly on temporal recency, update frequency, spatial precision and description quality simultaneously to reach it.

This bound reframes what the sensitivity analysis demonstrates. The invariance of the readiness proportion under weight perturbation is not evidence that the weights were well chosen; it is a consequence of the fact that, given the observed distribution of metadata presence, no admissible weighting could have produced a different classification. The zero-readiness result is an arithmetic consequence of universal gaps in two to four metadata fields rather than a product of weight calibration. It also has a bearing on the compensatory form of Equation 1 discussed in the "Common metadata failures" section: because the dimensions that would function as gating conditions in a non-compensatory model are precisely those absent for essentially all records, an additive and a gated specification return the same classification in this corpus.

**Table 4.** Prevalence of common metadata failures ($N$ = 2,176)

| Failure type | Prevalence (%) |
|---|---|
| No machine-readable format declared | 100.0 |
| Missing or unclear license | 99.95 |
| No EuroVoc/controlled vocabulary keywords | 99.4 |
| Monolingual metadata only | 99.4 |
| No temporal coverage metadata | 46.9 |
| No spatial coordinates/bounding box | 22.7 |

*Note*: Prevalence reflects metadata as mapped to the federated data.europa.eu endpoint.

**Table 5.** Sensitivity analysis: Readiness percentage under ±25% weight perturbation

| Task | Primary weight | −25% (%) | Baseline (%) | +25% (%) |
|---|---|---|---|---|
| Early warning | temporal_recency (0.25) | 0.0 | 0.0 | 0.0 |
| Compliance | temporal_coverage (0.25) | 0.0 | 0.0 | 0.0 |
| Cross-border | vocabulary_standard (0.25) | 0.0 | 0.0 | 0.0 |

*Notes*: Range (Δ): 0.0 pp for all tasks. The "Partial" category showed similarly negligible variation (Early Warning: 3.4–3.8%; Cross-Border: 0.4–0.6%; Compliance: 0.0% throughout).

**Table 6.** Maximum attainable sufficiency score given the observed structural gaps

| Task | Weight on absent dimensions | Ceiling | Ready threshold |
|---|---|---|---|
| Early warning | 0.30 | 0.70 | 0.70 |
| Compliance reporting | 0.35 | 0.65 | 0.70 |
| Cross-border | 0.65 | 0.35 | 0.70 |

Absent dimensions are format readability (0% present) and license openness (0.05% present, a single record), together with vocabulary standardization and multilingual availability (0.6% present). Ceilings are computed for the 99.4% of records lacking all four; the one record carrying a license declaration has a correspondingly higher ceiling on the compliance task and was nonetheless not classified Ready. For the 0.6% of records carrying vocabulary and multilingual metadata, the cross-border ceiling rises to 0.80, which is consistent with the 0.5% of cross-border records observed in the Partial category (Table 2).

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

### Pilot validation against real datasets

Manual validation of 25 real datasets from EU portals (EEA, JRC, Eurostat, Copernicus), conducted between March and April 2026, revealed critical findings about both extraction accuracy and infrastructure stability (Table 7).

A notable finding is that 12 of the 25 sampled dataset URLs were broken or redirected to retired pages at the time of verification. Because three of the sampled records resolve to a single shared landing page, this corresponds to 12 of 23 unique URLs, or 52%; on the nominal count of 25 records, the figure is 48%. We report the unique-URL denominator as the more appropriate one and note both for transparency. Breakage was concentrated in the floods domain (4 of 5 non-functional); each of the four remaining domains showed two of five non-functional.

For the 13 datasets with accessible URLs, agreement between cataloged metadata and manual portal verification across the four content dimensions was $32/52 = 61.5\%$. As set out in the "Pilot validation against real datasets" section, this combined figure mixes extraction fidelity with license presence. Decomposed, agreement on the three extraction-fidelity dimensions (description, temporal coverage, format) is $32/39 = 82.1\%$, while license presence is 0 of 13. The extraction procedure therefore reproduces portal content reasonably well where the portal can be reached, and the lower combined figure is driven by the universal absence of license declarations rather than by extraction error. Restricting to the 11 unique accessible pages changes these figures only marginally, to $27/44 = 61.4\%$ and $27/33 = 81.8\%$, respectively.

None of the 13 accessible source pages carried an explicit license declaration, despite all being freely accessible resources. We restrict this statement to the 13 pages that could actually be opened; the remaining 12 provide no evidence either way, and our earlier formulation over the full sample of 25 overstated what was observable. This systematic absence of license information, even for datasets published by EU agencies with explicit open-data mandates, represents a substantial barrier to automated reuse determination and legal clarity for downstream applications. Because the absence was confirmed by direct inspection of source portal pages and not only in the federated catalog, license metadata are the one dimension for which we can say that the gap does not originate solely in the federation pipeline.

### Country-level variation in metadata quality

Of 2,176 harvested datasets, country of origin could be identified for 1,430 (65.7%) across 20 EU Member States. Restricting to countries with $n \geq 10$ datasets yielded nine countries covering 1,400 records (97.9% of identified datasets). The distribution is highly skewed: France contributed 638 datasets, Germany 275, the Netherlands 159, Ireland 77, Italy 75, Belgium 69, Spain 53, Austria 35 and Denmark 19.

Mean composite sufficiency across the nine countries was 0.214 (SD = 0.039), with all countries falling within the "Insufficient" category. The range was narrow: from 0.138 (Netherlands) to 0.273 (Ireland). Spearman rank correlations between composite sufficiency and five socioeconomic indicators revealed no statistically significant associations at $p < 0.05$ (Table 8).

The absence of significant associations is itself informative. Neither national wealth (GDP), digital maturity (DESI), open data portal development (Open Data Maturity), e-government adoption nor water stress (Water Exploitation Index) predicts metadata quality on the federated portal. The Netherlands, one of Europe's most digitally advanced countries (DESI = 67.4, Open Data Maturity = 93%), achieved the lowest composite sufficiency in the sample (0.138), while Ireland (DESI = 56.4) achieved the highest (0.273). This pattern is consistent with a volume–quality trade-off: countries that publish large numbers of datasets through automated pipelines (e.g., WMS/WFS service layers) may achieve lower per-dataset metadata quality than countries that publish fewer but better-documented records. This exploratory observation warrants further investigation with larger samples and source-portal-level data, but it suggests that the metadata quality challenge is institutional and governance-driven rather than a function of economic or digital capacity.

### Priority fixes and high-impact improvements

Table 9 presents examples of high-priority metadata improvements identified by the framework, ranked by Impact × (1 − Sufficiency). Given that sufficiency scores are uniformly low, priority rankings are primarily driven by the impact proxy; datasets from authoritative publishers with operational keywords receive highest priority for metadata remediation.

The two highest-impact fixes across the corpus are (i) adding explicit open license declarations and (ii) declaring machine-readable format availability. These two metadata fields alone, if populated correctly across the corpus, would substantially increase sufficiency scores and enable basic automated processing capabilities, though full operational readiness would additionally require attention to data-level quality (sampling design, calibration, provenance).

**Table 7.** Pilot validation results ($N$ = 25 real datasets)

| Category | Yes | Partial | No | Not assessable |
|---|---|---|---|---|
| URL works | 13 | 0 | 12 | — |
| Description accurate | 12 | 0 | 1 | 12 |
| Temporal accurate | 10 | 2 | 1 | 12 |
| Format accurate | 5 | 8 | 0 | 12 |
| License stated | 0 | 0 | 13 | 12 |

Datasets whose URL did not resolve are recorded as Not assessable on the four content dimensions, since no observation of those fields was possible. Agreement is computed over the 13 accessible datasets: $(12+0)+(10+0.5\times2)+(5+0.5\times8)+0=32$, over a denominator of $4\times13=52$, giving 61.5%. Dataset-level coding for all 25 records is provided in the Appendix.

**Table 8.** Spearman rank correlations between country-level composite sufficiency and socioeconomic indicators (n = 9 countries with ≥10 datasets)

| Indicator | $\rho$ | $p$ value |
|---|---|---|
| GDP per capita (PPS) | +0.259 | 0.500 |
| DESI score | −0.050 | 0.898 |
| Water exploitation index | −0.067 | 0.865 |
| Open data maturity | −0.319 | 0.402 |
| E-government usage (%) | +0.042 | 0.915 |

*Notes*: No correlation is significant at $p < 0.05$. Countries with $n < 10$ datasets were excluded to ensure reliable means.

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

**Table 9.** Top 10 high-priority metadata improvements ranked by Impact × (1 − Sufficiency)

| Rank | Domain | Suggested fix | Est. gain |
|---|---|---|---|
| 1 | Floods | Add explicit open license (CC-BY) | +0.10 |
| 2 | Floods | Declare machine-readable format | +0.15 |
| 3 | Groundwater | Add explicit open license (CC-BY) | +0.10 |
| 4 | Floods | Add bounding box coordinates | +0.08 |
| 5 | Groundwater | Declare machine-readable format | +0.15 |
| 6 | Floods | Map keywords to EuroVoc | +0.05 |
| 7 | Groundwater | Add temporal coverage metadata | +0.08 |
| 8 | Floods | Add multilingual title/description | +0.06 |
| 9 | Groundwater | Map keywords to EuroVoc | +0.05 |
| 10 | Floods | Add temporal coverage metadata | +0.08 |

*Notes:* Est. Gain = estimated sufficiency score improvement from a single fix. These fixes enable automated metadata checks and reduce manual preprocessing; actual operational readiness may additionally require data-level quality checks. The complete ranked list of 60 priority improvements is available in the project repository.

## Discussion

The EU Water Dataset Observatory, applied to 2,176 datasets harvested from the EU's primary federated data portal, exposes a stark mismatch between the existence of water data and its operational usability. The finding that 0% of harvested datasets achieve operational readiness for any management task (early warning, compliance reporting or cross-border coordination) as represented in the federated endpoint at the time of harvest is the central result of this study. It is not a marginal shortfall but indicates systemic deficiencies in the metadata infrastructure at the federated level that substantially impair operational reuse.

The study should therefore be read as an assessment of whether the EU water data ecosystem is ready for the next generation of automated, AI-assisted and cross-border workflows, rather than as a claim that the underlying datasets have no value to expert human users. Our assessment does not imply that underlying datasets are unusable in absolute terms. Rather, it evaluates whether datasets are operationally usable through automated, federated workflows without manual intervention. In many cases, datasets may remain usable to expert users who manually navigate source portals, interpret documentation and resolve access issues. The results therefore reflect a gap between human-mediated usability and machine-operational readiness, which is increasingly critical for real-time systems, cross-border integration and reproducible analytics. This makes metadata quality a governance issue for machine readiness, not only a cataloging issue for human discoverability.

Methodologically, this work demonstrates both the feasibility and the necessity of task-specific metadata assessment. By operationalizing management tasks into weighted metric profiles and applying them to real harvested metadata, we move beyond generic repository scoring to measurements that directly predict usability for concrete operational tasks. The combination of automated scoring, a composite impact proxy and validation against real portal records creates a chain of evidence (from metadata attributes to operational outcomes) that is persuasive for both technical and policy audiences. Our reproducible SPARQL queries and Python tools make this approach transferable: portal managers can run periodic audits, and researchers can adapt profiles to other sectors or regions.

Three patterns in the results deserve particular attention. First, metadata deficiencies at the federated level are not heterogeneous but universal. Both floods and groundwater datasets show uniformly low sufficiency scores (means of 0.15–0.25), and the near-total failure rates for mapped license metadata and machine-readable format declarations indicate that the problem lies in portal-level metadata practices and federation infrastructure rather than in domain-specific traditions. This universality has an important policy implication: interventions should target the federation pipeline and portal ingestion workflows rather than individual dataset publishers, because even well-curated metadata at the national level may not survive the harvesting and mapping process that feeds data.europa.eu. This interpretation is consistent with prior findings that metadata quality degrades during cross-portal federation (Neumaier et al., 2016), and it reframes these failure rates not necessarily as an indictment of data producers but as a diagnosis of information loss in the metadata supply chain. We note, however, that the absence of license metadata was also confirmed on all 13 accessible source portal pages (Table 7), suggesting that for this particular field, the gap originates upstream of the federation process. For the remaining fields, we cannot attribute the loss to a stage: they may be absent at source or present but not propagated, and our design does not distinguish these.

Second, what we initially took to be a discovery gap turns out to be an availability failure of the discovery layer itself, which is a different finding and, for the argument of this paper, a more direct one. Our harvest queried all seven domain keyword sets and obtained records for only two. The post hoc audit reported in the "Overall metadata sufficiency" section establishes that the queries for the other five did not complete: on re-execution, all 30 of their keywords returned HTTP504 or failed to respond within 120 s, and no query in a 43-query re-run returned an empty result set. Records that those queries were intended to retrieve are demonstrably present in our own corpus, mislabeled under a later-iterating domain.

The consequence for the user is the same in either case, but the diagnosis is not. We cannot say that water quality, WFD metrics, utilities, agricultural runoff or cross-cutting datasets are absent from or invisible within the federated catalog, and we do not. What we can say is that the EU's principal federated data portal did not reliably answer ordinary substring queries about them, on two occasions 5 months apart, and that its behavior varied between those occasions in a way unrelated to what it holds. For a compliance officer or basin planner without prior knowledge of national catalog structures, an endpoint that times out is indistinguishable from one that holds nothing. This is precisely the distinction between availability and operability that the paper is about, arising at the discovery layer rather than in the metadata records themselves, and it suggests that assessments of open data infrastructure should treat endpoint reliability as a measured property rather than a precondition. We identify index-friendly paginated harvesting, multilingual query expansion and direct comparison against source catalogs as priorities for future investigation.

Third, infrastructure instability, manifested by the 52% rate of broken or retired URLs among the unique URLs in our manual validation, emerged as an independent and high-impact problem. Even where metadata fields exist and are well-formed, link rot prevents downstream use. In systems that presume persistent access (automated harvesters, ingestion pipelines, inter-agency workflows), this fragility translates into repeated human labor, loss of timeliness and potentially missed regulatory deadlines. We do not treat this rate as an estimate of prevalence across the corpus,

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

since the sample was purposive and small. Its significance lies elsewhere: these were among the most prominent and institutionally supported datasets in European water governance, deliberately chosen as a best case and more than half of their access links had nonetheless ceased to resolve. That is sufficient to establish that link persistence is not being systematically maintained, and it supports the case for persistent identifiers and active link monitoring as standard practice for operational data infrastructure.

A notable aspect of the results is that the two most consequential metadata failures, missing license declarations and missing format declarations (absent for 99.95% and 100% of harvested records, respectively), are arguably the easiest to remedy. Unlike description quality or temporal coverage, which require substantive knowledge of the underlying data, license and format fields can often be populated through administrative action at the portal level. That effectively no record in our harvest carries either field is consistent with ingestion and federation workflows that neither require nor validate them, though we cannot observe those workflows directly and other explanations are possible.

The country-level analysis reinforces the systemic interpretation. No socioeconomic indicator (GDP, digital maturity, open data portal development, e-government adoption or water stress) significantly predicts metadata quality across the nine Member States for which reliable estimates could be computed. Countries with the highest digital infrastructure scores do not produce better-described water data; indeed, the Netherlands (DESI = 67.4) achieved the lowest composite sufficiency in the sample (0.138), while Ireland (DESI = 56.4) achieved the highest (0.273). Although this exploratory finding should be interpreted cautiously given the small number of countries, it suggests that metadata quality gaps in the EU water domain are not driven by a lack of digital capacity or economic resources but by the absence of enforceable metadata quality standards at the portal level. This bears on the ongoing WFD revision in a specific way. If the gap were driven by uneven digital capacity, capacity-building targeted at lagging Member States would be the natural remedy. Our results give no support for that diagnosis, and point instead toward interventions applying uniformly across Member States, such as mandatory metadata fields validated at the point of ingestion. We state this as an implication of the observed pattern rather than as an established causal claim: we observe that variation in metadata quality is not explained by the capacity indicators available to us, which is not the same as demonstrating that a uniform mandate would close the gap.

These findings arrive at a particularly consequential moment for EU water policy. The European Commission's 2026 targeted revision of the Water Framework Directive aims to address regulatory bottlenecks while maintaining high environmental standards. Our results identify a constraint that receives little attention in current policy discussion: the data infrastructure underpinning WFD compliance reporting, cross-border coordination and early warning is impaired by metadata deficiencies that appear tractable and inexpensive to address relative to the monitoring investments they affect. One design response would be to attach minimum metadata requirements to datasets reported under Article 5 (characterization), Article 8 (monitoring) and Article 15 (reporting), covering license declaration, machine-readable format specification, temporal coverage and spatial referencing.

We advance this as a design implication rather than a policy prescription, and the distinction is material. Our evidence establishes the presence and distribution of metadata deficiencies in the federated catalog. It does not establish that any particular legislative instrument would remedy them, nor do we observe the ingestion and federation workflows through which such a requirement would have to operate. Whether ingestion-level validation rules would in practice raise metadata quality, and at what administrative cost, is an empirical question that could only be answered by evaluating an implementation. We set out the observed deficiencies, the points in the data supply chain at which they appear, and the estimated sufficiency gain associated with correcting each; the choice of instrument, and the assessment of whether its benefits justify its costs, properly rests with the institutions conducting the revision.

### Metadata requirements under increasingly capable AI systems

A reasonable objection to the framing of this paper is that it may be measuring a problem that is dissolving. General-purpose language models are increasingly able to parse heterogeneous and semi-structured material, extract structure from prose documentation and interpret records that would defeat a conventional parser. If a model can read a PDF landing page and infer what a dataset contains, its temporal extent and its spatial coverage, the practical distinction between human-readable and machine-readable material narrows considerably for discovery and preliminary interpretation. On this view, some of the deficiencies we document might be absorbed by improving inference rather than corrected at source.

We think this is partly right, and it changes which deficiencies matter rather than removing the need for metadata. The distinction turns on what can be inferred from content and what cannot. Description quality, keyword richness, and to some extent temporal and spatial extent are properties that a capable model can often reconstruct from the resource itself, and improving inference does reduce their marginal cost. Licensing status is not of this kind. A model cannot determine from a dataset's content whether its publisher permits redistribution; that is a legal fact about the publisher's intent which exists only if it has been recorded. Provenance is similar: a processing history that was never documented cannot be recovered by reading the output. And where results enter statutory reporting or inform decisions with legal consequence, an inferred attribute carries a different evidentiary status from a declared one, however accurate the inference. A compliance submission resting on a model's guess about license terms is not equivalent to one resting on a stated license, and would not survive audit.

This has a bearing on how our results should be read. The two dimensions absent for effectively every record in our corpus, license declaration and format specification, are precisely those least amenable to inferential substitution: the first because it encodes intent rather than content, the second because it governs automated ingestion at a point in the pipeline that precedes interpretation. The metadata deficiencies most likely to persist as AI capability improves are therefore the ones our framework identifies as binding. More capable inference raises the value of the fields that must be declared rather than reducing it, because it removes the surrounding friction that previously obscured their importance. The broader shift toward AI-assisted analysis in environmental governance (Raeissi and Knapen, 2025) accordingly strengthens rather than weakens the case for reliable access, explicit licensing, documented provenance and verifiable results as foundations of public decision-making. This argument connects to a broader concern now being voiced across economics and AI research. The *We Must Act Now* statement (Stanford Digital Economy Lab, 2026), signed by a large group of economists, policymakers and technology researchers including several Nobel laureates in Economic Sciences,

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

holds that AI may drive an economic transformation larger than the Industrial Revolution but unfolding over a far shorter period, and that the work of building the incentives, guardrails and institutions needed to steer it cannot be deferred. That debate has been conducted largely in terms of labor markets and macroeconomic adjustment. Our study speaks to a quieter and more specific part of the same agenda. If AI-assisted analysis is to support environmental governance, it must operate on public data infrastructure that records what can only be declared and not inferred. The institutional groundwork this requires is unglamorous: license fields populated at ingestion, format declarations validated before publication, access links that persist across reporting cycles. Our results indicate that in one policy domain of considerable regulatory and financial significance, this groundwork is largely absent. We offer the framework as a means of measuring that gap sector by sector, on the view that the institutional capacity the statement calls for will be built domain by domain rather than in the abstract.

### *Limitations and future research*

Several limitations should be acknowledged. First, the SPARQL harvest captures the federated metadata view as exposed by data.europa.eu, not the metadata as it may exist on original source portals. Some fields, particularly format and controlled vocabulary keywords, may be richer at the national level but lost during federation. Our results therefore characterize the European-level discovery experience rather than the full metadata landscape, although for cross-border applications this federated view is the operationally relevant one. Second, the harvest retrieved datasets from only two of seven queried water domains, and "Overall metadata sufficiency" section establishes that the other five failed at the endpoint rather than returning empty sets. Metadata patterns in those domains are therefore simply unobserved, and no inference about them should be drawn from this study. The original harvest did not log per-domain query outcomes, so its behavior had to be reconstructed after the fact; we record this as a methodological failing, and note that per-query outcome logging is a minimal requirement for any harvest whose null results are to carry interpretive weight. A further consequence is that every completed query returned the maximum 500 URIs, so the corpus is truncated at the keyword level and the reported counts are lower bounds. Third, the manual validation sample ($n = 25$ stratified records) is intentionally small, designed to check extraction fidelity and reveal obvious infrastructural problems; a larger validation ($n \geq 100$) would refine estimates of agreement and broken-link prevalence and enable domain-level comparisons. Fourth, the assessment focuses on metadata quality and accessibility, not on the intrinsic data quality (accuracy, sensor calibration, sampling design) that is critical for many scientific inferences. Fifth, we did not conduct systematic side-by-side comparisons of federated versus source-portal metadata for the same datasets; such comparisons would clarify where in the metadata supply chain specific fields are lost and should be a priority for future work. Sixth, the country-level analysis is exploratory, while the small number of countries ($n = 9$) limits statistical power, however including more countries (we also tried with 20) did not produce meaningful results as there were many of them with too few datasets. Currently, country-level comparisons should be interpreted with caution, as dataset counts vary substantially across Member States and reflect federated discoverability rather than true national data availability. Seventh, the manual validation sample was purposive rather than random, comprising prominent datasets from major EU publishers; the observed rates therefore describe a best-case selection and should not be read as population estimates, and three sampled records resolve to a single shared landing page, reducing the effective sample to 23 unique URLs. Eighth, our robustness testing varied one dominant weight at a time and the readiness threshold; we did not perturb multiple weights jointly, compare alternative functional forms for the individual scoring rules or test alternative treatments of missing values, each of which would further probe the specification. Ninth, the scoring rules for license openness and format readability have known deficiencies discussed in "Common metadata failures" section: the first conflates legal clarity with permissiveness, and the second imposes a single format ordering across tasks with differing requirements. Neither materially affects the results reported here, since format is zero for all records and license for all but one, but both would need revision before the framework were applied to a corpus in which these fields are populated. Tenth, our scoring model is additive and therefore compensatory, whereas some metadata requirements function as gating conditions; the two specifications coincide in this corpus but would diverge in one where license and format metadata are present.

Future research should build on this framework in several directions. Querying national portals directly (rather than through the federated layer) would establish whether the metadata gaps we observe originate at the source or are introduced during federation – a distinction with direct implications for where remediation efforts should be targeted. Expanding the harvest to additional water domains, potentially through direct API queries to the EEA Datahub and JRC Data Catalog, would test the generalizability of our findings beyond floods and groundwater. Future analyses could clarify whether the low means reflect universally poor metadata or are driven by a subset of particularly deficient records. Longitudinal monitoring of the same datasets over time would reveal whether metadata quality is improving, static or degrading in response to policy interventions such as the High-Value Datasets Regulation. Three methodological extensions follow directly from the limitations above. First, the weight derivation should be placed on firmer empirical footing through formal elicitation, either the Analytic Hierarchy Process as applied to metadata quality by Kubler et al. (2018), or a Delphi procedure with flood forecasters, compliance officers and basin managers; comparison against an equal-weight benchmark would additionally establish how much the task differentiation contributes. Second, the scoring rules should be refined by separating license clarity from license permissiveness into distinct dimensions, and by scoring format suitability separately for each task rather than through a single ordering. Third, an explicitly non-compensatory variant, in which license and format function as gating conditions rather than as weighted contributions, would allow the compensatory assumption to be tested directly.

The priority measure also admits a natural extension. As currently specified, it combines operational importance with unmet metadata need, but says nothing about what remediation would cost. A fuller formulation would incorporate the estimated repair cost of each metadata field, which differs substantially between administrative fields such as license declaration and substantive fields such as provenance documentation; expected usage intensity; statutory importance under the governing directive; the population affected by decisions the dataset informs; and the expected value of the improved decision quality that better metadata would enable. Such a formulation would allow data custodians to compare the expected benefit of remediation against its cost, and would move the framework from diagnosis toward investment appraisal, which is the form in which resourcing decisions are actually taken. Finally,

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press

collaboration with portal operators to implement the recommended fixes would provide an empirical test of whether the estimated sufficiency gains are realized in practice and whether they translate into measurable improvements in data reuse.

## Conclusions

Our analysis leads to a clear and pragmatic conclusion: improving the operational value of EU water data is not primarily a question of collecting more measurements but of fixing metadata and infrastructure failures that currently prevent reuse. The EU Water Dataset Observatory shows that none of 2,176 water datasets harvested from data.europa.eu achieve operational readiness for any management task – not for flood early warning, not for WFD compliance reporting and not for cross-border coordination – as assessed through the federated discovery layer at the time of harvest. This zero-readiness finding is driven by a small number of tractable problems, most critically the absence of machine-readable format metadata and license declarations across effectively the entire harvested corpus, and we show that it follows arithmetically from those gaps rather than from the particular weights we assign. The finding that these gaps are uniform across Member States regardless of economic or digital development underscores that the problem is one of governance, not capacity. Of course, the policy implication is not that EU water datasets are unusable, but that they are often not yet machine-ready in the form required by modern automated governance systems.

These appear to be fixable problems. The most direct design response would be metadata quality assurance at ingestion: automated checks flagging records that lack license fields, format declarations or temporal coverage, paired with lightweight workflows allowing publishers to correct deficiencies before records are released. The current situation, where 99.95% of harvested records lack mapped license metadata despite the underlying datasets being freely accessible, creates unnecessary legal uncertainty that discourages automated reuse and undermines the open-data mandate that these portals are designed to fulfill.

The reliability of the discovery layer is equally consequential. Five of seven water-management domains returned no records from the EU's primary federated data portal, and the audit establishes that their queries failed rather than completing empty: on re-execution, all 30 of their keywords returned gateway timeouts or no response, and across 43 queries not one empty result set was observed. We do not conclude that these datasets are absent. We conclude something narrower and, for an infrastructure assessment, harder to dismiss: that the portal did not answer ordinary queries about them, that its behavior changed between two runs 5 months apart, and that a user has no way to tell an endpoint that holds nothing from one that cannot answer in time. Closing this gap requires attention to endpoint capacity and query planning at the federated layer, alongside how national catalogs map their metadata to DCAT-AP and how the federation pipeline handles domain-specific vocabularies and multilingual keyword sets.

Infrastructure stability must also be institutionalized. Persistent identifiers, active link monitoring and clear archiving policies (or mirrored repositories for critical operational feeds) would reduce the brittle dependency on ephemeral URLs that left more than half the access links in our validation sample unresolvable. Given limited maintenance resources, improvements should be targeted where they deliver the greatest operational return: our priority metric provides an evidence-based triage that places high-importance, low-sufficiency records first.

Finally, metadata governance must become task-aware. One-size-fits-all FAIR checks are necessary but insufficient; portals should offer domain-specific templates and validation profiles that reflect the distinct needs of early warning (high temporal cadence and coordinate precision), compliance reporting (provenance and comprehensive temporal coverage) and cross-border coordination (controlled vocabularies and multilingual fields). The ongoing targeted revision of the Water Framework Directive offers an occasion on which requirements of this kind could be considered, for instance by attaching a minimum set of machine-readable metadata fields to datasets submitted under WFD reporting obligations. Embedding such templates into portal ingestion forms, procurement specifications and funding criteria would plausibly align publisher incentives with operational needs, though we emphasize that the effectiveness of any such measure is an empirical question that our study does not settle.

Taken together, ingestion-level metadata requirements, format and license declarations, federation quality assurance, link resilience and task-specific templates constitute a coherent agenda for raising the operational utility of existing EU water data. Our evidence establishes what is presently missing and where the largest gaps lie; whether addressing them delivers the improvements in flood forecasting, regulatory compliance and transboundary basin management that we anticipate is a proposition that implementation would test.

**Open peer review.** To view the open peer review materials for this article, please visit http://doi.org/10.1017/wat.2026.10030.

**Author contribution.** Conceptualization: A.A., P.K; methodology: A.A., G.F.; software: G.F.; validation: A.A., G.F.; formal analysis: A.A., G.F.; investigation: A.A.; data curation: G.F.; writing – original draft: A.A.; Writing – review and editing: A.A., G.F., P.K.; visualization: G.F.; supervision: P.K. All authors approved the final submitted draft.

**Competing interests.** The authors declare no competing interests.

**Data availability statement.** All materials required to reproduce the results reported here are available. The repository is public as of the date of this revision and comprises: the assessment framework code in Python; the exact SPARQL harvesting queries together with the RDF predicate mappings and fallback rules applied to each metadata dimension; the harvested metadata snapshot of 2,176 records, versioned and timestamped to the harvest date of 18 March 2026; the dataset-level validation coding for all 25 manually verified records, including URLs, publisher, per-dimension Yes/Partial/No/Not-assessable values and free-text notes; the country-level equity analysis script and merged indicator table; the interactive visualizations; the script reproducing the country-level correlations reported in Table 8 under the 10-dataset restriction; and the post hoc harvest audit, comprising the keyword attribution script, the Phase 1 re-execution script and its per-query outcome log for 22 August 2026. The repository is available at https://github.com/AE4RIA/ae4ria-eu-water-dataset-observatory.

Two constraints on reproducibility should be noted. The scoring, sensitivity, ceiling, validation and country-level analyses reported in this paper are fully reproducible from the deposited snapshot, since they operate on archived data. The harvest itself is reproducible in procedure but not necessarily in result: data.europa.eu is a live federated catalog whose contents and DCAT-AP mappings change over time, so re-running the harvesting queries at a later date will not return an identical corpus. This is why we deposit the March 2026 snapshot rather than relying on the queries alone, and why all reported figures are stated with reference to the harvest date.

**Ethics and approvals.** The research meets all ethical guidelines. No human subjects were involved in this research.

**Funding statement.** P.K. acknowledges funding from the European Research Council (ERC) under the ERC Synergy Grant Water-Futures grant agreement ID 951424.

## References

**Aguilar Gómez F and Bernal I** (2023) FAIR EVA: Bringing institutional multidisciplinary repositories into the FAIR picture. *Scientific Data* **10**(1), 764. https://doi.org/10.1038/s41597-023-02652-8.

**Alamanos A and Koundouri P** (2024) Future challenges of water resources management. In Koundouri P and Alamanos A (eds), *Future Challenges of Water Resources Management*. Edward Elgar Publishing. https://doi.org/10.4337/9781802202946.00028.

**Alamanos A, Latinopoulos D, Xenarios S, Tziatzios G, Mylopoulos N and Loukas A** (2019) Combining hydro-economic and water quality modeling for optimal management of a degraded watershed. *Journal of Hydroinformatics* **21**(6), 1118–1129. https://doi.org/10.2166/hydro.2019.079.

**Alamanos A and Linnane S** (2021) Estimating SDG indicators in data-scarce areas: The transition to the use of new technologies and multidisciplinary studies. *Earth* **2**(3), 635–652. https://doi.org/10.3390/earth2030037.

**Alamanos A, Rolston A and Papaioannou G** (2021) Development of a decision support system for sustainable environmental management and stakeholder engagement. *Hydrology* **8**(1), 40. https://doi.org/10.3390/hydrology8010040.

**Arnal L, Cloke HL, Stephens E, Wetterhall F, Prudhomme C, Neumann J, Krzeminski B and Pappenberger F** (2018) Skilful seasonal forecasts of streamflow over Europe? *Hydrology and Earth System Sciences* **22**(4), 2057–2072. https://doi.org/10.5194/hess-22-2057-2018.

**Bahim C, Cassorrán-Amilburu C, Dekkers M, Herczog E, Loozen N, Repanas K, Russell K and Stall S** (2020) The FAIR data maturity model: An approach to harmonise FAIR assessments. *Data Science Journal* **19**(1), 41. https://doi.org/10.5334/dsj-2020-041.

*Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources* (1991).

*Directive 2000/60/EC of the European Parliament and of the Council Establishing a Framework for Community Action in the Field of Water Policy* (2000).

*Directive 2007/2/EC of the European Parliament and of the Council Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)* (2007a).

*Directive 2007/60/EC of the European Parliament and of the Council on the Assessment and Management of Flood Risks* (2007b).

**European Commission** (2020) *tech. rep. DCAT-AP – DCAT Application Profile for Data Portals in Europe, Version 2.0.1.* Publications Office of the European Union. Available at https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe.

**European Commission** (2024a) *data.europa.eu: The Official Portal for European Data.* Available at https://data.europa.eu (accessed January 2026).

**European Commission** (2024b) *tech. rep. The EUWater Framework Directive: Implementation Report 2024.* COM(2024) 796 final. Brussels: European Commission.

**European Environment Agency** (2016) *tech. rep. WFD Reporting Guidance 2016.* Copenhagen: European Environment Agency.

**European Environment Agency** (2023) *EEA Datahub: Environmental Data for Europe.* Copenhagen. Available at https://www.eea.europa.eu/data-and-maps.

**Hall J, Arheimer B, Borga M, Brázdil R, Claps P, Kiss A, Kjeldsen TR, Krieger J, Kundzewicz ZW, Lang M, Aronica GT, Bilibashi A, Boháč M, Bonacci O, Burlando P, Castellarin A, Chirico GB, Fiala K, Gaál L, Gorbachova L, Gül A, Hannaford J, Kohnová S, Koskela JJ, Macdonald N, Mavrova-Guirguinova M, Ledvinka O, Mediero L, Merz B, Merz R, Molnar P, Montanari A, Osuch M, Parajka J, Perdigão RAP, Radevski I, Renard B, Rogger M, Salinas JL, Sauquet E, Šraj M, Szolgay J, Viglione A, Volpi E, Wilson D, Zaimi K and Blöschl G** (2015) A European Flood Database: Facilitating comprehensive flood research beyond administrative boundaries. *Proceedings of the International Association of Hydrological Sciences* **370**, 89–95. https://doi.org/10.5194/piahs-370-89-2015.

**Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Holve E, Johnson SG, Goerg C, Liaw ST, Hamiton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN and Schilling L** (2016) A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs* **4**(1), 18. https://doi.org/10.13063/2327-9214.1244.

**Kubler S, Robert J, Neumaier S, Umbrich J and Le Traon Y** (2018) Comparison of metadata quality in open data portals using the analytic hierarchy process. *Government Information Quarterly* **35**(1), 13–29. https://doi.org/10.1016/j.giq.2017.11.003.

**Neumaier S, Umbrich J and Polleres A** (2016) Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality* **8**(1), 1–29. https://doi.org/10.1145/2964909.

**Nogueras-Iso J, Lacasta J, Ureña-Cámara MA and Ariza-López FJ** (2021) Quality of metadata in open data portals. *IEEE Access* **9**, 60364–60382. https://doi.org/10.1109/ACCESS.2021.3073455.

**Ochoa X and Duval E** (2009) Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries* **10**(2–3), 67–91. https://doi.org/10.1007/s00799-009-0054-4.

**OECD** (2020) *tech. rep. Financing Water Supply, Sanitation and Flood Protection: Challenges in EU Member States and Policy Options*. Paris: OECD Publishing. https://doi.org/10.1787/6893cdac-en.

**Pekel JF, Cottam A, Gorelick N and Belward AS** (2016) High-resolution mapping of global surface water and its long-term changes. *Nature* **540**(7633), 418–422. https://doi.org/10.1038/nature20584.

**Publications Office of the European Union** (2024) *tech. rep. Open Data Maturity Report 2024.* data.europa.eu. Available at https://data.europa.eu/en/publications/open-data-maturity.

**Raeissi MM and Knapen R** (2025) Applications of generative large language models in environmental science: A systematic review. *Advances in Environmental and Engineering Research* **6**(3), 1–15. https://doi.org/10.21926/aeer.2503028.

**Stanford Digital Economy Lab** (2026) *We Must Act Now: A Statement on AI's Transformation of the Economy.* Available at https://www.wemustactnow.ai/. Statement signed by economists, policymakers and technology researchers, including multiple Nobel laureates in Economic Sciences (accessed 20 August 2026).

**Thielen J, Bartholmes J, Ramos MH and de Roo A** (2009) The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences* **13**(2), 125–140. https://doi.org/10.5194/hess-13-125-2009.

**Wang RY and Strong DM** (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* **12**(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.

**Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC 't, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J and Mons B** (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3**(1), 160018. https://doi.org/10.1038/sdata.2016.18.

**Wu M, Brandhorst H, Marinescu MC, Lopez JM, Hlava M and Busch J** (2023) Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence* **5**(1), 122–138. https://doi.org/10.1162/dint_a_00162.

## Appendix: Dataset-level validation results

**Table A1.** Manual validation coding for all 25 sampled datasets

| ID | Group | Dataset | Publisher | URL | Desc. | Temp. | Fmt. | Lic. |
|---|---|---|---|---|---|---|---|---|
| 1 | Flood | River flood hazard maps, Europe | JRC | N | n/a | n/a | n/a | n/a |
| 2 | Flood | EFAS historical simulations | Copernicus | N | n/a | n/a | n/a | n/a |
| 3 | Flood | Copernicus EMS rapid mapping | Copernicus | Y | Y | Y | Y | N |
| 4 | Flood | European past floods database | EEA | N | n/a | n/a | n/a | n/a |
| 5 | Flood | Floods directive reported data | EEA | N | n/a | n/a | n/a | n/a |
| 6 | WQ | Waterbase – water quality ICM | EEA | Y | Y | Y | Y | N |
| 7 | WQ | Waterbase – biology | EEA | N | n/a | n/a | n/a | n/a |
| 8 | WQ | Bathing water quality | EEA | Y | Y | Y | Y | N |
| 9 | WQ | Waterbase – transit./coastal | EEA | N | n/a | n/a | n/a | n/a |
| 10 | WQ | Nutrients in freshwater | EEA | Y | Y | Y | Y | N |
| 11 | WFD | WISE WFD database[a] | EEA | Y | Y | Y | P | N |
| 12 | WFD | WISE spatial data (Bodies) | EEA | N | n/a | n/a | n/a | n/a |
| 13 | WFD | UWWTD reported data | EEA | Y | Y | Y | P | N |
| 14 | WFD | Ecological Status Dashboard | EEA | Y | Y | Y | P | N |
| 15 | WFD | Chemical status of water bodies | EEA | N | n/a | n/a | n/a | n/a |
| 16 | GW | Groundwater bodies (WISE)[a] | EEA | Y | Y | Y | P | N |
| 17 | GW | Nitrates in groundwater | EEA | N | n/a | n/a | n/a | n/a |
| 18 | GW | Groundwater quantitative status | EEA | Y | N | N | P | N |
| 19 | GW | Hydrogeological map, Europe | JRC/ESDAC | N | n/a | n/a | n/a | n/a |
| 20 | GW | GW-dependent ecosystems[a] | EEA | Y | Y | Y | P | N |
| 21 | Util | Eurostat water statistics | Eurostat | N | n/a | n/a | n/a | n/a |
| 22 | Util | SDG 6 indicators | Eurostat | N | n/a | n/a | n/a | n/a |
| 23 | Util | Water exploitation index | EEA | Y | Y | Y | Y | N |
| 24 | Util | Population connected to WWTP | Eurostat | Y | Y | P | P | N |
| 25 | Util | Water abstraction by source | Eurostat | Y | Y | P | P | N |

*Notes*: Y = Yes (1.0), P = Partial (0.5), N = No (0.0), n/a = Not assessable. Content dimensions are coded n/a where the access URL did not resolve, since no observation was possible. URL = access URL resolves; Desc. = description accurate; Temp. = temporal coverage accurate; Fmt. = declared formats available; Lic. = license explicitly stated on portal page. Dataset names are the descriptive labels used internally by the author team. The corresponding record identifiers and access URLs, as they appear in the harvested corpus, together with the free-text inspection notes, are given in the project repository. Agreement over the 13 accessible datasets: $(12+0)+(10+0.5\times2)+(5+0.5\times8)+0=32$ over $4\times13=52$ assessments $=61.5\%$. Excluding the license dimension, which measures presence rather than extraction fidelity, agreement is $32/39=82.1\%$.

[a]Records 11, 16 and 20 resolve to the same EEA landing page. On unique URLs, the sample comprises 23 records, of which 11 are accessible; agreement is then $27/44=61.4\%$ and $27/33=81.8\%$, respectively.

https://doi.org/10.1017/wat.2026.10030 Published online by Cambridge University Press