Revista de Științe Politice. Revue des Sciences Politiques • No. 92 • 2026: 364 - 379

ORIGINAL PAPER

# More Data, Less Transparency? A Data Pollution Index for Romania's National Open Data Portal

**Oana Stăiculescu[1)], Veronica Gheorghiță[2)]**

**Abstract:**
Open government data portals are usually assessed by the volume of data they publish. This article argues that the growing volume of open government data does not necessarily translate into effective transparency: when published datasets are outdated, redundant, inconsistent or poorly documented, data pollution may undermine their reliability, accessibility and practical usability. Drawing on the ecological approach to information and on the distinction between nominal and effective transparency, we define data pollution as the accumulation of unusable residues in an open data ecosystem, whose costs fall on re-users rather than publishers. To examine the gap between data availability and data quality, we develop a Data Pollution Index (DPI) and apply it to the full catalogue of Romania's national open data portal, data.gov.ro: 5,269 datasets and 39,899 resources from 131 institutions, harvested in November 2026. The index combines indicators of staleness, broken links, format barriers, missing metadata and confirmed redundancy. Three datasets in five show at least one severe form of pollution, and every publication cohort since 2013 contributes polluted datasets. Staleness dominates: 69% of continuously updated datasets have not changed in over two years, and only one recurring series in three is still continued. Almost all files from the portal's first two years are no longer accessible, CSV files account for 6.4% of resources, and 70.5% of datasets give no contact address. Heavy pollution on several dimensions is rarer (6% of datasets) and sensitive to how the

[1)] Lecturer, Ph.D., University of Craiova, Faculty of Law, Department of Public Law and Administrative Sciences, Craiova, Romania, Email: oana.staiculescu@edu.ucv.ro, Phone: +40 773 821 350, https://orcid.org/0000-0001-7740-2241.
[2)] Lecturer, Ph.D., University of Craiova, Faculty of Social Sciences, Department of Sociology, Philosophy and Social Work, Craiova, Romania, Email: veronica.gheorghita@edu.ucv.ro, Phone: +40 740 458 527.

index is specified, whereas institution rankings are reasonably robust. High-value datasets are less stale and better documented, but more often duplicated. The findings complement existing open data maturity assessments by documenting dataset-level barriers to accessibility and re-use. They suggest that open data policy should monitor the condition and usability of published datasets alongside their volume.
**Keywords:** *open government data, data pollution, transparency, data quality, composite index, Romania*
**JEL Classification: H83, C43, D83, O38**

## 1. Introduction

Open government data portals are usually judged by how much they publish. European policy reinforces this logic: the Open Data Directive (EU) 2019/1024 and the High-Value Datasets Implementing Regulation (EU) 2023/138 oblige public bodies to release more data, in more domains, through more channels. Yet volume is a weak proxy for transparency. Datasets that are outdated, duplicated, contradictory or poorly documented can satisfy disclosure obligations while remaining practically unusable (Lourenço, 2015; Vetrò et al., 2016).

Romania illustrates this tension. Its national portal, data.gov.ro, has operated since 2013, and its catalogue has grown through successive Open Government Partnership (OGP) action plans; in a single cycle, public institutions added about 700 datasets, a 41% increase (OGP, n.d.-a). International benchmarks have improved as well. In 2025, Romania was placed among the "followers" of the European Open Data Maturity assessment and scored 0.59 on the OECD OURdata Index, above the OECD average of 0.53 (data.europa.eu, 2025; OECD, 2026). At the same time, re-users and official evaluations have repeatedly reported uneven quality and missing metadata, and the government itself acknowledges that the portal's publication guidelines are rarely applied by institutions (OGP, n.d.-b; OGP, n.d.-c).

This article argues that the growing volume of open government data does not necessarily translate into effective transparency. When published datasets are outdated, redundant, inconsistent, or poorly documented, data pollution may undermine their reliability, accessibility, and practical usability. To examine the gap between data availability and data quality, we develop and apply a Data Pollution Index (DPI) to Romania's national open data portal.

Three research questions guide the analysis:

1. **RQ1:** Does the growing volume of open data on data.gov.ro translate into effective transparency, or mainly into nominal transparency?
2. **RQ2:** Which forms of data pollution prevail, and how are they distributed across institutions, domains and publication cohorts?
3. **RQ3:** What does the DPI reveal about the gap between Romania's declared open data maturity and the actual state of its data?

The article makes two contributions. Conceptually, it reframes low-quality open data as data pollution: an ecosystem-level phenomenon defined by accumulation, persistence, exposure and externalities, rather than a property of individual datasets. Methodologically, it builds the DPI from the complete metadata of data.gov.ro, harvested through the portal's API on 11 November 2026, and offers an initial validation of the index. Section 2 develops the theoretical framework, and Section 3 outlines the Romanian

context. Section 4 presents the construction of the DPI, and Section 5 reports the results. Section 6 discusses the findings in light of the research questions, and Section 7 concludes with policy recommendations.

## 2.Theoretical framework: from data quality to data pollution

### 2.1. Public data as an informational ecosystem

The ecological metaphor treats information not as a static stock but as a living environment of people, practices, technologies and values (Nardi and O'Day, 1999). Open government data research has adopted this lens to describe how data move between publishers, intermediaries and re-users, and how value emerges only when these actors interact (Harrison, Pardo and Cook, 2012; Zuiderwijk, Janssen and Davis, 2014).

We use three ecological concepts. *Data vitality* describes whether published data remain alive: updated, accessible and relevant over time. *Data recycling* describes the transformation of administrative information into new resources, such as research, applications, journalism and civic monitoring tools. *Data pollution* describes what happens when the life cycle breaks down and unusable residues accumulate in the ecosystem. This article measures the third concept and uses the first two to interpret it.

### 2.2. Nominal versus effective transparency

Heald (2006) distinguishes nominal transparency, the formal availability of information, from effective transparency, which requires that information can be received, processed and used by its audience. When the two diverge, a "transparency illusion" emerges: disclosure indicators rise while the public's capacity to scrutinise government does not. Fung, Graham and Weil (2007) make a similar point: disclosure works only when information is usable and fits into users' decisions.

Open data policy is particularly exposed to this illusion. Janssen, Charalabidis and Zuiderwijk (2012) list as a persistent myth the belief that publishing data automatically yields public value. Lourenço (2015) shows that government portals often publish data that cannot support accountability. Our first research question follows from this distinction: if catalogue growth is accompanied by growing pollution, volume measures nominal, not effective, transparency.

### 2.3. Why "pollution" and not only "poor quality"

Data quality is classically defined as fitness for use (Wang and Strong, 1996). It is assessed along dimensions such as accuracy, completeness, consistency and timeliness (Batini et al., 2009). In open data research, it has been operationalised at dataset level (Vetrò et al., 2016; Raca et al., 2022) and, for metadata, through the FAIR principles (Wilkinson et al., 2016) and the European Metadata Quality Assessment (MQA).

The term "pollution" is not new, and we use it in a distinct sense. Ben-Shahar (2019) introduced "data pollution" to describe the negative externalities of collecting personal data. Wardle and Derakhshan (2017) use "information pollution" for mis- and disinformation. We borrow Ben-Shahar's externality logic but apply it to public data. We define *data pollution* as the accumulation of outdated, redundant, inconsistent, inaccessible or poorly documented public data in an open data ecosystem, whose costs fall on re-users rather than on publishers.

Four properties, taken from environmental science, distinguish pollution from poor quality:

- **Accumulation:** pollution is a stock. What matters is the load of unusable residues in the whole catalogue, not the average score of a dataset.
- **Persistence:** pollutants remain. A dead link or an abandoned dataset that has stayed uncorrected for years weighs more than one that failed last week.

• **Exposure:** harm depends on who is exposed. A polluted dataset that many people look for matters more than an obscure one, much as an air quality index weighs pollution by the population that breathes it.

• **Externality:** the publisher does not pay the cost. Re-users absorb it in time spent searching, cleaning and verifying, or in errors carried into their own work.

Table 1 summarises the distinction.

**Table 1. Data quality and data pollution compared**

| Aspect | Data quality | Data pollution |
|---|---|---|
| Unit of analysis | Individual dataset | Catalogue, institution, ecosystem |
| Core question | Is this dataset fit for use? | How much unusable residue has accumulated, and who is exposed to it? |
| Time | Snapshot | Accumulation and persistence over time |
| Who bears the cost | The user of the dataset | Third parties: re-users, citizens, downstream services |
| Typical measure | Average quality score | Pollution load, persistence, exposure |
| Remedy | Correct the dataset | Clean up, retire and govern the data life cycle |

*Source: Authors' conceptual synthesis based on Wang and Strong (1996), Heald (2006), Ben-Shahar (2019), and the theoretical framework developed in this study.*

This distinction is what the DPI must capture to be more than another quality index. Section 4 shows how each property enters the index, and Section 5 tests the index with the validation checks set out in Section 4.5.

## 3. The Romanian context

Romania has a complete legal and strategic framework for open data, but evidence on implementation points to a persistent quality gap. Law no. 109/2007 transposed the first Public Sector Information Directive. Government Emergency Ordinance no. 41/2016 required institutions to publish datasets on the national portal. Law no. 179/2022 on open data and the re-use of public sector information transposed the Open Data Directive. In 2024, the government adopted a National Open Data Strategy and Action Plan for 2024–2028, whose objectives include publishing high-value datasets with quality assurance and encouraging re-use (data.europa.eu, 2025). Governance is shared by the General Secretariat of the Government and the Authority for the Digitalisation of Romania (data.europa.eu, 2024).

International benchmarks show progress, but an uneven one. In the OECD OURdata Index 2025, Romania scored 0.85 on data accessibility and 0.52 on government support for re-use, both above the OECD averages of 0.67 and 0.40. However, it scored 0.41 on data availability, below the OECD average of 0.53 (OECD, 2026). The European Open Data Maturity assessment placed Romania in the "followers" cluster in 2025 (data.europa.eu, 2025).These international assessments provide valuable but different perspectives on open data maturity. The OECD OURdata Index primarily evaluates the governance, accessibility and re-use of open government data, while the European Open Data Maturity framework also examines quality and impact. In 2025, Romania scored 71.4% on the European quality dimension and 49.1% on impact, indicating that weaknesses are already recognised by existing assessments (data.europa.eu, 2025). The DPI complements these benchmarks by directly examining the observable condition of individual datasets and resources within the national catalogue.

On that condition, the record is critical. During the 2014–2016 OGP cycle, civil society and hackathon participants complained that the portal favoured quantity over quality and that usable metadata were almost absent (OGP, n.d.-b). The 2020 Open Data Maturity assessment named irregular, non-standardised and non-interoperable publication as a barrier to re-use (data.europa.eu, 2020). The 2025–2027 OGP action plan acknowledges that the guidelines available on data.gov.ro are rarely applied by institutions (OGP, n.d.-c). Romania therefore offers a suitable case for testing whether a growing catalogue can coexist with growing pollution.

## 4. Methodology: constructing the Data Pollution Index

### 4.1. Data

The analysis covers the full public catalogue of data.gov.ro. On 11 November 2026, we harvested the metadata of all 5,269 public datasets, their 39,899 resources (files) and the 131 institutions that publish them, through the portal's CKAN Action API. We then tested every resource URL with an HTTP request: HEAD first, then a partial GET when the server refused HEAD, with a 25-second timeout. An external validation against the European Metadata Quality Assessment (MQA) scores was initially planned but could not be completed for the November 2026 snapshot. This limitation is discussed in Sections 4.5 and 6.5.

### 4.2. Indicators

The DPI combines five indicators grouped in the three dimensions named in our thesis: reliability, accessibility and usability. Each indicator is scored between 0 (no pollution) and 1 (maximum pollution) for every dataset. Because all indicators are shares or bounded scores, no further normalisation is needed, and the index does not depend on the sample's minimum and maximum.

**Table 2. Structure of the Data Pollution Index**

| Dimension | Indicator | Operationalisation (0 = clean, 1 = polluted) | Property captured |
|---|---|---|---|
| Reliability | R1 Staleness | The portal has no update-frequency field, so fixed thresholds apply. Continuous datasets: time since a file was last added or modified; 0 below 12 months, 0.5 at 12–24 months, 1 above 24 months. Year-specific datasets (a reference year in the title) are assessed through their series: the latest edition of a recurring series scores 0 if its reference year is 2025 or 2026, 0.5 if 2024 and 1 if 2023 or earlier. Earlier editions and one-off year-specific datasets are historical by nature and excluded. Two datasets belong to the same series when they come from the same institution and their titles are identical once reference years, diacritics, punctuation and letter case are removed | Persistence |
| Accessibility | A1 Broken links | Share of the dataset's resources that return an HTTP error (400 or above) or cannot be reached; resources without a link count as inaccessible | Accumulation |

| Dimension | Indicator | Operationalisation (0 = clean, 1 = polluted) | Property captured |
|---|---|---|---|
| Accessibility | A2 Format barriers | Format-related reusability barriers, averaged over the dataset's resources: open, structured formats (CSV, JSON, XML, RDF, ODS, geospatial formats, APIs) = 0; structured but proprietary (XLS, XLSX) = 0.5; unstructured or opaque formats, whose content cannot be processed or identified without opening them (TXT, ZIP, RAR) = 0.75; documents, images and undeclared formats (PDF, DOC, DOCX, ODT, HTML, empty) = 1 | Accumulation |
| Usability | U1 Missing metadata | Share of five elements missing: description of at least 50 characters, at least one tag, an open licence, a contact e-mail, a declared format for every resource | Externality |
| Usability | U2 Redundancy | 1 if redundancy is confirmed: the dataset repeats a file (same SHA-1 hash in the portal's archiver records) or a URL among its own resources, or contains a file identical to one in exactly one other dataset. Files shared by three or more datasets are treated as auxiliary files (nomenclatures, field descriptions) and not counted. Probable redundancy (similar titles, URLs shared with other datasets, same name, format and size) is reported separately. Otherwise 0 | Accumulation |

*Source: Authors' own operationalisation of the Data Pollution Index (DPI), based on the theoretical framework and methodological decisions presented in Section 4.*

Inconsistency, the fourth form of pollution in our definition, cannot be detected from metadata. Measuring it requires manual inspection of file contents, so it is outside the scope of this first version of the index (Section 6.5).

### 4.3. Aggregation and weights

The three dimensions receive equal weights, fixed before the analysis. Within each dimension, indicators are averaged:

$$DPI_i = \frac{1}{3} R1_i + \frac{1}{3} \cdot \frac{A1_i + A2_i}{2} + \frac{1}{3} \cdot \frac{U1_i + U2_i}{2}$$

When R1 is not defined because a dataset is historical by nature, the DPI is the mean of the two remaining dimensions, whose weights rise from one third to one half each. Section 5.6 reports the alternative of scoring these datasets 0 on R1.

The dataset scores are then aggregated in two ways that reflect the ecological properties in Table 1. The *pollution load* of an institution or of the whole catalogue is the share of its datasets with a DPI of 0.5 or above (accumulation). DPI and pollution load are also computed for each publication cohort (year of first publication), which shows whether the stock of pollution grows with the catalogue. Data.gov.ro exposes no view or download counts, so the exposure-weighted variant of the index could not be computed; exposure is discussed qualitatively in Section 6.

To describe how widespread pollution is beyond the index, we also count the datasets that show at least one severe form of it: staleness of more than 24 months or an

abandoned series (R1 = 1), at least one broken link, only documents, images or undeclared formats (A2 = 1), at least two missing metadata elements, or confirmed redundancy.

**4.4. Scope of the index**

The DPI covers the forms of pollution that can be measured for the whole catalogue: staleness, inaccessibility, format barriers, missing metadata and redundancy. Inconsistency (totals that do not add up, classifications that change between editions, values that differ from official statistics) requires reading the files themselves and is left for future work.

**4.5. Validation strategy**

The DPI is a formative index: its indicators define the construct rather than reflect a common latent cause. Internal consistency measures such as Cronbach's alpha therefore do not apply. Instead, we check that indicators are not redundant (Diamantopoulos and Winklhofer, 2001) and follow the OECD and JRC (2008) guidance on composite indicators with four tests:

1. **Non-redundancy:** the indicators should be only weakly correlated with each other (Spearman's rho), since each is meant to capture a distinct form of pollution.
2. **Discriminant validity against metadata quality:** a comparison with the European MQA scores of the same datasets was planned, to show that the DPI does not simply reproduce metadata quality. It could not be completed for this snapshot and is left for future work (Section 6.5).
3. **Known groups:** high-value datasets (identified through their HVD tags on the portal), which are subject to stricter EU publication rules, were expected to exhibit lower DPI scores than other datasets, providing a theoretically motivated comparison for assessing the index's ability to distinguish between groups. We report the Mann–Whitney U test with the rank-biserial correlation as effect size, for the index and for each indicator. HVD tags are added by publishers, so they may not cover every dataset that legally falls in this category.
4. **Robustness:** institution rankings are recalculated with equal weights across all five indicators, with weights derived from principal component analysis and by dropping one indicator at a time. Rank stability is reported as Spearman's rho between the alternative rankings. We also vary the staleness thresholds, the treatment of historical datasets and the definitions of broken links and redundancy, and report how the share of heavily polluted datasets changes.

## 5. Results

Data pollution on data.gov.ro is widespread but mostly partial. Three datasets in five (60.0%) show at least one severe form of pollution, and 15.8% show two or more. Heavy pollution across dimensions is rarer: 6.1% of datasets (322) have a DPI of 0.5 or above, a share that ranges from 2% to 24% depending on how the index is specified (Section 5.6). The mean DPI is 0.26 (median 0.23, interquartile range 0.18–0.37).

**5.1. The catalogue**

On 11 November 2026, data.gov.ro held 5,269 datasets with 39,899 resources, published by 131 institutions. Publication is highly concentrated: the Ministry of Regional Development and Public Administration alone accounts for 27.7% of datasets, and the five largest publishers for 50.7%. Most datasets (80.8%) are year-specific editions, with a reference year in the title, rather than continuously updated resources. The portal offers no field for declaring an update frequency and no usage statistics.

### 5.2. Forms of pollution

The prevalence of the five forms of data pollution identified through the Data Pollution Index (DPI) is summarised in Table 3. The results provide an overview of the main deficiencies affecting the accessibility, reliability and usability of datasets published on Romania's national open data portal.

**Table 3. Prevalence of the forms of data pollution on data.gov.ro**

| Indicator | Measure | Value |
|---|---|---|
| R1 Staleness | Continuous datasets not updated for more than 24 months | 69.0% (median: 51 months since the last update) |
| R1 Staleness | Recurring series not continued in 2025 or 2026 | 66.2% of 621 series |
| A1 Broken links | Unique links that could not be reached | 1,243 of 39,801 (3.1%) |
| A1 Broken links | Resources that cannot be opened, including 82 without a link | 1,329 of 39,899 (3.3%); 289 datasets affected, 145 entirely |
| A2 Format barriers | Resources in XLS or XLSX | 42.4% |
| A2 Format barriers | Resources in CSV | 6.4% |
| A2 Format barriers | Resources in open machine-readable formats, excluding the environment agency | 21.2% (53.9% including it) |
| U1 Missing metadata | Datasets without a contact e-mail | 70.5% |
| U1 Missing metadata | Datasets with no description, or one under 50 characters | 35.2% |
| U1 Missing metadata | Datasets without tags | 20.2% |
| U1 Missing metadata | Datasets without an open licence | 3.7% |
| U2 Redundancy | Datasets with confirmed redundancy | 8.6% (452) |
| U2 Redundancy | Datasets containing a file identical to one in another dataset | 316 |
| Probable redundancy | Datasets with probable redundancy only | 5.7% (300) |

*Note: the 39,899 resources include 82 without a link and 16 that repeat a link already used by another resource, which leaves 39,801 unique links. The 1,243 unreachable links are used by 1,247 resources, because four of them appear twice; together with the 82 resources without a link, 1,329 resources cannot be opened.*
*Source: Authors' calculations based on metadata harvested from data.gov.ro through the CKAN Action API and automated resource accessibility checks, snapshot of 11 November 2026.*

Staleness is the dominant form of pollution. Of the 1,647 datasets whose timeliness can be assessed, 62.1% are stale, and the typical continuous dataset was last updated more than four years ago. Files are rarely replaced on the portal: in 90.6% of the resources with both dates, creation and last modification fall on the same day. Updates therefore appear as new files, which is why R1 uses the most recent date of any file in the dataset. Format pollution is less visible in the aggregate than it is in practice. The National Environmental Protection Agency publishes 42.0% of all resources, almost all as XML air-quality files; without it, only one resource in five is in an open machine-readable format, and CSV

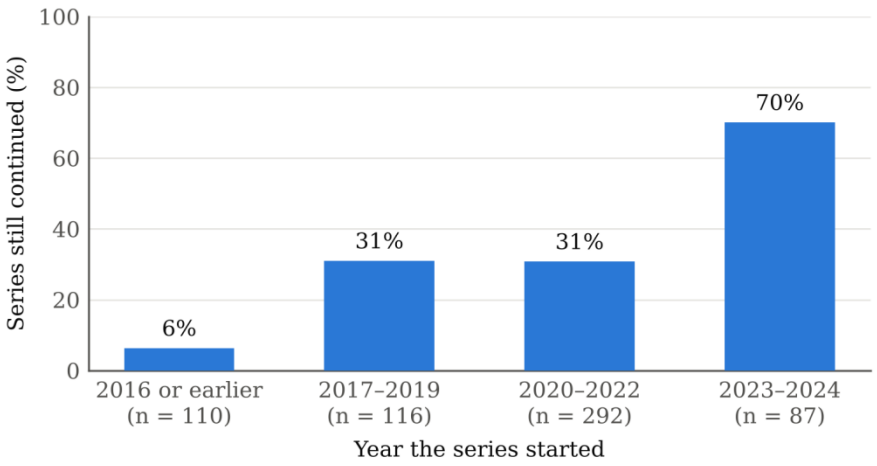
accounts for 6.4% of all files. Missing metadata mostly affects accountability: seven datasets in ten name no contact address for questions or corrections. Redundancy is smaller but concrete: 316 datasets contain a file byte-for-byte identical to one in another dataset. Inspection of these pairs shows the same data published twice under slightly different titles, an updated file published next to the old one, or a file published under another agency's dataset.

Broken links are rare in recent data but almost universal in the oldest. Of the datasets first published in 2013 and 2014, 91% and 95% contain at least one dead link; from 2016 onwards the share stays below 3% in every cohort. The failures differ in kind: 929 are 503 errors, all from the portal's original storage; 212 are refused connections, mostly from a Ministry of Finance server; 58 are timeouts, 17 are 404 errors and 27 are other errors. Counting only failures that are certainly permanent (404 or 410 errors, refused connections, unknown domains, invalid or missing links) leaves 322 resources and changes the index little (Section 5.6).

**5.3. Vitality: series that stop**

Of the 621 recurring series on the portal, only 33.8% are still alive, meaning that their latest edition refers to 2025 or 2026. Survival declines steeply with age (Figure 1). Only 6.4% of the series started in 2016 or earlier are still continued, against 31% of those started in 2017–2022 and 70.1% of those started in 2023–2024; the 16 series started in 2025–2026 are continued by definition. Because the snapshot is cross-sectional, these shares describe how many series of each age are alive today, not how each series evolved.

**Figure 1. Share of recurring series still continued in 2025–2026, by the year the series started**



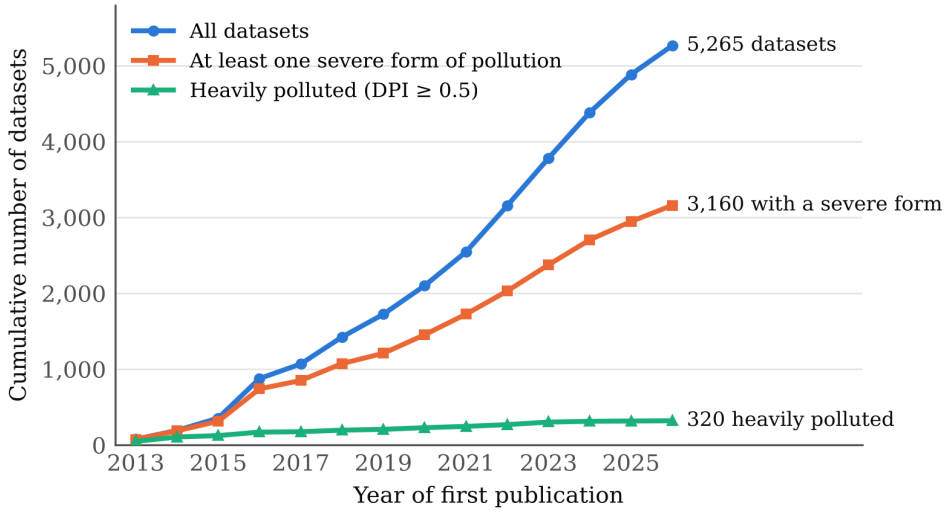

*Note: The 16 series initiated in 2025–2026 are excluded from the figure because they are classified as continued by definition. The four groups presented account for 605 of the 621 identified recurring series.*
*Source: Authors' calculations based on data.gov.ro metadata, snapshot of 11 November 2026.*

**5.4. Pollution by publication cohort**

Grouping datasets by year of first publication shows where the polluted datasets in today's catalogue come from (Figure 2). This is a retrospective view from a single snapshot: it counts the datasets published up to each year that are polluted in November

2026, not how many were polluted at the time. Every cohort contributes polluted datasets. Datasets with at least one severe form of pollution make up 97% of the 2013 and 2014 cohorts, whose files were largely lost, and between 46% and 83% of each later cohort. Heavily polluted datasets come disproportionately from the first two cohorts, which account for 106 of the 320 with a known publication year. Today's pollution therefore has two sources: the loss of early data and the ageing of what was published later. Whether the share of polluted data has grown over time can only be established with repeated snapshots.

**Figure 2. Datasets published up to each year, by their pollution status in the November 2026 snapshot (cumulative by year of first publication, 2013–2026)**




*Note: Pollution status was assessed using the November 2026 snapshot. The figure presents a retrospective distribution by publication cohort and does not represent historical year-by-year measurements of data pollution. Four datasets without a recorded creation date are excluded. Source: Authors' calculations based on data.gov.ro metadata and the Data Pollution Index (DPI), snapshot of 11 November 2026.*

### 5.5. Institutions

Among the 37 institutions with at least 20 datasets, the pollution load ranges from 0% to 46% and the mean DPI from 0.10 to 0.42 (Table 4). The largest publishers have low loads, 2% for the Ministry of Regional Development and Public Administration and 1% for the General Inspectorate for Emergency Situations, although many of their assessable datasets are stale. The highest loads belong to the "Dimitrie Gusti" National Village Museum and the National Institute of Heritage (46% each), the Ministry of Justice (36%) and the National Administration of Penitentiaries (34%); their datasets combine staleness with format barriers, missing metadata or dead links. The National Employment Agency shows that a large publisher can keep its catalogue clean: 197 datasets, a mean DPI of 0.10 and no heavily polluted dataset.

**Table 4. Data pollution by institution (institutions with at least 20 datasets; selection)**

| Institution | Datasets | Mean DPI | Pollution load |
|---|---|---|---|
| Ministry of Regional Development and Public Administration | 1,460 | 0.26 | 2% |
| General Inspectorate for Emergency Situations | 612 | 0.24 | 1% |
| National Environmental Protection Agency | 234 | 0.23 | 6% |
| National Employment Agency | 197 | 0.10 | 0% |
| National Institute of Administration | 167 | 0.25 | 1% |
| Ministry of Public Finance | 109 | 0.41 | 20% |
| National Trade Register Office | 82 | 0.13 | 0% |
| Ministry of Justice | 81 | 0.40 | 36% |
| General Anticorruption Directorate | 74 | 0.38 | 23% |
| National Agency for Fiscal Administration | 50 | 0.13 | 8% |
| Tulcea County Council | 50 | 0.10 | 0% |
| National Institute of Statistics | 49 | 0.27 | 12% |
| "Dimitrie Gusti" National Village Museum | 37 | 0.38 | 46% |
| National Administration of Penitentiaries | 32 | 0.42 | 34% |
| National Institute of Heritage | 28 | 0.36 | 46% |

*Source: Authors' calculations based on data.gov.ro metadata and the Data Pollution Index (DPI), snapshot of 11 November 2026.*

### 5.6. Validation

The five indicators are almost uncorrelated (Spearman's rho between −0.13 and 0.09), as expected for a formative index whose indicators capture distinct forms of pollution. The known-groups test gives a mixed result. The 99 datasets tagged as high-value have a slightly lower mean DPI than the rest (0.23 against 0.26), but the difference is not significant (Mann–Whitney U, one-sided $p = 0.07$) and the effect is small (rank-biserial $r = 0.09$). The indicators show why. High-value datasets are much less often stale (mean R1 of 0.42 against 0.72; $r = 0.46$) and much better documented (U1 of 0.10 against 0.27; $r = 0.50$). Their format barriers are slightly lower, but not significantly ($p = 0.06$), and they do not differ in broken links. But they are twice as often redundant (17% against 8%), because several of them repeat files published elsewhere. HVD tags are added by publishers, so the comparison covers only the datasets that institutions chose to label.

Institution rankings are moderately to highly robust. Their correlation with the baseline ranking is 0.996 with staleness thresholds of 18 and 36 months, 0.98 when historical datasets are scored 0 on staleness or when probable redundancy is added, 0.96 with equal weights for all five indicators or with only definitive link failures, 0.84 when one-off year-specific datasets are treated as series and 0.81 with weights from principal component analysis. Dropping one indicator at a time gives correlations between 0.74 and 0.93, with staleness the most influential. The share of heavily polluted datasets is far less stable. It ranges from 3.5% to 12.7% across these specifications, and from 1.5% without staleness to 24.2% without broken links. That share should therefore be read as an order of magnitude; the prevalence of each form of pollution, the institution rankings and the cohort patterns are the more robust results.

## 6. Discussion

### 6.1. Nominal or effective transparency? (RQ1)

The growth of data.gov.ro has produced a large share of nominal transparency. Between 2013 and 2026 the catalogue grew almost seventy-fold, and in today's snapshot three datasets in five, from every publication cohort, show at least one severe form of pollution. Almost two thirds of the datasets that can be assessed for timeliness are stale, only one recurring series in three is still continued, and almost all files from the portal's first two years can no longer be opened. A citizen or journalist who looks for current data therefore meets, more often than not, an edition that stopped years ago. These findings are consistent with Heald's (2006) distinction between nominal and effective transparency: an expanding catalogue may coexist with persistent limitations in the accessibility, timeliness and usability of published data.

Pollution is widespread but mostly partial: about one dataset in sixteen is heavily polluted on several dimensions at once, and that share depends strongly on how the index is specified. The answer to the question in our title is therefore qualified. A single snapshot cannot show that transparency has declined, because it does not observe the past state of the catalogue. It does show that more data has not meant proportionally more usable data, because published datasets begin to age immediately and nothing removes them when they do.

The way progress has been measured reinforces the illusion. Romania's OGP commitments set targets in numbers of datasets, such as a minimum of 535 new or updated datasets in one action plan (OGP, n.d.-a). Targets of this kind reward publication and are blind to what happens to data afterwards.

### 6.2. The forms of pollution (RQ2)

The findings suggest that a substantial component of data pollution is associated with the persistence of outdated or discontinued datasets within the national portal. However, because the analysis relies on a single snapshot, it cannot establish whether these datasets were fully adequate at the time of their initial publication. The observed patterns nevertheless highlight the importance of data life-cycle management, including regular updates, a clear archival status and the identification of authoritative dataset versions. The portal itself makes this harder: it has no field for declaring an update frequency, so neither publishers nor users can tell an abandoned dataset from a historical one.

The other forms point to specific externalities. The links to the portal's original storage now return errors, yet the datasets that list them remain in the catalogue, so users still find data that no longer exist. The dominance of XLS and XLSX files, with CSV at 6.4% of resources, shifts the cost of conversion to every re-user. The absence of a contact address in seven datasets out of ten removes the feedback channel through which re-users could report errors, so pollution cannot be corrected from outside. Redundancy, including 316 datasets that repeat a file published in another dataset, multiplies search costs and creates doubt about which version is authoritative.

### 6.3. Declared versus real maturity (RQ3)

Romania's performance in international open data benchmarks presents a mixed picture. In the 2025 OECD OURdata Index, Romania scored above the OECD average overall (0.59 compared with 0.53), particularly in data accessibility and government support for re-use. However, its data availability score remained below the OECD average (0.41 compared with 0.53). Similarly, the European Open Data Maturity assessment

classified Romania among the "followers" in 2025, while recording substantially weaker results in the quality (71.4%) and impact (49.1%) dimensions (OECD, 2026; data.europa.eu, 2025). These assessments already identify important weaknesses, but primarily examine governance arrangements, publication practices and monitoring capabilities. The DPI complements them with direct evidence on the condition of published datasets: 60% exhibit at least one severe pollution indicator. Rather than contradicting existing benchmarks, the findings show how progress in open data governance can coexist with persistent barriers to effective data re-use.

The known-groups analysis provides a similarly qualified result. High-value datasets have a slightly lower mean DPI than other datasets (0.23 versus 0.26), but the overall difference is not statistically significant (one-sided $p = 0.07$; rank-biserial $r = 0.09$). At the indicator level, however, HVD-tagged datasets are substantially less stale and better documented, while showing higher levels of redundancy (17% versus 8%). This pattern is consistent with differences in publication and maintenance practices, but does not establish a causal effect of EU regulatory obligations. The findings suggest that stronger data life-cycle governance should combine requirements for timeliness and documentation with measures to prevent unnecessary duplication and maintain authoritative dataset versions.

### 6.4. Recycling: an illustration

Public procurement data provide an example of data recycling within the open government data ecosystem. The OpenTender platform collects, transforms and republishes procurement information from European and national sources, combining data from Tenders Electronic Daily (TED) with Romanian procurement datasets published on data.gov.ro (Open Contracting Partnership, n.d.). By integrating information from different sources and making it available in structured formats, the platform illustrates how administrative data can acquire additional value beyond their original publication context.

This example also highlights the importance of maintaining accessible, well-documented and non-redundant source data. Broken links, ambiguous dataset versions and format-related barriers can increase the effort required to integrate and verify information from multiple public sources. However, the present study does not assess OpenTender's processing activities or measure the costs incurred by its operators. The example therefore illustrates the potential consequences of data pollution for downstream re-use rather than providing direct empirical evidence of those costs.

### 6.5. Limitations

The analysis relies on metadata and on file-level checks, not on the content of the files. Inconsistency, the fourth form of pollution in our definition, is therefore not measured; a manual audit of a stratified sample, coded by two researchers, is the natural next step. Dates may understate the true age of some data: about 13,000 resources carry creation dates from September–December 2016, which suggests a platform migration, so staleness is a lower bound. Series are identified from titles, which may split or merge some series, and historical datasets are separated from continuous ones by the presence of a reference year in the title. Links were checked once, so some failures, such as 503 errors, may be temporary. The portal offers no usage data, so exposure could not be measured. The planned comparison with the European MQA scores could not be completed, so the distinction between pollution and metadata quality rests on the conceptual argument and on the low correlations between indicators. Finally, the DPI describes a single snapshot; the longitudinal analysis of vitality is left for future work.

## 7. Conclusions and policy recommendations

More data has not meant proportionally more usable data on data.gov.ro. The catalogue has grown steadily, but three datasets in five now show at least one severe form of pollution, mostly datasets and series that were published and then abandoned. Heavy pollution on several dimensions affects a much smaller and less certain share: about 6% in our baseline, and between 2% and 24% across specifications. The DPI makes this residue visible and comparable across institutions and publication cohorts. Its rankings are reasonably robust, but its validity as a measure distinct from metadata quality still has to be confirmed against the European MQA scores.

Four policy implications follow:

1. **Measure the stock, not the flow.** Open data targets in OGP action plans and in the National Open Data Strategy 2024–2028 should track the share of up-to-date, usable datasets alongside the number published.
2. **Govern the life cycle.** The portal should require an update frequency for each dataset, flag datasets that miss it, and let institutions mark editions as historical or retire them.
3. **Make feedback possible.** A contact point should be mandatory, so that re-users can report errors to the publisher.
4. **Prevent redundancy and format barriers at upload.** Simple checks, such as file hashes and a requirement to add CSV alongside XLS files, would stop part of the pollution at its source.

Future work should complete the ecological framework with a Data Ecosystem Vitality Index, follow the DPI through repeated snapshots, compare it with the European MQA scores and measure exposure through usage data where portals provide it.

**Acknowledgement:**
This article is based on research conducted for the project ”Receptivitate strategică și inovare la nivelul instituțiilor publice din România” (Strategic Receptivity and Innovation in Romanian Public Institutions), „C.S. Nicolăescu-Plopșor” Institute for Research in Social Sciences and Humanities.

**Authors’ Contributions:**
The authors contributed equally to this work.

**References:**

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1–52.

Ben-Shahar, O. (2019). Data pollution. *Journal of Legal Analysis*, 11, 104–159.

data.europa.eu (2020). *Open Data Maturity 2020: Romania country factsheet*. Luxembourg: Publications Office of the European Union. https://data.europa.eu/sites/default/files/country-factsheet_romania_2020.pdf

data.europa.eu (2024). *Open Data Maturity 2024: Romania country factsheet*. Luxembourg: Publications Office of the European Union.

https://data.europa.eu/sites/default/files/2025-06/2024_odm_factsheet_romania.pdf
data.europa.eu (2025). *Open Data Maturity 2025: Romania country factsheet*. Luxembourg: Publications Office of the European Union.
https://data.europa.eu/sites/default/files/2025-12/2025_odm_factsheet_romania.pdf
Diamantopoulos, A. and Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2), 269–277.
Fung, A., Graham, M. and Weil, D. (2007). *Full Disclosure: The Perils and Promise of Transparency*. Cambridge: Cambridge University Press.
Harrison, T. M., Pardo, T. A. and Cook, M. (2012). Creating open government ecosystems: A research and development agenda. *Future Internet*, 4(4), 900–928.
Heald, D. (2006). Varieties of transparency. In Hood, C. and Heald, D. (eds.), Transparency: The Key to Better Governance? Proceedings of the British Academy, vol. 135. Oxford: Oxford University Press, pp. 25–43.
https://doi.org/10.5871/bacad/9780197263839.003.0002
Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258–268.
Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, 32(3), 323–332.
Nardi, B. A. and O'Day, V. L. (1999). *Information Ecologies: Using Technology with Heart*. Cambridge, MA: MIT Press.
OECD (2026). *Digital Government Outlook 2026: Romania*. Country note. Paris: OECD Publishing.
https://www.oecd.org/en/publications/digital-government-outlook-2026_d46c0555-en/romania_858f5765-en.html
OECD and JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris: OECD Publishing.
OGP (n.d.-a). *Romania: Open Data (RO0065)*. Open Government Partnership. https://opengovpartnership.org/members/romania/commitments/RO0065
OGP (n.d.-b). *Romania: Increasing the Quality and Quantity of Published Open Data (RO0027)*. Open Government Partnership.
https://opengovpartnership.org/members/romania/commitments/RO0027
OGP (n.d.-c). *Romania: Implementation of Law on open data (RO0092)*. Open Government Partnership. https://www.opengovpartnership.org/members/romania/commitments/RO0092/
Open Contracting Partnership (n.d.). Romania: OpenTender. OCP Data Registry. Available at: https://data.open-contracting.org/en/publication/75
Raca, V., Velinov, G., Dzalev, S. and Kon-Popovska, M. (2022). A framework for evaluation and improvement of open government data quality: Application to the Western Balkans national open data portals. *SAGE Open*, 12. doi:10.1177/21582440221104813
Vetrò, A., Canova, L., Torchiano, M., Orozco Minotas, C., Iemma, R. and Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), 325–337.
Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33.

Wardle, C. and Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Strasbourg: Council of Europe.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

Zuiderwijk, A., Janssen, M. and Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, 19(1–2), 17–33.

*Legal sources:* Directive (EU) 2019/1024 on open data and the re-use of public sector information; Commission Implementing Regulation (EU) 2023/138 on high-value datasets; Law no. 109/2007; Government Emergency Ordinance no. 41/2016; Law no. 179/2022 on open data and the re-use of public sector information.

**Article Info**

***Received:*** November 01 2026
***Accepted:*** November 08 2026

**How to cite this article:**
Stăiculescu, O., Gheorghiță, V. (2026). More Data, Less Transparency? A Data Pollution Index for Romania's National Open Data Portal. *Revista de Științe Politice. Revue des Sciences Politiques,* no. 92, pp. 364 -379.