Metadata Quality Governance in a Data Lake: A Case Study of Lithuania's State Open Data Infrastructure

Tomas Danikauskas[1], Algirdas Šukys[1], Milda Aksamitauskas[2]

1 Kaunas University of Technology, Kaunas, Lithuania
2 University of Wisconsin, Madison, United States

Abstract

To comply with Directive (EU) 2019/1024 on the re-use of public sector data, EU member states are building national-level solutions—one of which is the national data lake. These systems are meant to generate value in both public and private sectors. In practice, though, their usefulness depends not just on whether data exist, but on whether the metadata describing those data is accurate and well managed. This paper examines metadata quality governance in Lithuania's state open data infrastructure, drawing on a metadata quality audit, a DCAT-AP compliance appendix, and related implementation materials. Five dimensions guide the analysis: completeness, consistency, uniqueness, semantic standardization, and governance-process adequacy. The audit covered 51,042 unique dataset-level records, 241,299 rows in the broader category-linked dataset context, and roughly 2.5 million dataset column records. The findings point to problems with identifiers, descriptions, keywords, naming conventions, and the alignment between metadata and actual data content. These are not cataloging mistakes in isolation — they are signs of deeper governance shortcomings in stewardship, vocabulary control, validation, and publication-readiness workflows. The paper argues that progress is achievable through stronger DCAT-AP-aligned governance, persistent identifiers, controlled vocabularies, SKOS-based concept schemes, and recurring quality assurance cycles. The broader contribution is an empirically grounded public-sector case showing that metadata quality is best treated as an operational governance capability—one that directly shapes interoperability and reuse.

Keywords

metadata governance, metadata quality, data lake, DCAT-AP

1. Introduction

National open data programmes are expected to deliver transparency, innovation, interoperability, and wider data reuse, but these outcomes do not follow automatically from datasets publication. They depend on the quality of the metadata used to describe the available datasets. Data lake approaches have grown more prominent as governments and public institutions work to store and manage increasingly large and heterogeneous collections. The literature is consistent on one point: without effective metadata management, data lakes risk becoming "data swamps"—repositories where information exists but cannot be found, understood, or reused [1–8]. The same risk applies to public-sector open data infrastructures, where discovery, comprehension, and dataset exchange all rely on well-structured, standardized descriptive metadata [9–17].

Lithuania offers a useful case for examining this issue. The state open data environment brings together a national portal, a broader state data management context, and a stated strategic interest in semantic interoperability and data reuse. Programme materials reference the portal data.gov.lt, the State Data Agency's metadata and data repository, and an EU-level interoperability context. Those materials also signal a clear intention to use DCAT-AP, controlled vocabularies, and semantic technologies for future reuse scenarios—including semantic search and ontology-based integration. Even so, the metadata quality audit underlying this paper reveals a number of practical weaknesses in the current metadata environment.

IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

tomas.danikauskas@ktu.lt (T. Danikauskas); algirdas.sukys@ktu.lt (A. Sukys); aksamitauska@wisc.edu (M. Aksamitauskas)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper17.pdf

The paper addresses three research questions: Which metadata quality problems are most significant in Lithuania's state open data and data lake metadata environment? In what ways do these problems reflect broader metadata governance weaknesses in a public-sector infrastructure? How might DCAT-AP-aligned governance, controlled vocabularies, and semantic standardization improve metadata quality and interoperability?

The paper makes three main contributions. First, it treats metadata quality governance as an important link between data lake metadata management and public-sector open data governance. Second, it presents an empirically grounded Lithuanian case build on a real metadata quality audit and associated implementation materials. Third, it proposes a compact governance improvement model that may be relevant not only for Lithuania, but to other national open data and data lake environments facing similar challenges.

2. Related Work and Conceptual Framing

Metadata has long been recognized as a important component of data lake architectures. Giebler et al. emphasize that comprehensive data lake architecture must address multiple interacting architectural concerns, including metadata management [1]. Sawadogo and Darmont likewise show that data lake architectures and metadata management are closely related, while also noting that operationally mature solutions remain limited [2]. Similarly, Eichler et al. argue for generic metadata models capable of supporting multiple metadata-related use cases rather than fragmented, ad hoc implementations [3].

Other researchers focus on more specific metadata functions within data lakes. Alserafi et al. treat information profiling as a tool for managing content metadata in data lake settings [4]. Zhao et al. extend this perspective by introducing analysis-oriented metadata that can be reused across ingestion, transformation, and analytical stages [5]. Halevy et al., in their description of Google's GOODS system, demonstrate how metadata at scale becomes essential for organizing datasets and enabling discoverability [6]. Langenecker et al. highlight the possible role of automation in metadata extraction [7], while Quix et al. address the same challenge through the GEMMS approach [8]. These studies collectively demonstrate that metadata is not just description layer—it is part of the operating logic of the data lake itself.

Research on open data reaches a very similar conclusion from a different direction. Quarati identifies metadata quality as one of the major factors shaping actual use of open public data [9], while Schauppenlehner and Muhar pointing to metadata management as a decisive factor in open data portals [10]. Neumaier et al. show that automated assessment of metadata quality across portals is both feasible and necessary, particularly because many metadata problems recur systematically rather than appearing as isolated exceptions [11]. Haberl et al. further broaden the governance perspective by identifying key governance requirements in open data ecosystems, including data usability, lineage, access conditions, and organizational design [12]. Their contribution is especially relevant here because it connects metadata quality and standardization with broader platform-level governance capabilities.

In this paper, metadata quality governance is understood as the combination of roles, policies, standards, controlled vocabularies, validation rules, and review processes through which metadata is kept complete, consistent, uniquely identifiable, semantically meaningful, and reusable across systems. In formal terms, DCAT and DCAT-AP provide the descriptive model needed for interoperable data catalogues [13,14], while SKOS offers a practical mechanism for representing controlled vocabularies and concept schemes [15]. From an organizational perspective, DAMA-DMBOK remains a useful reference because it links metadata management and data quality within a broader governance framework rather than treating them as separate technical concerns [18]. More recent portal studies reinforce the same point: even comparatively mature open data portals often struggle to maintain completeness and consistency against their own publishing rules, which suggests that metadata quality control must be institutionalized as an ongoing governance function rather than handled as a one-time compliance exercise [20].

The gap addressed by this paper is therefore not the absence of metadata standards or conceptual models. Those already exist. Just few studies examine metadata quality issues as governance problems within a national public-sector data lake or open data environment. Much of the literature stays conceptual, architecture-driven, or focused on single organizations. The Lithuanian case offers a concrete,

policy-relevant setting where audit findings can be tied directly to interoperability requirements and implementation practice.

3. Research Context and Methodology

This study is situated in Lithuania's state open data infrastructure and its wider state data lake context. The materials underlying the case show that metadata management is expected to support dataset discovery, lifecycle control, and, in the longer term, semantic search. They also point to the use of standards such as DCAT-AP, ISO 19115, and Dublin Core. Lithuania has additionally formalized a national metadata profile through DCAT-AP-LT, an adaptation of DCAT-AP for Lithuanian datasets and services [19]. This is important because it indicates that the key problem is not the lack of an available standard. Rather, the challenge lies in achieving consistent operational uptake of that standard across public institutions and publication workflows. It is also relevant that metadata maturity varies considerably across institutions. In some cases, data are still handled through paper-based or spreadsheet-centric practices, and many information systems were originally developed without an explicit data strategy.

Methodologically, the paper adopts an interpretive case study design based on structured documentary evidence. The aim is not simply to restate the audit, but to reinterpret its findings through the lens of metadata governance. Five analytical dimensions guide the analysis: completeness, consistency, uniqueness, semantic standardization, and governance-process adequacy. Completeness refers to whether required metadata elements are present. Consistency concerns uniformity in naming, syntax, and classification practices. Uniqueness relates to stable identifiers and duplication control. Semantic standardization addresses the use of controlled vocabularies and explicit metadata models. Governance-process adequacy concerns the existence of roles, rules, validation workflows, and recurring review mechanisms.

The dataset audit covered 51,042 unique records at dataset level. Identifier- and category-related checks were performed on a broader set of 241,299 rows, and the column-level audit covered approximately 2.5 million entries. These volumes provide a sufficiently broad empirical basis for drawing governance-level conclusions from repeated patterns rather than from isolated anomalies.

4. Results

The audit findings suggest that Lithuania's metadata environment already has several important building blocks of a national metadata infrastructure. At the same time, those elements do not yet operate with enough governance discipline to ensure coherent metadata quality at scale.

4.1. Metadata quality problems

Some of the clearest problems concern missing values, duplication, naming inconsistency, and weakly controlled keywords. The audit identified 2,091 dataset descriptions recorded as NULL or "nan", which points to a noticeable completeness problem in one of the most important descriptive fields for dataset understanding and discovery.

Consistency issues are also widespread. The audit points to irregular use of uppercase and lowercase letters, punctuation, quotation marks, abbreviations, and institutional naming formats. Examples include variation between full legal names and abbreviated institutional forms, as well as semantically overlapping labels such as "Medicines," "Drugs," and "Pharmaceuticals." These inconsistencies are not merely cosmetic. In a metadata environment intended to support search, filtering, joining, and semantic interoperability, non-standard naming directly reduces retrieval quality and makes downstream harmonization more difficult.

Keyword quality is another weak area. The materials report 14,496 unique keywords, many of them redundant, irrelevant, or excessively long, with some strings exceeding 300 characters. This strongly suggests that keyword fields have not been governed through adequate input constraints, vocabulary control, or publication-readiness checks. From a governance perspective, the keyword field appears to have functioned largely as an uncontrolled free-text container rather than as a managed semantic access

point. The fact that keyword generation was partly automated may help explain this outcome, but it does not remove the need for quality control.

The audit also reports 2,173 datasets declared with zero columns and 11,222 datasets with zero records. Such values do not necessarily mean that every one of these datasets is invalid. However, they do indicate weak alignment between catalog metadata, structural metadata, and the actual publication status of data resources. When metadata is expected to guide users toward reusable data, these discrepancies increase uncertainty and raise the cost of interpretation.

4.2. Governance weaknesses revealed by the audit

Taken together, these quality defects point to broader governance issues. The first concerns identifier governance. The audit recommends introducing a unique and persistent identifier for each dataset and also suggests that such identifiers could form part of a stable URI strategy. This is not a minor technical refinement. Persistent identifiers are a basic requirement for traceability, stable referencing, and cross-system interoperability.

A second weakness concerns vocabulary governance. The audit materials recommend the standardized use of fields for access rights, language, publication date, update frequency, geographic coverage, and high-value dataset categories, often in ways that map directly to DCAT-AP properties and controlled values. They also recommend controlled vocabularies for themes, access rights, licenses, and frequencies, together with SKOS-based concept schemes for taxonomy and thesaurus management. The fact that such recommendations are necessary indicates that these semantic controls either do not yet exist in a sufficiently standardized form or are not being applied consistently in practice.

A third weakness lies in stewardship and review procedures. The materials recommend clearer governance policies, periodic reviews, deduplication procedures, improved description practices, and stronger quality controls. This implies that routine governance mechanisms of this kind are still not fully embedded in day-to-day publication practice.

4.3. Summary of findings

Overall, the Lithuanian case shows that metadata quality problems accumulate across four connected layers: record-level defects, field-level standardization gaps, vocabulary-level semantic fragmentation, and process-level governance weaknesses. Table 1 summarizes the main issues and their implications.

Table 1
Main metadata quality findings and their governance implications.

Dimension	Observed issue	Governance implication
Completeness	2,091 descriptions marked as NULL or “nan”.	Mandatory-field enforcement and publication readiness checks are insufficient.
Uniqueness	Approx. 43,929 repeated dataset identifiers in the wider dataset context.	Persistent identifier policy and deduplication rules are missing or weak.
Consistency	Institution names, abbreviations, punctuation, and casing vary.	Naming conventions and normalization controls are not consistently enforced.
Semantic control	14,496 unique keywords include redundant, irrelevant, and overlong values.	Keyword fields are not governed through controlled vocabularies or semantic review.
Catalog–data coherence	Some datasets are declared with zero columns or zero records.	Metadata publication is weakly synchronized with structural and operational data status.

5. Discussion

The Lithuanian case aligns with the broader literature in several important respects. First, it confirms that metadata is an architectural enabler of the data lake rather than a secondary catalog add-on. This is

consistent with studies that view metadata management as a necessary condition for discoverability, traceability, reuse, and the long-term operability of data lake environments [1–8]. Where persistent identifiers, controlled fields, and semantically coherent descriptions are weak, the catalogue ceases to function as an effective interface to the underlying data assets.

Second, the case shows that metadata quality in public-sector infrastructures needs to be understood at multiple levels. The visible problems begin at record level, for example in the form of NULL descriptions, repeated identifiers, and poorly controlled keyword values. Yet these are only the surface manifestations of deeper issues at field, vocabulary, and workflow level. In that sense, the Lithuanian evidence makes concrete a point that is often stated more abstractly in the literature: a data lake becomes swamp-like not only because data are heterogeneous, but because governance arrangements are too weak to preserve meaning, trust, and usability over time [2,3,17].

Third, the case strengthens the link between open data governance and semantic interoperability. Public-sector metadata is not created only for internal cataloging convenience. It must also support publication, exchange, and reuse across institutional and technical boundaries. This makes DCAT-AP alignment, vocabulary governance, and concept-scheme management strategically important rather than optional descriptive refinements [13–17,20]. In other words, the governance challenge is dual: institutions must maintain both technical metadata discipline and semantic coherence.

At the same time, the case adds a useful empirical nuance to the literature. A number of studies emphasize the importance of governance but leave it relatively abstract [2,3,12]. The Lithuanian material shows more concretely what governance operationalization looks like in practice: identifier rules, mandatory-field policies, controlled vocabularies, validation routines, stewardship assignments, review cycles, and publication-readiness controls. This is significant because one-off cleanup efforts may improve metadata temporarily, but without institutionalized controls the same issues are likely to recur.

More broadly, the case suggests that metadata quality governance should be understood as a public-sector interoperability capability. Lithuania's metadata environment connects a national catalogue, a state data repository, and a broader EU exchange context. In such an environment, local metadata weaknesses do not remain local. They propagate outward into broader discovery, comparison, and reuse problems. By the same logic, improvements in metadata governance can produce cumulative gains for findability, comparability, machine processing, and cross-portal reuse. For that reason, metadata quality should be governed as a continuing infrastructure function rather than treated as a narrow catalog maintenance task.

6. Recommendations

The case makes it possible to outline a compact governance model for national data lake metadata.

First, every dataset, distribution, and relevant organizational entity should have a stable identifier that remains valid across publication cycles, category assignments, and system migrations. Without this, deduplication, traceability, and URI-based interoperability remain fragile.

Second, DCAT-AP should be used not only as an external publication target, but also as an internal governance profile. At a minimum, mandatory fields should include identifier, title, description, publisher, theme, access rights, language, update frequency, and distribution references. Domain-specific extensions may then be added where justified.

Third, controlled vocabularies should be governed explicitly rather than left to diffuse institutional practice. A national metadata governance function should maintain approved term lists, concept schemes, change procedures, and mappings between legacy labels and preferred terms.

Fourth, validation should combine automated and human controls. Automated validation can detect missing mandatory fields, malformed dates, identifier collisions, uncontrolled free text in controlled fields, punctuation irregularities, overlong keywords, and duplication patterns. Human review remains necessary for semantic adequacy, terminology quality, and cross-institutional consistency.

Fifth, stewardship responsibilities should be assigned clearly. Metadata quality tends to improve when named stewards, data owners, and catalog managers are accountable for the completeness and semantic coherence of the metadata they publish. This emphasis is also consistent with the open data governance literature, particularly where organizational design and usability are concerned [12].

Sixth, recurring audit cycles should be institutionalized. For example, weekly automated checks could be complemented by monthly reviews of vocabulary use and missing fields, together with annual semantic audits. Most importantly, these checks should be integrated into ordinary publication workflows and monitored through simple quality dashboards.

These recommendations are realistic in technical and organizational terms because they build on standards and governance practices that are already visible in the Lithuanian materials and are well aligned with the wider metadata and open data literature.

7. Conclusion

This paper examined metadata quality governance in Lithuania's state open data infrastructure through an interpretive case study grounded in a metadata audit. The findings suggest that recurring issues such as missing descriptions, inconsistent naming, and weak keyword control are better understood as governance signals than as isolated data-cleaning defects. Taken together, they point to insufficient field standardization, incomplete vocabulary governance, weak validation routines, and limited institutionalization of stewardship and review.

The paper contributes in three ways. From a theoretical perspective, it positions metadata quality governance at the intersection of data lake metadata management, semantic interoperability, and public-sector open data governance. Empirically, it documents a Lithuania-based case in which audit findings can be linked to specific governance design choices and implementation gaps. Practically, it proposes a governance model centered on persistent identifiers, DCAT-AP-aligned field rules, controlled vocabularies, machine-assisted validation, and recurring quality assurance cycles.

The study also has limitations. It is based on a single national case and relies on project documentation rather than longitudinal observation of governance change over time. The findings should therefore be read as an analytically grounded case interpretation rather than as a statistical generalization. Future research could compare multiple national portals, assess improvements after governance reforms using repeated quality metrics, and test machine-assisted validation methods, including SHACL-based conformance checking in DCAT-AP publication workflows.

8. Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, B. Mitschang, The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture, in: Proceedings der 19. Fachtagung für Datenbanksysteme für Business, Technologie und Web (BTW 2021), 2021, pp. 351–370. DOI: 10.18420/btw2021-19.

[2] P.N. Sawadogo, J. Darmont, On data lake architectures and metadata management, Journal of Intelligent Information Systems 56 (2021) 97–120. doi:10.1007/s10844-020-00608-7.

[3] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, B. Mitschang, Modeling metadata in data lakes—A generic model, Data & Knowledge Engineering 136 (2021) 101931. doi:10.1016/j.datak.2021.101931.

[4] A. Alserafi, A. Abelló, O. Romero, T. Calders, Towards Information Profiling: Data Lake Content Metadata Management, in: 2016 IEEE 16th International Conference on Data Mining Workshops, 2016, pp. 178–185. doi:10.1109/ICDMW.2016.0033.

[5] Y. Zhao, J. Aligon, G. Ferrettini, I. Megdiche, F. Ravat, C. Soulé-Dupuy, Analysis-oriented Metadata for Data Lakes, in: 25th International Database Engineering & Applications Symposium (IDEAS 2021), ACM, 2021. doi:10.1145/3472163.3472273.

[6] A. Halevy, F. Korn, N.F. Noy, C. Olston, N. Polyzotis, S. Roy, S.E. Whang, GOODS: Organizing Google's Datasets, in: Proceedings of the 2016 International Conference on Management of Data (SIGMOD 2016), 2016, pp. 795–806. doi:10.1145/2882903.2903730.

[7] S. Langenecker, C. Sturm, C. Schalles, C. Binnig, Towards Learned Metadata Extraction for Data Lakes, in: BTW 2021, 2021, pp. 325–336. doi:10.18420/btw2021-17.

[8] C. Quix, R. Hai, I. Vatov, Metadata extraction and management in data lakes with GEMMS, Complex Systems Informatics and Modeling Quarterly 9 (2016) 67–83. doi:10.7250/csimq.2016-9.04.
[9] A. Quarati, Open Government Data: Usage trends and metadata quality, Journal of Information Science 49 (2023) 1–24. doi:10.1177/01655515211027775.
[10] T. Schauppenlehner, A. Muhar, Theoretical Availability versus Practical Accessibility: The Critical Role of Metadata Management in Open Data Portals, Sustainability 10 (2018) 545. doi:10.3390/su10020545.
[11] S. Neumaier, J. Umbrich, A. Polleres, Automated Quality Assessment of Metadata across Open Data Portals, Journal of Data and Information Quality 8 (2016) 1–29. doi:10.1145/2964909.
[12] A. Haberl, C.D. Malin, S. Thalmann, A Framework to Identify Data Governance Requirements in Open Data Ecosystems, in: 35th Bled eConference, University of Maribor Press, 2022, pp. 360–375.
[13] SEMIC, DCAT-AP 3.0.1, Interoperable Europe / SEMIC, 2025. URL: https://semiceu.github.io/DCAT-AP/releases/3.0.1/.
[14] W3C, Data Catalog Vocabulary (DCAT) - Version 3, W3C Recommendation, 2024. URL: https://www.w3.org/TR/vocab-dcat-3/.
[15] W3C, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, 2009. URL: https://www.w3.org/TR/skos-reference/.
[16] European Union, Directive (EU) 2019/1024 on open data and the re-use of public sector information, Official Journal of the European Union, 2019. URL: https://eur-lex.europa.eu/eli/dir/2019/1024/oj/eng.
[17] European Union, Regulation (EU) 2022/868 on European data governance, Official Journal of the European Union, 2022. URL: https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng.
[18] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge, 2nd ed., Technics Publications, Basking Ridge, NJ, 2017.
[19] VSSA, DCAT-AP-LT - Version 2, Technical Specification, 22 Oct. 2024. URL: https://ivpk.github.io/DCAT-AP-LT/en/.
[20] B. Šlibar, E. Mu, OGD metadata country portal publishing guidelines compliance: A multi-case study search for completeness and consistency, Government Information Quarterly 39(4) (2022) 101756. doi:10.1016/j.giq.2022.101756.