Are the Provided Keywords on Open Data Portals Semantically Complete and Relevant? An Exploratory Study*

Deperias Kerre[1,2,*], Christophe Debruyne[1]

[1]Montefiore Institute of Electrical Engineering and Computer Science, University of Liège, Belgium
[2]SCES, Strathmore University, Nairobi, Kenya

Abstract

Open data portals maintain collections of datasets and metadata that support dataset discovery. Descriptive metadata, such as titles, descriptions, and keywords, summarize the underlying data and facilitate metadata-driven discovery. Keywords enable indexing, filtering, and thematic organization of datasets, yet their quality in relation to descriptive metadata remains largely unexplored. This work introduces the notion of *Semantic Keyword Quality (SKQ)*, comprising *Semantic Relevance* and *Semantic Completeness*. We propose an initial framework for analyzing the semantic quality of keyword annotations as a component of metadata quality. Our findings show that keywords exhibit low-to-moderate semantic relevance with respect to titles and descriptions (mean aggregated similarity ≈ 0.42), while semantic completeness remains limited (title mean = 0.33; description mean = 0.28). The results indicate that keyword sets are generally semantically related to descriptive metadata but do not fully represent the concepts expressed in titles and descriptions, particularly within richer descriptive text. These findings highlight the importance of evaluating semantic keyword quality in metadata-driven discovery environments and provide a foundation for future work on keyword enrichment and its implications for dataset discoverability.

Keywords

Data Discoverability, Metadata Quality, Metadata Enrichment, Open Data Portals, Semantic Keyword Quality

1. Introduction

Open data portals, such as the European Data Portal and data.gov.uk, contain large volumes of data collections that hold potential for data-driven decision making and analytics. However, as the volume of data increases, so does the need to make this data discoverable. The discoverability of data is based on the quality of the metadata that describes it [1, 2]. From a data governance perspective, metadata quality constitutes a critical control point for ensuring data usability, accessibility, and interoperability within data ecosystems [3] .

The role of discoverability is emphasized by the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, which require data to be findable through rich and well-structured metadata [4]. In practice, metadata fields such as titles, descriptions, and keywords play a crucial role in enabling this process. In particular, keywords support metadata-based dataset discovery by organizing the data into categories, such as themes, topics, and locations.

Metadata standards such as the W3C Data Catalog Vocabulary (DCAT)[1] are widely adopted across open data portals and are used to define structured descriptions for datasets, including fields for titles, descriptions, and keywords [5]. Despite the adoption of these standards, datasets' metadata on open data portals, as noted by [6], still exhibit inconsistencies and other metadata quality issues, such as incompleteness. It is thus reasonable to assume that these issues can have a negative impact on data discoverability, highlighting the need for richer metadata.

NeXt-generation Data Governance workshop 2026 (NXDG 2026), co-located with SEMANTiCS 2026: 22nd International Conference on Semantic Systems, 15-17 September, 2026, Ghent, Belgium

*Corresponding author.
deperias.kerre@uliege.be (D. Kerre); c.debruyne@uliege.be (C. Debruyne)
0000-0002-7437-6735 (D. Kerre); 0000-0003-4734-3847 (C. Debruyne)

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

[1]https://www.w3.org/TR/vocab-dcat-2/

In open data portals, dataset content is typically not directly accessible or processable at scale. Consequently, descriptive metadata (e.g., titles, keywords, and descriptions) are supposed to serve as a proxy for dataset content, as it is intended to summarize and represent the underlying data. Keywords further support important discovery mechanisms such as filtering, metadata search, faceted navigation, and thematic browsing [7]. Since dataset discovery is largely mediated through metadata representations rather than raw dataset contents, we argue that evaluating semantic keyword quality in relation to descriptive metadata becomes essential for determining whether keyword annotations adequately reflect the semantic content expressed in descriptive metadata.

Prior studies have established that metadata quality plays a significant role in dataset discoverability, with keyword annotations contributing to dataset findability and search effectiveness, but with little emphasis on alignment evaluation in relation to other descriptive dataset metadata used for dataset discovery [6, 8]. There is therefore a need to adapt metadata quality analysis to keyword annotations and analyze whether assigned keywords adequately align with the descriptive metadata fields used during dataset discovery.

In this work, we introduce the notion of *Keyword Semantic Quality (SKQ)* to analyze the semantic alignment between keywords and descriptive metadata used for dataset discovery. This is operationalized into *semantic relevance* and *semantic completeness*. This work contributes to data governance by adapting metadata quality perspectives to analyze SKQ as a component of dataset discoverability and the FAIR *findability* principle. The study, therefore, focuses on the following two main research questions:

- RQ1: How can established metadata quality dimensions be operationalized for analyzing keyword annotations in open data portals?
- RQ2: Do the keyword annotations on open data portals adequately semantically represent the concepts expressed in descriptive metadata?

The main contributions of this paper are therefore as outlined below.

- **Operationalizing Semantic Keyword Quality (SKQ):** Building on established metadata quality dimensions such as accuracy and completeness, we adapt and formalize these concepts for keyword annotations as *semantic relevance* and *semantic completeness*, enabling the analysis of how well keywords align with and represent the concepts expressed in other descriptive metadata fields.
- **Analysis of Semantic Keyword Quality.** We conduct an experimental analysis showing that keywords exhibit *low-to-moderate semantic alignment* and often *limited semantic completeness* in relation to the dataset descriptive metadata, highlighting variability in metadata quality across the evaluated dataset sample on the European Open Data portal.

This work represents an exploratory initial step toward broader research on semantic keyword quality in metadata-driven dataset discovery. It focuses on the semantic alignment and representational completeness of keyword annotations relative to descriptive metadata, examining how well assigned keywords semantically represent information expressed in dataset titles and descriptions rather than the underlying dataset contents. This reflects the metadata-centric nature of dataset discovery in open data portals. Before investigating the impact of semantic keyword quality on retrieval performance and discoverability, it is necessary to establish how the concept can be formally defined and measured. Accordingly, this study operationalizes semantic keyword quality as a measurable metadata quality construct, while its implications for discoverability and retrieval effectiveness are left for future work.

The remainder of this paper is organized as follows: Section 2 reviews related work on metadata quality and dataset discoverability, the semantic keyword quality anbalysis approach is introduced in Section 3, and Section 4 describes the experimental setup. Section 5 presents and analyzes the experimental results, and finally, Section 6 concludes the paper and outlines future research directions.

2. Metadata Quality and Data Discoverability

Discoverability refers to the degree to which datasets can be located and accessed by users through search or browsing mechanisms [9]. In open data portals, discoverability is largely determined by the quality and structure of metadata representations, which serve as the primary interface for indexing and retrieval [7, 10]. Several dimensions for analysis metadata quality have been proposed in the literature [11]. These include completeness, weighted completeness[2], accuracy, richness of information, and accessibility.

Prior work has examined metadata quality and dataset discovery from different perspectives. Ahmed et al. [6] investigated automatic keyword generation for open datasets and showed that poor or missing keywords negatively affect dataset findability, but did not analyze the semantic alignment of keywords with descriptive metadata. Yang [8] analyzed metadata field contributions to discoverability and found that titles and descriptions strongly influence search engine discovery, without evaluating keyword quality relative to these fields. Similarly, Gan et al. [12] studied metadata importance for conversational and natural-language dataset retrieval, demonstrating the strong influence of descriptive metadata, particularly descriptions, on retrieval performance. However, while these studies demonstrate the importance of keywords and descriptive metadata for dataset discovery, they do not analyze whether assigned keywords semantically represent the information expressed in titles and descriptions. Understanding this relationship is important because semantically inadequate keywords may limit metadata-driven dataset discoverability by failing to represent information contained in descriptive metadata.

Several frameworks have also been proposed for evaluating metadata quality based on metrics such as accuracy, consistency and readability [10, 13]. Another study proposes a metadata quality evaluation framework using ISO-based quality dimensions [7]. Neumaier et al. [14] further proposed an automated metadata quality analysis framework for Open Data portals based on DCAT metadata mappings and quality dimensions such as existence, conformance, retrievability, accuracy, and openness. These frameworks mainly evaluate structural and catalog-level dimensions such as completeness, consistency, accessibility, interoperability, and metadata field existence, while the semantic quality of keyword annotations in relation to descriptive metadata fields remains largely unexplored.

In summary, prior studies have established the importance of metadata in general and keywords in particular for dataset discoverability, while proposed metadata quality frameworks primarily focus on structural, catalog-level, or retrieval-oriented evaluation. However, the semantic quality of keyword annotations themselves, particularly in relation to the descriptive metadata that serves as the primary representation of datasets in open data portals, remains largely unexplored. This work addresses this gap by evaluating whether keywords adequately align with and semantically represent the content in titles and descriptions through the dimensions of semantic relevance and semantic completeness. By focusing on the metadata layer, our approach directly reflects how datasets are represented and discovered in practice.

3. Operationalizing Semantic Keyword Quality Analysis

Existing metadata quality frameworks commonly define metadata quality through dimensions such as completeness, accuracy, consistency, accessibility, and interoperability [10, 11, 13]. These dimensions primarily analyze whether metadata is present, structurally correct, and operationally usable across catalogs and data portals. However, such definitions do not explicitly capture whether keyword annotations semantically align with the descriptive metadata used during dataset discovery. There is therefore a need to establish how semantic keyword quality can be formally defined and measured, before analyzing its impact on data discovery.

[2]Completeness measures the proportion of metadata fields that are populated, treating all fields as equally important, whereas weighted completeness assigns different importance weights to metadata fields and evaluates completeness based on the presence of more critical metadata elements.

Our adaptation of metadata quality dimensions to semantic keyword quality analysis is motivated by two cases:

- *Keywords as Thematic Descriptors:* Prior work on Folksonomy and Information Retrieval observes that "documents are linked thematically with each other when they have been indexed with the same tags" [15], indicating that tags act as thematic descriptors connecting semantically related resources. Extending this intuition to dataset metadata, we argue that keyword annotations should be semantically aligned with the descriptive metadata they accompany and should adequately represent the semantic content expressed therein. This is vital since dataset discovery is metadata-driven, and keywords contribute to how datasets are filtered and discovered.
- *Metadata-driven dataset discovery in open data portals:* Datasets are typically discovered through metadata rather than direct access to their contents, making keywords important semantic descriptors of information expressed in descriptive metadata such as titles and descriptions. If assigned keywords are misaligned with this metadata, relevant datasets may be overlooked because important aspects of their descriptions are not adequately represented. Evaluating semantic keyword quality therefore requires analyzing the extent to which assigned keywords align with and semantically represent the content expressed in descriptive metadata.

In this work, we treat *semantic keyword quality* as a component of metadata quality, focusing on how established dimensions such as accuracy and completeness can be interpreted in the context of keyword annotations. This focus is important because keyword annotations constitute a key metadata element used to support dataset findability and discovery. Accordingly, we operationalize these metadata quality dimensions for keyword evaluation as *semantic relevance* and *semantic completeness*.

3.1. Semantic Relevance

Metadata accuracy refers to the extent to which metadata correctly and meaningfully represents the underlying dataset [7, 11]. Building on this, we extend the notion of metadata accuracy to *semantic relevance*, operationalized as the *degree to which keyword annotations semantically align* with the descriptive metadata they are associated with. Thus, metadata accuracy is reflected not only in correctness but also in semantic consistency between keyword annotations and descriptive metadata. Semantic relevance, therefore, evaluates this semantic alignment between keywords and dataset descriptive metadata. We operationalize semantic relevance through the following metrics.

3.1.1. Aggregated Semantic Relevance:

The aggregated semantic relevance measures the *overall semantic alignment* between the entire keyword set and a descriptive metadata field associated with a dataset. This reflects practical dataset discovery scenarios in open data portals, where users may formulate search queries using combinations of terms rather than relying solely on individual keywords. In such cases, metadata-driven discovery may depend not only on the relevance of isolated keywords, but also on the extent to which the complete keyword set collectively represents the semantic content of the dataset metadata.

Let $K_d = \{k_1, \dots, k_n\}$ be the set of keywords assigned to dataset d, and let $M_d \in \{t_d, desc_d\}$ represent the associated descriptive metadata, where M_d consist of the dataset title or description. Let $f(\cdot)$ denote the embedding function, and let $g(\cdot)$ denote an aggregation function that combines the semantic representations of all keywords into a collective representation suitable for comparison with descriptive metadata. The aggregated relevance is defined as:

$$Rel_{\text{agg}}(K_d, M_d) = \cos\left(g(\{f(k_1), \dots, f(k_n)\}), f(M_d)\right) \tag{1}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. The aggregation function $g(\cdot)$ may be instantiated using different strategies for combining text embeddings into a unified semantic representation, such as mean pooling, max pooling, or weighted aggregation. Since keywords in metadata constitute an unordered

set of descriptors rather than coherent natural language text, aggregating at the embedding level avoids imposing artificial syntactic structure that may arise from text-based concatenation [16].

In this work, we therefore adopt mean pooling, which computes the centroid representation of the keyword set and provides a simple, widely used mechanism for deriving fixed-size semantic representations suitable for similarity comparison [16, 17]. This is computed as shown in Equation 2 below.

$$g(K_d) = \frac{1}{|K_d|} \sum_{k \in K_d} f(k) \tag{2}$$

This is computed separately for titles and descriptions across all dataset metadata and later analyzed collectively to examine the various distributions for this metric. High aggregated semantic relevance implies that keywords collectively represent the metadata field well, while a low value indicates global semantic misalignment with the descriptive metadata. Since mean pooling may be sensitive to noisy or semantically unrelated keywords, investigating alternative pooling and aggregation strategies is left for future work.

3.1.2. Average Keyword-Level Relevance:

This measures the average semantic alignment of individual keywords with a metadata field. It is determined by computing the similarity between each keyword and the metadata field, then taking the average.

Let $d \in D$ denote a dataset, $K_d = \{k_1, \dots, k_n\}$ the set of keywords assigned to d, and $|K_d|$ the number of keywords. Each $k \in K_d$ represents an individual keyword, and $M_d \in \{t_d, desc_d\}$ denotes the dataset metadata, where M_d is either the dataset title or description. The formal definition of average keyword relevance is as given in Equation 3 below.

$$Rel_{\text{avg}}(K_d, M_d) = \frac{1}{|K_d|} \sum_{k \in K_d} \cos\big(f(k), f(M_d)\big) \tag{3}$$

A high average keyword-level relevance implies that most keywords are individually relevant, while a low one implies the presence of noisy or unrelated keywords.

We adopt cosine similarity as the measure of semantic alignment between keyword and metadata representations for both aggregated and average keyword-level relevance. Since cosine similarity ranges from -1 to 1, small negative values may occur when keyword and metadata embeddings are weakly misaligned in the embedding space. In the context of this study, such values are interpreted as weak semantic mismatch rather than strong semantic opposition.

3.2. Semantic Completeness

Metadata completeness refers to the extent to which metadata provides sufficient coverage of the characteristics of the resource/data being described [7, 11, 14]. Building on this, we extend the notion of metadata completeness to *semantic completeness*, operationalized as the *degree to which keyword annotations semantically represent the concepts expressed in a metadata field.* Semantic completeness therefore measures the extent to which the concepts expressed in a metadata field are semantically represented by the assigned keywords.

Let $d \in D$ denote a dataset and $M_d \in \{t_d, desc_d\}$ represent the associated descriptive metadata, where M_d is either the dataset title or description. Let $C_d = \{c_1, \dots, c_m\}$ denote the set of concepts expressed in metadata field M_d, and let $K_d = \{k_1, \dots, k_n\}$ denote the set of assigned keywords. Ideally, semantic completeness should analyze the extent to which the assigned keywords semantically represent the concepts in C_d. However, since these concepts are not directly observable at scale, we operationalize them using a set of metadata-derived keywords $K'_{M_d} = \{k'_1, \dots, k'_m\}$ extracted from M_d, which serves as a proxy representation of C_d. Consequently, the proposed metric should be interpreted as an approximation of semantic completeness rather than an absolute measure of concept coverage. Semantic completeness is

analyzed separately for titles and descriptions to enable an independent evaluation of each metadata field. The semantic completeness metric can then be computed as shown in Equation 4.

$$SemComp(K_d, M_d) = \frac{1}{|K'_{M_d}|} \sum_{k' \in K'_{M_d}} \mathbb{I}\left(\max_{k \in K_d} \cos\left(f(k'), f(k)\right) \geq \tau\right) \tag{4}$$

The operator $\max_{k \in K_d}$ selects the highest semantic similarity between a metadata-derived keyword k' and any assigned keyword in K_d, while $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the similarity exceeds a threshold τ, and 0 otherwise. Under this operationalization, semantic completeness estimates the extent to which assigned keywords represent concepts expressed in the descriptive metadata through the proxy set K'_{M_d}.

We adopt cosine similarity as the similarity measure for both semantic relevance and semantic completeness, as it is a standard metric in the vector space model for comparing textual representations [17], and is commonly employed in modern embedding-based approaches to capture semantic similarity between text representations [18]. Together, these metrics capture semantic relevance and semantic completeness of keyword annotations in relation to descriptive metadata.

4. Experimental Setup

In open data portals, descriptive metadata serve as the primary textual representations used for indexing and retrieval. Access to raw dataset content is often heterogeneous or limited. In our experimental setup, we treat descriptive metadata (title and description) as a proxy for dataset content, as this metadata is intended to summarize the underlying data. This assumption is consistent with practical dataset discovery workflows in open data portals, where users primarily interact with metadata representations rather than raw dataset contents during search and browsing.

4.1. Data Source

In this work, the European Open Data Portal is used as a case study. We downloaded the metadata dump available from the portal[3] and extracted 23,777 datasets from the *data.gov.uk* collection, which was selected due to its comparatively large number of datasets. From this collection, we randomly sampled 10,000 datasets and retained the dataset ID, title, description, and keywords. The analysis was restricted to English-language metadata records containing at least one keyword. While multilingual metadata presents an interesting avenue for future research, it is beyond the scope of this study.

4.2. Evaluation Metrics and Variable Choices

4.2.1. Evaluation Metrics:

We analyze the semantic keyword quality in terms of semantic relevance and semantic completeness as conceptualized in Section 3, in relation to the descriptive metadata they are associated with. We examine the extent to which keywords semantically align with and semantically represent the descriptive metadata.

4.2.2. Keyword Baseline:

Ideally, semantic completeness would be analyzed against a human-validated gold standard of metadata concepts. However, creating such annotations at the scale of the dataset under investigation is impractical. Consequently, metadata concepts are approximated using NLP-based keyword extraction. Specifically, concepts are extracted from dataset titles and descriptions using KeyBERT, a transformer-based keyword extraction method [19], and used as a proxy representation of the information expressed in the descriptive metadata. To analyze robustness to the choice of extraction method, we additionally

[3] https://data.europa.eu/en/download-datasets, accessed on 16 April 2026.

evaluated the approach using YAKE [20]. The results exhibited consistent qualitative trends despite minor differences in absolute semantic completeness values (Appendix B), suggesting that the findings are not specific to the selected keyword extraction technique.

4.2.3. Choice of the Threshold Value (τ):

For the semantic completeness metric, we set the similarity threshold τ to 0.7 as a practical criterion for identifying strong semantic similarity. This threshold is informed by prior empirical evidence in the literature [21, 22], which indicates that this threshold offers a balanced trade-off for determining strong semantic alignment without imposing overly strict matching constraints. We additionally perform sensitivity analysis under alternative thresholds ($\tau \in \{0.6, 0.8\}$) to analyze the robustness of the observed trends. While absolute semantic completeness scores vary as expected, the qualitative trends remain consistent, indicating that the observed limitations in keyword representation are not an artifact of a single threshold choice. The results of this analysis are presented in Table 3.

4.2.4. Choice of Embedding Model:

Lastly, for computing embeddings, we use the all-MiniLM-L6-v2[4] model from the SentenceTransformers framework. It has been widely adopted for semantic similarity and information retrieval tasks due to its balance between computational efficiency and strong semantic representation performance [23]. A robustness analysis using an alternative sentence embedding model (`all-mpnet-base-v2`) [24] yielded consistent qualitative trends for both semantic relevance and semantic completeness. While threshold-based semantic completeness values showed modest sensitivity to encoder choice, the comparative pattern of higher title semantic completeness relative to description semantic completeness remained stable across thresholds (Appendix A, Tables 5 and 6).

5. Results and Discussion

In this section, we present and discuss the results of semantic keyword quality evaluation on a collection of datasets from the European Open Data Portal.

5.1. Semantic Relevance of Keywords

As semantic relevance is measured using cosine similarity, values theoretically range from -1 to 1; consequently, small negative values should be interpreted as weak semantic mismatch rather than semantic contradiction. The identical minimum values in Table 1 arise from datasets with a single assigned keyword ($k = 1$), for which the aggregated and average keyword-level relevance measures are equivalent.

Table 1
Summary statistics of semantic relevance metrics in relation to dataset titles and descriptions.

Metric	Agg Title	Agg Description	Avg Title	Avg Description
Mean	0.435	0.407	0.298	0.278
Median	0.440	0.417	0.284	0.273
Std	0.177	0.145	0.135	0.105
Min	-0.084	-0.081	-0.084	-0.081
Max	1.000	1.000	1.000	1.000

The evaluation results indicate that the keywords exhibit low-moderate semantic relevance to associated descriptive metadata. As shown in Table 1, the mean aggregated similarity scores of approximately 0.44 for titles and 0.41 for descriptions, together with median values close to 0.44 and 0.42, suggest that

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

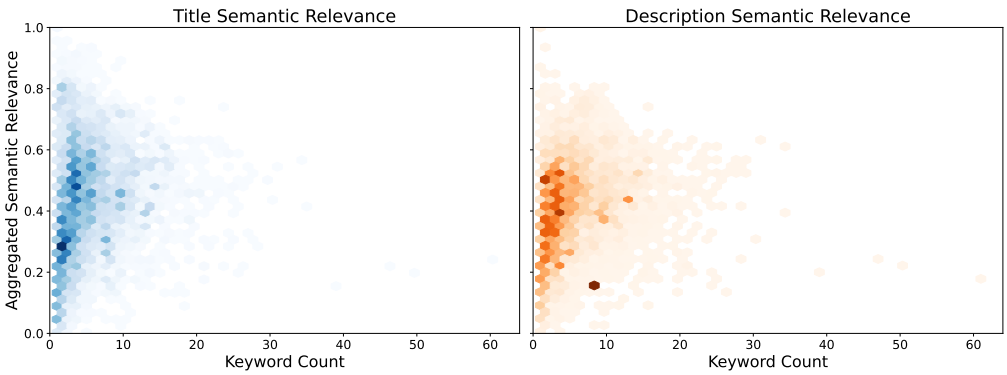

Figure 1: Density plot showing the relationship between the number of assigned keywords and aggregated semantic relevance.

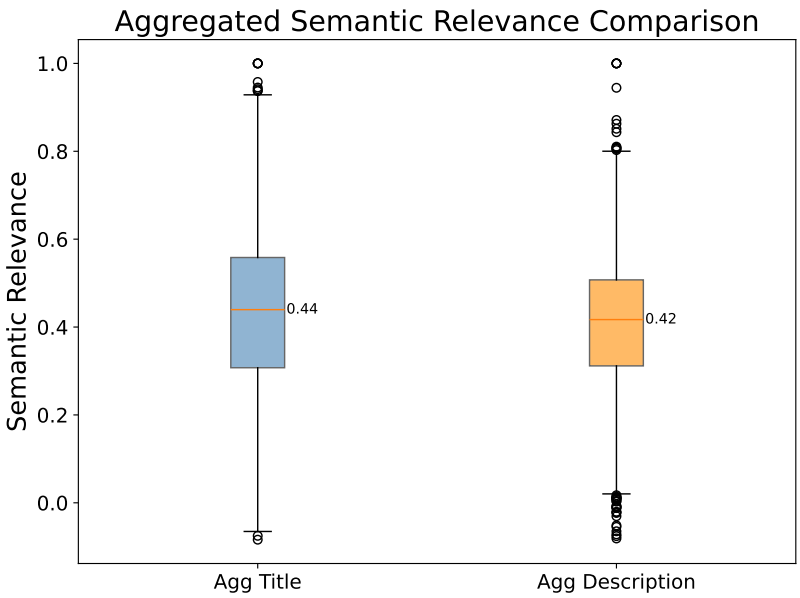

Figure 2: Box plot comparison of aggregated semantic relevance scores for dataset titles and descriptions.

keywords are on average semantically related to the descriptive metadata, although this alignment is not particularly strong. The results further show that keywords are slightly more aligned with titles than descriptions. The average keyword-level similarities presented in Table 1 are lower than the aggregated similarities, with mean values of 0.30 for titles and 0.28 for descriptions. This indicates that while keyword sets collectively capture the semantic meaning of metadata reasonably well, individual keywords are often only partially aligned with the metadata semantics.

Figure 1 shows that a higher number of assigned keywords does not necessarily correspond to higher aggregated semantic relevance. While some datasets with relatively few keywords achieve high relevance, datasets with larger keyword sets exhibit considerable variation in similarity scores. Consistent with the summary statistics in Table 1, aggregated relevance remains only low-to-moderate across keyword counts, suggesting that semantic alignment is more important than keyword quantity. The figure also indicates that title relevance is generally slightly higher than description relevance, particularly for datasets with fewer keywords. The box plots in Figures 2 and 3 further show that most datasets fall within a low-to-moderate relevance range rather than exhibiting strong semantic alignment. Although a small number of datasets achieve relevance scores close to 1.0, particularly for titles, such cases are uncommon. The presence of low and negative similarity values further suggests that some datasets contain poorly aligned or potentially noisy keywords that do not adequately correspond to the

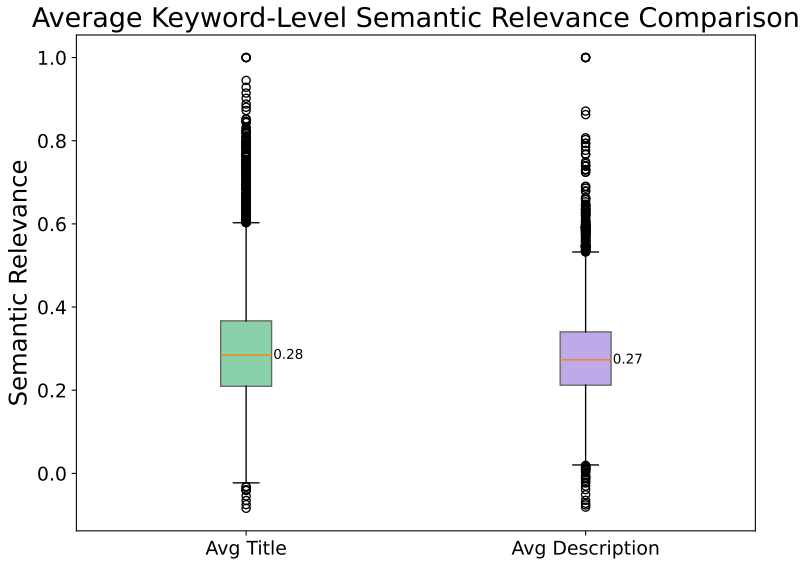

Figure 3: Box plot comparison of average keyword-level semantic relevance scores for dataset titles and descriptions.

Table 2
Illustrative examples of semantic relevance.

Dataset	Assigned keywords	Relevance	Observation
Heritage Partnerships Agreements	Heritage Agreements	0.90	The assigned keyword is closely aligned with the dataset topic and expresses a highly related concept. This strong semantic correspondence results in high semantic relevance.
Heritage at Risk Register	historic environment	0.05	The assigned keyword is only loosely aligned with the dataset topic and represents a broader heritage concept. This weak semantic correspondence results in low semantic relevance.

descriptive metadata. Table 2 provides illustrative examples of high and low semantic relevance. The examples highlight that high relevance is typically associated with keywords that closely match the dataset topic, whereas low relevance often results from broad thematic keywords that fail to capture the specific concepts expressed in the metadata.

The relatively large number of outliers observed in the semantic relevance boxplots reflects substantial variation in keyword assignment practices across datasets. An analysis of the outlier results shows that high-relevance outliers often occur when highly specific keywords closely match the dataset title (e.g., *Strategic Industrial Land* assigned the keyword *Strategic Industrial Land*, relevance score 1.0), whereas low-relevance outliers arise when broad thematic keywords only partially represent the dataset content (e.g., *All Sources (within agglomerations) Noise Contours Lden Round 2* assigned *Human health and safety*, relevance score −0.084). This pattern is particularly evident for the average keyword-level relevance metric, which is more sensitive to individual keyword similarities than the aggregated relevance metric. Overall, while keywords are generally semantically related to dataset metadata, their relevance remains only low to moderate, highlighting limitations in the precision and consistency of keyword assignment, particularly for richer description metadata.

5.2. Semantic Completeness of Keyword Annotations

Table 3 summarizes the semantic completeness results for $\tau = 0.7$ together with a robustness analysis across alternative threshold values. Although absolute scores vary with the threshold, the overall qualitative trends remain consistent.

Table 3
Summary statistics and threshold robustness analysis of semantic completeness.

(a) Summary Statistics ($\tau = 0.7$)

Metric	Title	Description
Mean	0.325	0.278
Median	0.200	0.200
Std	0.314	0.317
Min	0.000	0.000
Max	1.000	1.000

(b) Mean Semantic Completeness Across Threshold Values

τ	Title	Description
0.6	0.418	0.394
0.7	0.325	0.278
0.8	0.207	0.150

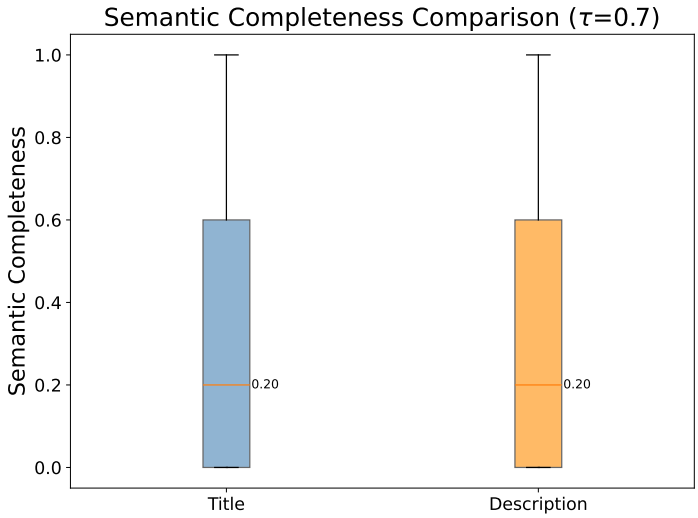

Figure 4: Box plot comparison of semantic completeness scores for dataset titles and descriptions at $T = 0.7$.

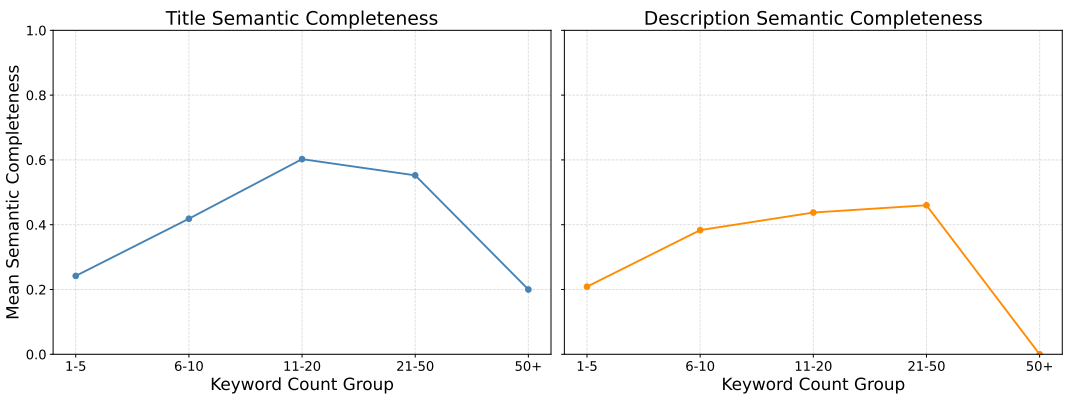

Figure 5: Mean semantic completeness across keyword-count groups for dataset titles and descriptions ($\tau = 0.7$).

The semantic completeness results indicate that the assigned keywords are limited in their semantic representation of the concepts expressed in the descriptive metadata. As shown in Table 3, the mean semantic completeness values are approximately 0.33 for titles and 0.28 for descriptions, with median

Table 4
Illustrative examples of semantic completeness.

Dataset	Assigned keywords	Completeness	Observation
Woodland Grant Scheme	Woodland, grants, forestry	0.79	The assigned keywords represent the principal concepts expressed in the dataset title and include relevant related terminology. Because most of the semantic content is covered, the dataset exhibits high, though not complete, semantic completeness.
Air Quality Management Areas	Air Quality	0.07	The assigned keyword represents the broader environmental theme of the dataset but omits the management-area concept. As a result, only limited semantic content is covered, leading to low semantic completeness.

values of 0.20 for both metadata fields. This indicates that only a small fraction of the concepts expressed in the descriptive metadata are semantically represented by assigned keywords, as estimated through the metadata-derived keyword proxy. The results further show that concepts expressed in title metadata are generally better represented than those expressed in description metadata. Table 4 provides illustrative examples of high and low semantic completeness. The examples show how differences in semantic coverage arise from the extent to which assigned keywords represent concepts expressed in the metadata.

Figure 4 further illustrates the distribution of semantic completeness scores across the datasets. Notably, the median semantic completeness is 0.2 for both titles and descriptions, indicating that a typical dataset has only a small proportion of metadata concepts semantically covered by its assigned keywords. The low median values and concentration of datasets in lower semantic completeness ranges indicate a systematic lack of completeness in keyword assignments. Although some datasets achieve high semantic completeness values approaching 1.0, these cases are relatively limited and do not represent the general trend observed across the datasets. In Figure 5, it is clear that a higher number of keywords does not necessarily guarantee an improvement in semantic completeness. While there is a slight upward trend in some cases, the overall pattern suggests a weak relationship between keyword quantity and semantic completeness. This observation suggests that additional keywords may introduce semantically overlapping or redundant terms rather than capturing additional concepts expressed in the descriptive metadata.

In summary, the keyword sets across the datasets provide incomplete semantic representations of the descriptive metadata, particularly for dataset descriptions where many concepts expressed in the metadata remain semantically unrepresented. The results indicate limitations in the completeness and representativeness of keyword assignment practices across open data portals.

5.3. Semantic Relevance Versus Semantic Completeness

The results reveal a clear divergence between semantic relevance and semantic completeness across the datasets. While relevance exhibits low alignment with descriptive metadata (mean aggregated similarity of approximately 0.42), semantic completeness remains substantially lower. This suggests that although assigned keywords are generally related to dataset metadata, they capture only a limited portion of the concepts expressed in descriptive metadata, as estimated through the metadata-derived keyword proxy. The median values of 0.43 and 0.20 further support this observation. For example, the dataset *Heritage Partnerships Agreements* is annotated with the keyword *Heritage Agreements*. Although the

keyword exhibits high semantic relevance due to its close semantic alignment with the dataset topic, it achieves low semantic completeness because it represents only a subset of the concepts expressed in the metadata. This example illustrates that semantically relevant keywords do not necessarily provide comprehensive semantic coverage, highlighting the need to consider both semantic relevance and semantic completeness when evaluating semantic keyword quality.

From a governance perspective, these findings highlight limitations in current metadata curation practices within open data portals. The observed gap suggests that existing keyword annotation processes may not adequately capture the full semantic scope of metadata representations, potentially affecting metadata-driven dataset discovery. This underscores the need for governance strategies that promote both semantic relevance and semantic completeness through guidelines, validation mechanisms, and automated keyword enrichment.

In summary, the results indicate that semantic keyword quality depends on both semantic relevance and semantic completeness. The substantial gap between these dimensions suggests that improving keyword quality requires not only ensuring keyword–metadata alignment but also increasing the coverage of metadata concepts represented by keywords.

6. Conclusion and Future Work

This paper presents an operational adaptation of established metadata quality dimensions for analyzing keyword annotations and analyzes semantic keyword quality in open data portals using the European Open Data Portal as a case study. We propose a structured approach that operationalizes semantic keyword quality through semantic relevance and semantic completeness to assess how well keywords align with descriptive metadata. Rather than evaluating retrieval effectiveness, this work focuses on defining and measuring semantic keyword quality, providing a foundation for future studies on its impact on dataset discoverability.

The experimental analysis highlights low-to-moderate semantic relevance of assigned keywords in relation to dataset descriptive metadata, particularly for dataset titles. The presence of low similarity scores further indicates inconsistencies in keyword annotation across datasets. In contrast, the semantic completeness results reveal that keyword annotations provide only limited semantic representation of the concepts expressed in descriptive metadata, particularly for richer dataset descriptions. This highlights a clear gap between semantic relevance and semantic completeness in current keyword assignment practices. Our findings therefore provide valuable insights regarding keyword annotations on Open Data Portals and emphasize the need to better capture the semantic content of descriptive metadata fields. Our findings further suggest that analyzing semantic keyword quality provides an additional lens for evaluating metadata quality, particularly in metadata-driven discovery environments where keyword annotations play an important role in dataset representation and access. Overall, the results indicate that keyword annotations in open data portals often provide an incomplete semantic representation of dataset metadata, highlighting substantial opportunities for improving metadata-driven dataset discovery. The proposed SKQ metrics offer a systematic approach to assessing keyword annotation quality and may support metadata governance in open data ecosystems; however, future work is needed to establish their relationship with discoverability, retrieval effectiveness, and FAIR findability.

Although this study analyzes semantic alignment and semantic completeness as indicators of semantic keyword quality, it does not establish their impact on dataset retrieval performance in realistic discovery scenarios. Furthermore, the aggregated semantic relevance metric relies on mean pooling and may be sensitive to noisy or semantically unrelated keywords, while the analysis assumes that descriptive metadata is a meaningful proxy for dataset content without directly validating metadata–content alignment. Semantic completeness is estimated using automatically extracted keywords as proxy concepts rather than human-validated annotations; consequently, the reported scores should be interpreted as estimates whose accuracy depends on the concept extraction process. Future work could evaluate the effectiveness of different keyword extraction methods in approximating dataset concepts through comparison with

expert-annotated concept sets. It will also investigate the relationship between semantic keyword quality and dataset discoverability, multilingual keyword settings, alternative concept extraction and aggregation strategies, and automatic keyword enrichment approaches for improving semantic keyword quality and metadata-driven dataset discovery.

Acknowledgments

This work was primarily funded by the Fonds de la Recherche Scientifique – FNRS under Grant No. MIS F.4016.24. Partial mobility support covering the initial travel costs to Belgium was provided to the first author by Strathmore University, through the Staff Development Programme.

Declaration on Generative AI

During the preparation of this work, ChatGPT was used in proofreading, grammar check and condensing sections of the manuscript to attain the page limit. ChatGPT was also used as an assistive tool in code debugging and optimization. All scientific content, results, and conclusions were developed solely by the authors. The authors proofread the work and take full responsibility of the content.

Availability of Data and Materials.

The datasets and code used in this study are publicly available at our GitHub repository: https://github.com/DeperiasKerre/keyword_quality_eval. The authors commit to making those assets available on Zenodo if the paper is accepted.

References

[1] J. Attard, F. Orlandi, S. Scerri, S. Auer, A systematic review of open government data initiatives, Government Information Quarterly 32 (2015) 399–418. doi:10.1016/j.giq.2015.07.006.

[2] J. Attard, F. Orlandi, S. Auer, Value creation on open government data, in: 49th Hawaii International Conference on System Sciences, 2016, pp. 2605–2614. doi:10.1109/HICSS.2016.326.

[3] A. Quarati, Open government data: Usage trends and metadata quality, Journal of Information Science 49 (2023) 887–910. URL: https://doi.org/10.1177/01655515211027775. doi:10.1177/01655515211027775.

[4] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 1–9. doi:10.1038/sdata.2016.18.

[5] R. Albertoni, D. Browning, S. Cox, A. González Beltrán, A. Perego, P. Winstanley, Data catalog vocabulary (dcat) – version 2, W3C Recommendation, 2020. URL: https://www.w3.org/TR/vocab-dcat-2/.

[6] U. Ahmed, C. Alexopoulos, M. Piangerelli, A. Polini, Bryt: Automated keyword extraction for open datasets, Intelligent Systems with Applications 23 (2024) 200421. doi:10.1016/j.iswa.2024.200421.

[7] J. Nogueras-Iso, J. Lacasta, M. A. Ureña-Cámara, F. J. Ariza-López, Quality of metadata in open data portals, IEEE Access 9 (2021) 60364–60382. doi:10.1109/ACCESS.2021.3073455.

[8] L. Yang, Metadata effectiveness in internet discovery, College and Research Libraries 77 (2016) 7–19. doi:10.5860/crl.77.1.7.

[9] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: A survey, VLDB Journal 29 (2020) 251–272. doi:10.1007/s00778-019-00564-x.

[10] J. Martinez-Gil, Framework to automatically determine the quality of open data catalogs, Expert Systems with Applications 289 (2025) 128379. doi:10.1016/j.eswa.2025.128379.

[11] K. J. Reiche, E. Höfig, Implementation of metadata quality metrics and application on public government data, in: IEEE COMPSAC Workshops, 2013, pp. 236–241. doi:10.1109/COMPSACW.2013.32.

[12] L. Y. Gan, A. Das, J. Walker, E. Simperl, Keywords are not always the key: A metadata field analysis for natural language search on open data portals, in: J. F. Krems, H. P. da Silva, P. Cipresso (Eds.), Computer-Human Interaction Research and Applications, volume 2835 of *Communications in Computer and Information Science*, Springer, Cham, 2026, pp. 421–440. doi:10.1007/978-3-032-16451-3_26.

[13] L. Wenige, C. Stadler, M. Martin, R. Figura, R. Sauter, C. W. Frank, Open data and the status quo–a fine-grained evaluation framework for open data quality and an analysis of open data portals in germany, arXiv preprint arXiv:2106.09590 (2021). URL: https://doi.org/10.48550/arXiv.2106.09590. doi:10.48550/arXiv.2106.09590.

[14] 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.

[15] I. Peters, W. G. Stock, Folksonomy and information retrieval, Proceedings of the American Society for Information Science and Technology 44 (2007) 1–28. doi:10.1002/meet.1450440226.

[16] J. Xing, D. Luo, C. Xue, R. Xing, Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective, arXiv preprint arXiv:2411.14654 (2024). URL: https://doi.org/10.48550/arXiv.2411.14654. doi:10.48550/arXiv.2411.14654.

[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[18] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, S. Quarteroni, An introduction to information retrieval, in: Web Information Retrieval, 2013, pp. 3–11. doi:10.1007/978-3-642-39314-3_1.

[19] M. Grootendorst, Keybert: Minimal keyword extraction with bert, GitHub Repository, 2020. URL: https://github.com/MaartenGr/KeyBERT.

[20] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289. doi:10.1016/j.ins.2019.09.013.

[21] E. A. Chimoto, B. A. Bassett, Very low resource sentence alignment: Luhya and swahili, in: Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022), Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 1–8. URL: https://aclanthology.org/2022.loresmt-1.1/.

[22] T. J. B. Cann, B. Dennes, T. Coan, S. O'Neill, H. T. P. Williams, Using semantic similarity to measure the echo of strategic communications, EPJ Data Science 14 (2025). URL: https://doi.org/10.1140/epjds/s13688-025-00538-w. doi:10.1140/epjds/s13688-025-00538-w.

[23] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: NeurIPS, volume 33, 2020, pp. 5776–5788.

[24] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, in: NeurIPS, volume 33, 2020, pp. 16857–16867.

Appendix

A. Robustness Analysis Across Alternative Embedding Models

For this robustness analysis, only the sentence embedding model was varied, while the keyword extraction method (KeyBERT) and all other experimental settings were kept fixed.

Table 5
Semantic relevance robustness across embedding models. Semantic relevance scores remain largely consistent across encoder architectures, suggesting that the observed trends are robust to embedding model choice.

(a) Baseline: all-MiniLM-L6-v2

Metric	Mean	Median
Agg Title	0.4354	0.4396
Agg Description	0.4067	0.4170
Avg Title	0.2979	0.2844
Avg Description	0.2776	0.2730

(b) Alternative: all-mpnet-base-v2

Metric	Mean	Median
Agg Title	0.4331	0.4393
Agg Description	0.4232	0.4322
Avg Title	0.2979	0.2849
Avg Description	0.2910	0.2827

Table 6
Robustness analysis of semantic completeness across semantic similarity thresholds (τ) and sentence embedding models. Although absolute values vary, the overall trends remain consistent across thresholds and embedding models.

(a) Baseline: all-MiniLM-L6-v2

τ	Statistic	Title	Desc.
0.6	Mean	0.4184	0.3935
0.6	Median	0.4000	0.4000
0.7	Mean	0.3251	0.2777
0.7	Median	0.2000	0.2000
0.8	Mean	0.2067	0.1495
0.8	Median	0.2000	0.0000

(b) Alternative: all-mpnet-base-v2

τ	Statistic	Title	Desc.
0.6	Mean	0.3901	0.3621
0.6	Median	0.4000	0.4000
0.7	Mean	0.2740	0.2204
0.7	Median	0.2000	0.2000
0.8	Mean	0.1665	0.1106
0.8	Median	0.2000	0.0000

B. Metadata Concept Extraction Robustness Analysis

For this robustness analysis, only the keyword extraction method was varied, while the sentence embedding model (`all-MiniLM-L6-v2`) and all other experimental settings were kept fixed.

Table 7
Semantic completeness robustness across similarity thresholds and keyword extraction methods. Although absolute values vary, the overall trends remain consistent across thresholds and extraction methods.

(a) Baseline: KeyBERT

τ	Statistic	Title	Desc.
0.6	Mean	0.4184	0.3935
0.6	Median	0.4000	0.4000
0.7	Mean	0.3251	0.2777
0.7	Median	0.2000	0.2000
0.8	Mean	0.2067	0.1495
0.8	Median	0.2000	0.0000

(b) Alternative: YAKE

τ	Statistic	Title	Desc.
0.6	Mean	0.3735	0.2705
0.6	Median	0.4000	0.2000
0.7	Mean	0.2912	0.2089
0.7	Median	0.2000	0.2000
0.8	Mean	0.1959	0.1341
0.8	Median	0.2000	0.0000