technologies

MDPI

Article

A Governance Aware Ontology Based Analytics Framework for Scholarly Knowledge Graphs Using Budgeted Semantic Coverage Optimization

Esteban Inga [1,*], Vladimir Robles [2,†], Daniel Lopez [1,†] and Daniel Pinargo [1,†]

[1] Research Vice-Rectorate Department, Universidad Politécnica Salesiana, Cuenca 010102, Ecuador; elopezg@ups.edu.ec (D.L.); dpinargo@ups.edu.ec (D.P.)
[2] Postgraduate Vice-Rectorate Department, Universidad Politécnica Salesiana, Cuenca 010102, Ecuador; vrobles@ups.edu.ec
* Correspondence: einga@ups.edu.ec; Tel.: +593-992-808-415
† These authors contributed equally to this work.

Abstract

This study presents GAGER, a governance-aware ontology-based framework for scholarly knowledge graph analytics evaluated as a single-university case study. The framework integrates bibliometric normalization, semantic organization, thematic clustering, provenance control, governance validation, and budgeted semantic coverage optimization. It transforms Scopus-based institutional bibliographic records into an operational knowledge infrastructure in which documents, authors, institutions, countries, publication venues, document types, knowledge areas, thematic clusters, and Sustainable Development Goal tags are jointly represented through a traceable semantic graph. GAGER uses the ontology to define candidate semantic anchors and applies a maximum-coverage optimization model to select the most informative governance-valid anchors under explicit budget constraints. Experiments using a real single-university corpus show that the proposed governance-validated ontology-plus-optimization strategy achieves higher semantic coverage and aggregate decision quality than ontology-only selection, random selection, and frequency-based selection baselines evaluated under the same budget constraints. The governance layer also has a measurable operational effect, reducing the initial candidate-anchor space from 27,542 anchors to 8320 validated anchors before optimization. The contribution is therefore a methodologically reproducible single-institution analytical framework that connects semantic governance, provenance, graph-based bibliometric representation, and budgeted optimization to support institutional research analytics, thematic prioritization, and evidence-based decision-making within the analyzed university context.

Keywords: scholarly knowledge graphs; ontology based analytics; bibliometric analytics; semantic coverage optimization; budgeted maximum coverage; provenance graph; research governance; thematic clustering; institutional analytics; higher education decision making

Check for updates

Academic Editors: Arash Habibi Lashkari and Zhipeng Wu

Received: 15 April 2026
Revised: 20 July 2026
Accepted: 22 July 2026
Published: 24 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Universities and research organizations increasingly rely on bibliometric data to monitor scientific production, identify thematic strengths, evaluate collaboration, and support evidence-based decision-making. However, bibliographic datasets are rarely ready for direct analytical use. They usually contain heterogeneous metadata, inconsistent keywords,

fragmented affiliations, uneven citation patterns, duplicated or ambiguous entities, and dispersed thematic structures. These limitations reduce the reliability of institutional dashboards, ranking analyses, accreditation evidence, and research-governance instruments. Recent studies on scholarly knowledge graphs and intelligent academic data management show that structured semantic representations can improve the integration, interpretation, and reuse of scientific information in institutional and knowledge-management environments [1–3].

Scholarly knowledge graphs offer a structured way to organize this complexity by connecting documents, authors, institutions, countries, topics, publication venues, document types, and impact dimensions within a unified semantic representation. Nevertheless, semantic representation alone does not determine which parts of a large academic knowledge space are the most informative or strategically relevant. A graph can describe relationships among entities, but it does not automatically select the subset of themes, knowledge areas, SDG tags, document types, or collaboration configurations that maximizes coverage and diversity under limited analytical resources. This limitation becomes more evident when scholarly graphs are affected by author-name ambiguity, institution-name inconsistency, record linkage issues, entity alignment problems, and heterogeneous metadata structures [4–9].

The research gap addressed in this study is the limited integration between ontology-based scholarly knowledge graph construction and budgeted semantic coverage optimization. Existing approaches provide valuable mechanisms for author disambiguation, entity matching, record linkage, entity alignment, multimodal alignment, ontology mapping, and knowledge graph construction [10–27]. However, fewer studies treat the selection of semantic anchors as an explicit optimization problem that maximizes coverage, diversity, and institutional utility within a bibliometric knowledge graph.

This study proposes GAGER, a governance-aware, ontology-based framework for scholarly knowledge graph analytics that leverages budgeted semantic coverage optimization. The framework converts a real bibliometric corpus into a semantic graph, derives thematic clusters and knowledge-area relationships, defines candidate semantic anchors, and solves a maximum-coverage selection problem under explicit budget constraints. The model is not presented as a full entity-resolution algorithm. Instead, it uses semantic normalization and a graph-based representation to support decision-oriented research analytics, while drawing on prior advances in governance-aware deduplication, ontology-based entity matching, RDF knowledge graph construction, and large-language-model-assisted knowledge graph generation [28–30].

The objective of this work is to develop and evaluate a reproducible computational framework that connects bibliometric normalization, ontology-based semantic organization, provenance control, governance validation, graph construction, semantic-anchor selection, and baseline-based evaluation. The empirical evidence is limited to a Scopus-based single-university case study. Therefore, the framework is not claimed as a general model for all higher education systems or science-policy contexts; rather, it is evaluated as an institutional analytics approach for identifying which semantic components of the analyzed university corpus provide the highest coverage and diversity under constrained analytical conditions.

The main contributions of this study are fourfold. First, it introduces an ontology-based representation of a single-university scholarly corpus that integrates documents, knowledge areas, thematic clusters, SDG tags, document types, countries, and publication venues. Second, it formulates governance-valid semantic-anchor selection as a budgeted maximum-coverage problem. Third, it evaluates the proposed strategy against three constrained baselines: ontology-only selection, random anchor selection, and frequency-based selection. These baselines are included to determine whether the optimization layer produces gains

beyond semantic annotation, arbitrary selection, or simple frequency-driven prioritization. Fourth, it provides an interpretable analytical structure for institutional evaluation, thematic prioritization, and evidence-based decision-making within the analyzed university case.

The remainder of the article is organized as follows. Section 2 reviews scholarly knowledge graphs, entity-resolution and alignment methods, ontology-based academic analytics, and optimization-based selection in knowledge systems. Section 3 presents the proposed formulation and methodology. Section 4 reports the experimental results and baseline comparisons. Section 5 discusses the implications of the findings, and Section 6 presents the study's limitations. Section 7 concludes the article and outlines future work.

2. Related Works

2.1. Scholarly Knowledge Graphs and Bibliometric Knowledge Infrastructures

Scholarly knowledge graphs have become relevant instruments for representing scientific production, institutional collaboration, thematic structures, and citation-based relationships in a unified analytical environment. These graphs support the integration of publications, authors, organizations, venues, topics, and impact indicators, allowing bibliometric information to be reused for discovery, evaluation, monitoring, and strategic planning. Verma et al. provide a broad review of scholarly knowledge graphs and show their role in structuring scholarly communication [1]. Liu proposes a knowledge graph-based intelligent data-management model for university library systems, highlighting the relevance of graph structures for academic information services [2]. Rajabi et al. demonstrate the construction of a knowledge graph for open government disease datasets, showing that semantic graph structures can support data integration and reuse beyond purely bibliometric environments [3].

Despite these advantages, scholarly knowledge graphs face significant challenges in data quality and semantic coherence. Bibliographic records often contain inconsistent names, incomplete affiliations, redundant keywords, heterogeneous document types, and uneven disciplinary descriptions. Author-name ambiguity is one of the most studied problems in scholarly data. Santini et al. propose a knowledge graph embedding-based approach for author-name disambiguation using literals [4], while Rodrigues et al. provide a consolidated meta-analytic review of the author-name disambiguation literature [5]. Zhou et al. address author-name disambiguation through hybrid attention mechanisms [6], and Choi et al. propose a name-disambiguation scheme based on heterogeneous academic sites [7]. These studies show that local identity consistency remains a core requirement for reliable scholarly data infrastructures.

Institutional-name consistency and record linkage also affect the reliability of scholarly analytics. Jia et al. address the normalization of Web of Science institution names using deep learning [8]. Azeroual et al. propose a record linkage-based data deduplication framework supported by a DataCleaner extension [9]. Ali Omar et al. examine training data selection for record linkage classification, emphasizing the importance of labeled examples and the design of classification methods for matching tasks [10]. Shah et al. analyze efficient record linkage in the age of large language models and highlight the critical role of blocking as a scalability mechanism [11]. Karapiperis and Verykios review partitioning techniques for scaling entity resolution using k-means, showing that blocking and partitioning remain essential as the number of candidate comparisons increases [12]. Nananukul et al. explore cost-efficient prompt engineering for unsupervised entity resolution in product matching, illustrating the increasing use of language models in matching tasks [13].

The present study differs from these works because it does not formulate the problem as full pairwise entity resolution, record linkage, or duplicate-record matching. Instead, these methods are treated as relevant foundations for understanding the data-quality chal-

lenges that affect scholarly knowledge graphs. The proposed framework uses bibliometric normalization and semantic mapping as preprocessing steps, then focuses on constructing a decision-oriented scholarly knowledge graph and optimizing semantic-anchor selection.

2.2. Entity Alignment, Multimodal Alignment, and Entity Linking

Entity alignment has become a central topic in knowledge graph research because it supports the integration of equivalent entities across heterogeneous graphs. Shu et al. study entity alignment using graph neural networks and provide a component-level analysis of the mechanisms underlying alignment performance [14]. Zhu et al. review knowledge graph entity-alignment research based on graph embedding, showing the relevance of embedding representations for matching entities across graphs [15]. Zhou et al. propose a knowledge graph-based entity alignment method with unified representation for auditing, connecting alignment with governance-sensitive analytical settings [16]. These works demonstrate that entity alignment is essential for integrating heterogeneous knowledge structures consistently.

Recent entity-alignment research has also incorporated multimodal information. Li et al. propose a multimodal entity-alignment method based on image enhancement and credibility iteration [17]. Fang and Yan introduce MDSEA, a knowledge graph entity-alignment approach based on multimodal data supervision [19]. Wang et al. develop a fusion-optimized multimodal entity-alignment method with textual descriptions [21]. Li et al. propose a multimodal entity-alignment approach based on enhanced relationship learning and multi-layer feature fusion [23]. Zhu et al. introduce UMEAD, an unsupervised multimodal entity-alignment method for equipment knowledge graphs using dual-space embedding [24]. These approaches show that alignment increasingly depends on integrating structural, textual, visual, and relational information.

Other studies extend alignment through advanced architectures or specialized benchmarks. Cai et al. propose an efficient and effective unsupervised entity-alignment method for large knowledge graphs [20]. Cai and Ma introduce a learnable convolutional attention network for unsupervised knowledge graph entity alignment [22]. Li et al. present QLGAN, a quantum-lineage graph attention network for temporal knowledge graph entity alignment [18]. Gu et al. propose an attribute-enhanced relationship-aware neighborhood matching model with dual attention [27]. In the field of entity linking, Qu et al. propose an end-to-end knowledge graph entity-linking method for civil aviation emergencies [25], and Karpov et al. introduce SynEL as a synthetic benchmark for entity linking [26]. Together, these works show the breadth of current entity-alignment and entity-linking research.

Although these studies are highly relevant for knowledge graph integration, the present framework addresses a different decision layer. It does not attempt to align equivalent entities across multiple knowledge graphs or to link mentions to canonical entities. Instead, it assumes that a normalized scholarly corpus has been projected into a semantic graph and asks which semantic anchors should be selected to maximize coverage, diversity, and institutional utility under a constrained budget.

2.3. Ontology-Based Semantic Organization and Governance

Ontology-based approaches provide explicit structures for representing entities, attributes, relationships, evidence, provenance, and validation states. In academic analytics, this is important because institutional decisions require not only numerical indicators but also traceable definitions of how topics, fields, affiliations, and impact dimensions are assigned. Galán-Mena et al. propose OntoDup, a governance-aware entity-matching approach for scholarly knowledge graph deduplication, showing how governance criteria can be embedded into scholarly graph cleaning and consolidation processes [28].

Mavridis et al. study the use of large language models for intelligent construction of RDF knowledge graphs via medical ontology mapping [29]. Ghanem and Cruz evaluate fine-tuning and prompting strategies for knowledge graph construction tasks using large language models [30]. These works show that ontology-based and LLM-assisted methods can improve semantic construction, mapping, and interpretability.

Nevertheless, ontology-based representation does not automatically solve the problem of prioritization. A semantic graph can describe the meaning of entities and relations, but it does not decide which subset of semantic components should be selected when the analytical space is large and resources are limited. This limitation is especially relevant for research-governance environments, where institutions need to identify representative themes, knowledge areas, SDG alignments, and collaboration patterns without reducing the analysis to raw bibliometric counts. Therefore, ontology provides the semantic and governance layer, but an additional optimization layer is needed to support strategic selection.

2.4. Optimization-Based Selection and Semantic Coverage

Optimization-based methods enable transforming semantic representations into decision-oriented analytics. In particular, maximum-coverage models are useful when the objective is to select a limited number of elements that jointly cover the largest possible portion of a relevant universe. In the context of scholarly knowledge graphs, candidate elements can be interpreted as semantic anchors, including thematic clusters, knowledge areas, SDG tags, document types, countries, or publication venues. The universe to be covered corresponds to combinations of semantic states that describe the structure of the academic corpus.

This perspective differs from purely descriptive bibliometric analysis because it not only reports what exists in the corpus. Instead, it determines which semantic anchors to prioritize to maximize coverage and diversity under explicit budget constraints. It also differs from local entity-resolution methods because the main decision is not whether two records refer to the same entity, but which semantic structures provide the most useful and representative coverage of the scholarly knowledge space. In this sense, the proposed framework does not replace the contributions of author disambiguation, record linkage, entity alignment, entity linking, ontology mapping, or knowledge graph construction. Rather, it uses them as conceptual foundations and moves the contribution toward a budgeted semantic selection problem within an already-structured scholarly knowledge graph.

2.5. Research Gap

The reviewed literature provides four relevant foundations. First, scholarly knowledge graphs organize academic information and support semantic integration of scientific production [1–3]. Second, author disambiguation, institution-name normalization, record linkage, and entity-resolution methods improve the consistency of scholarly metadata [4–13]. Third, entity alignment, multimodal alignment, and entity linking support integration across heterogeneous knowledge structures [14–27]. Fourth, ontology-based and LLM-assisted approaches improve semantic construction, governance, provenance, and interpretability [28–30].

However, these components are rarely integrated into a single framework for scholarly knowledge graph analytics. The main gap is the lack of a reproducible model that connects semantic graph construction, provenance-aware governance, budgeted semantic-anchor selection, diversity-aware evaluation, and baseline-based comparison within a single scholarly knowledge graph analytics framework. The present study addresses this gap by proposing GAGER, a governance-aware ontology-based analytics framework that formulates semantic-anchor selection as a budgeted maximum-coverage problem over a

scholarly knowledge graph. Table 1 summarizes the position of the proposed framework with respect to representative prior works.

Table 1. Comparison of representative studies and the proposed framework.

Work	SKG	Ont.	Gov.	Prov.	Opt.	Base.
Galán-Mena et al. [28]		✓	✓	✓		
Santini et al. [4]	✓					
Rodrigues et al. [5]						
Zhou et al. [6]					✓	
Choi et al. [7]						
Jia et al. [8]						
Shah et al. [11]					✓	
Shu et al. [14]	✓				✓	
Zhu et al. [15]	✓					
Zhou et al. [16]	✓	✓	✓		✓	
Mavridis et al. [29]	✓	✓		✓		
Ghanem and Cruz [30]	✓	✓				
Verma et al. [1]	✓	✓				
Liu [2]	✓	✓	✓			
Rajabi et al. [3]	✓	✓				
Present work	✓	✓	✓	✓	✓	✓

Note: SKG = scholarly knowledge graph; Ont. = ontology; Gov. = governance; Prov. = provenance; Opt. = optimization; Base. = baseline evaluation. A checkmark indicates that the corresponding feature is present.

3. Problem Formulation and Methodology

The notation is defined in Table 2, and the computational sequence is summarized in Algorithm 1.

Figure 1 summarizes the complete workflow of the proposed framework. The process starts with Scopus-based bibliographic records and continues with data cleaning, semantic normalization, ontology and provenance construction, thematic clustering, scholarly knowledge graph generation, semantic-anchor definition, budgeted coverage optimization, and coverage and diversity evaluation.

The problem addressed in this study is the construction of a governance-aware semantic and analytical infrastructure from real-world bibliometric records. The framework does not perform full entity resolution or pairwise duplicate matching. Instead, it applies bibliometric normalization, semantic mapping, thematic clustering, provenance control, and budgeted semantic-anchor selection to transform a heterogeneous scholarly corpus into a decision-oriented knowledge graph.

The methodological challenge is to represent documents, authors, institutions, countries, publication venues, document types, knowledge areas, thematic clusters, and SDG tags within a unified graph, and then determine which semantic anchors provide the highest coverage and diversity under explicit budget constraints. Therefore, the proposed model should be understood as a scholarly knowledge graph analytics framework supported by budgeted semantic coverage optimization.

Table 2. Variables used in the proposed scholarly knowledge graph analytics framework.

Symbol	Description
$\mathcal{R}$	Set of real bibliographic records ($\lvert\mathcal{R}\rvert = n$)
$\mathcal{V}$	Set of semantic entities induced from the dataset
$\mathcal{E}$	Set of semantic relations in the bibliometric graph
$\mathcal{A}$	Set of authors
$\mathcal{I}$	Set of normalized institutions
$\mathcal{C}$	Set of inferred countries
$\mathcal{K}$	Set of unified keywords
$\mathcal{T}$	Set of thematic clusters
$\mathcal{S}$	Set of knowledge areas
$\mathcal{D}$	Set of document types
$\mathcal{G}$	Set of SDG labels
$\mathbf{x}_i$	Semantic–bibliometric attribute vector for record i
$x_i^{(m)}$	Value of attribute m for record i
α_m	Weight assigned to attribute m
s_i	Composite semantic score for record i
$\Pi_{\mathcal{T}}$	Partition of thematic clusters
$G = (\mathcal{V}, \mathcal{E}, w)$	Weighted bibliometric knowledge graph
w	Edge-weight function in the semantic graph
Ω	Universe of semantic states to be covered
$\mathcal{U}$	Set of candidate semantic anchors
$\Gamma(u_j)$	Coverage set induced by anchor u_j
z_j	Binary selection variable for anchor u_j
y_ω	Coverage activation variable for semantic state ω
B	Optimization budget
$\mathrm{Cov}(B)$	Coverage ratio reached under budget B
$\mathrm{Div}_{\mathcal{T}}(B)$	Topic diversity under budget B
$\mathrm{Div}_{\mathcal{S}}(B)$	Knowledge-area diversity under budget B
$\mathrm{Div}_{\mathcal{G}}(B)$	SDG diversity under budget B
$Q(B)$	Composite optimization score

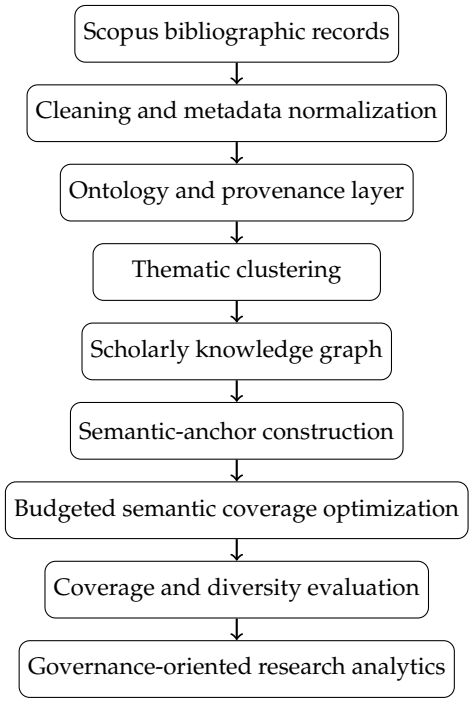


Figure 1. Overall workflow of the proposed governance-aware scholarly knowledge graph analytics framework.

Algorithm 1 GAGER: Governance-Aware Scholarly Knowledge Graph Analytics with Budgeted Semantic Coverage Optimization

Require: Real bibliographic dataset $\mathcal{R}$, semantic mapping rules, provenance rules, text fields, keyword fields, affiliation fields, budget set $\mathcal{B}$
Ensure: Scholarly knowledge graph G, provenance graph $\mathcal{P}$, thematic partition $\Pi_{\mathcal{T}}$, selected semantic anchors, coverage scores, diversity indicators, and baseline comparisons
Read and validate the real bibliographic dataset
Standardize text strings, separators, missing values, and categorical labels
Normalize document types, open-access status, language, publishers, and venues
Extract authors, affiliations, institutions, and inferred countries
Unify author keywords and index keywords into a semantic vocabulary
Derive bibliometric variables, collaboration variables, and impact profiles
Infer knowledge areas and SDG labels from titles, abstracts, and normalized keywords
Build textual representation for thematic analysis
Induce thematic partition $\Pi_{\mathcal{T}}$ using unsupervised clustering
Construct the weighted scholarly knowledge graph $G = (\mathcal{V}, \mathcal{E}, w)$
Construct the provenance graph $\mathcal{P}$ linking semantic assignments to source fields, rules, and validation states
Define the universe of semantic states Ω
Build the candidate semantic-anchor set $\mathcal{U}$ and the coverage map $\Gamma(u_j)$
Apply governance validation and remove anchors with missing provenance
for each budget level $B \in \mathcal{B}$ **do**
    Select semantic anchors using greedy maximum coverage under budget B
    Compute $\mathrm{Cov}(B)$, $\mathrm{Div}_{\mathcal{T}}(B)$, $\mathrm{Div}_{\mathcal{S}}(B)$ and $\mathrm{Div}_{\mathcal{G}}(B)$
    Evaluate the composite score $Q(B)$
end for
Compare the proposed strategy against ontology-only, random-selection, and frequency-based baselines
Export graphs, heatmaps, multicurves, and optimization summaries
return G, $\mathcal{P}$, $\Pi_{\mathcal{T}}$, selected anchors, coverage and diversity profiles, and baseline-comparison outputs

Let $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ be the set of actual bibliographic records. Each record r_i contains a set of attributes directly observed in the source and preserved following a phase of reading, validation, cleaning, and normalization. Rather than reducing the problem to a single equivalence decision, each record is projected onto a broader semantic space that includes authors, institutions, countries, unified keywords, document type, publisher, conference, thematic cluster, ESC area, SDGs, and impact profile. The purpose of this projection is not merely descriptive. Its goal is to convert a bibliometric CSV into a reusable knowledge structure for analysis, governance, and optimization.

Based on the actual records, a set of semantic entities is defined

$$\mathcal{V} = \mathcal{A} \cup \mathcal{I} \cup \mathcal{C} \cup \mathcal{K} \cup \mathcal{T} \cup \mathcal{S} \cup \mathcal{D} \cup \mathcal{G}, \tag{1}$$

where $\mathcal{A}$ represents authors, $\mathcal{I}$ institutions, $\mathcal{C}$ countries, $\mathcal{K}$ normalized keywords, $\mathcal{T}$ thematic clusters, $\mathcal{S}$ fields of knowledge according to the knowledge area criteria, $\mathcal{D}$ document types, and $\mathcal{G}$ SDG tags. This union forms the basis of the analytical semantic graph. Unlike a formulation focused solely on correspondences, the problem here is to construct a system of relationships among these entities, preserving provenance and traceability, to subsequently identify strategic subsets that maximize coverage of the knowledge space.

Each record r_i is transformed into a general vector of observable semantic attributes

$$\mathbf{x}_i = \left[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}\right]^{\top}, \tag{2}$$

where p represents the number of bibliometric and semantic features actually used by the algorithm. In the final implementation, these features include cleaned textual variables,

authorship counts, affiliation lists, inferred countries, unified keywords, normalized document type, open access status, language, citations, publisher, conference, and attributes derived from thematic clustering. The expression (2) keeps the formulation open so that new features can be added without modifying the system's overall logic.

The local semantic evidence for each record is summarized using a composite function

$$s_i = \sum_{m=1}^{p} \alpha_m x_i^{(m)}, \qquad \alpha_m \geq 0, \qquad \sum_{m=1}^{p} \alpha_m = 1, \tag{3}$$

where α_m controls the relative contribution of each bibliometric or semantic dimension. This expression does not act here as a paired classifier but rather as an aggregation function that allows weighting heterogeneous information on productivity, impact, semantics, and context. On this basis, derived variables are constructed for collaboration, thematic density, concentration by source, temporal evolution of themes, and relationships between fields, SDGs, and document types.

The thematic phase of the model projects the title, abstract, author keywords, and index keywords onto a vector space and then performs a thematic partition

$$\Pi_{\mathcal{T}} = \{T_1, T_2, \ldots, T_q\}, \tag{4}$$

where each T_ℓ represents a thematic cluster automatically inferred from textual co-occurrence and semantic proximity among documents. This partition does not replace the dataset's original tags, but rather introduces a latent layer of thematic organization that is subsequently linked to knowledge areas, SDGs, sources, document types, and collaboration networks. In the final implementation, the derivation of $\Pi_{\mathcal{T}}$ relies on a TF-IDF representation and an unsupervised clustering algorithm that produces interpretable, traceable clusters.

Next, the corpus is modeled as a bibliometric knowledge graph

$$G = (\mathcal{V}, \mathcal{E}, w), \tag{5}$$

where $\mathcal{V}$ is the set of semantic entities defined in (1), $\mathcal{E}$ is the set of observed or inferred relationships between these entities, and $w : \mathcal{E} \to \mathbb{R}_{\geq 0}$ is a weight function that quantifies frequency, co-occurrence, or link strength. Under this formulation, a relationship can express, for example, co-authorship, shared affiliation, co-occurrence of keywords, association between a field and SDGs, relationship between a conference and a thematic cluster, or link between a document and a document type. The resulting graph serves as the basis for visualizations, heatmaps, and structural metrics generated by the algorithm.

Governance and Provenance Layer

To make the framework explicitly governance-aware, each semantic assignment is linked to a provenance graph

$$\mathcal{P} = (\mathcal{V}_P, \mathcal{E}_P), \tag{6}$$

where $\mathcal{V}_P$ contains provenance entities associated with source fields, extraction dates, normalization rules, semantic-mapping rules, thematic-clustering outputs, SDG-assignment rules, and validation states. The set $\mathcal{E}_P$ links each analytical entity or semantic anchor to the evidence and rule that generated it.

For each semantic anchor u_j, a governance validation variable is defined as

$$v_j = \begin{cases} 1, & \text{if anchor } u_j \text{ has valid provenance and passes governance checks,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The selection of semantic anchors is constrained by provenance availability and governance validation:

$$z_j \leq v_j, \qquad j = 1, \ldots, M, \tag{8}$$

which prevents the optimization layer from selecting anchors without traceable evidence or a valid semantic assignment. Additionally, anchors with no provenance support are excluded:

$$z_j = 0 \qquad \text{if prov}(u_j) = \emptyset. \tag{9}$$

This layer ensures that semantic coverage is not optimized only by frequency or structural relevance, but also by traceability, auditability, and rule-based validation.

The optimization layer does not operate on pairs of duplicates, but rather on a set of candidate semantic anchors

$$\mathcal{U} = \{u_1, u_2, \ldots, u_M\}, \tag{10}$$

where each u_j can correspond to a knowledge area, an SDG label, or a thematic cluster. Each anchor covers a subset of the universe of semantic states

$$\Omega = \{\, \omega = (t, s, g, d, c) \,\}, \tag{11}$$

where t is a thematic cluster, s is a knowledge area, g is an SDG tag, d is a document type, and c is a country or collaboration configuration. In operational terms, Ω represents the space of relevant combinations that describe the structure of institutional knowledge. The coverage of an anchor u_j is denoted by $\Gamma(u_j) \subseteq \Omega$.

The main binary decision variable in the optimization model is

$$z_j = \begin{cases} 1, & \text{if the semantic anchor } u_j \text{ is selected,} \\ 0, & \text{otherwise,} \end{cases} \qquad j = 1, \ldots, M. \tag{12}$$

Furthermore, a coverage activation auxiliary variable is introduced

$$y_\omega = \begin{cases} 1, & \text{if the semantic state } \omega \in \Omega \text{ is covered,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The first variable determines which anchors are retained under a budget constraint; the second indicates what fraction of the knowledge space is effectively represented by that selection.

The basic objective function is to maximize semantic coverage under a finite budget

$$\max_{\mathbf{z},\mathbf{y}} \sum_{\omega \in \Omega} y_\omega, \tag{14}$$

subject to

$$y_\omega \leq \sum_{j:\, \omega \in \Gamma(u_j)} z_j, \qquad \forall \omega \in \Omega, \tag{15}$$

$$\sum_{j=1}^{M} z_j \leq B, \tag{16}$$

where B is the selection budget. The constraint (15) ensures that a semantic state is considered covered only if at least one selected anchor represents it. The constraint (16) imposes the budget constraint that transforms the problem into a maximum coverage formulation with a cardinality constraint.

Since coverage alone does not fully capture the structural quality of the selection, the algorithm also computes a composite performance score

$$Q(B) = \beta_1 \,\mathrm{Cov}(B) + \beta_2 \,\mathrm{Div}_{\mathcal{T}}(B) + \beta_3 \,\mathrm{Div}_{\mathcal{S}}(B) + \beta_4 \,\mathrm{Div}_{\mathcal{G}}(B), \tag{17}$$

where $\mathrm{Cov}(B)$ is the coverage ratio achieved under budget B, $\mathrm{Div}_{\mathcal{T}}(B)$ measures the diversity of covered thematic clusters, $\mathrm{Div}_{\mathcal{S}}(B)$ measures the diversity of knowledge areas, and $\mathrm{Div}_{\mathcal{G}}(B)$ measures the diversity of SDGs. The coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ are non-negative and sum to one. This function summarizes the joint contribution of ontology and optimization into a single interpretable indicator.

The complete structure of the problem can therefore be interpreted as a search for strategic subsets within an enriched semantic graph. The solution does not consist of activating correspondences between records, but of constructing a computational ontological layer, deriving thematic partitions, projecting institutional relationships, and selecting, under explicit budget constraints, those anchors that maximize coverage, diversity, and institutional utility. In this sense, optimization complements ontology rather than replacing it. Ontology organizes the system's meaning; optimization identifies the most informative configuration within that system.

The methodology implemented translates this formulation into a reproducible procedure applied to the actual dataset. First, the records are read and validated in tabular format. Second, null values, inconsistencies, mixed formats, and missing columns are identified. Third, strings are cleaned, and delimiters are standardized. Fourth, a computational ontology is constructed in which each relevant column of the CSV is mapped to classes, attributes, and semantic relationships. Fifth, each record in the dataset is mapped to entities and links within the bibliometric graph. Sixth, semantic and bibliometric-derived variables are generated, including the number of authors per document, collaboration intensity, affiliation lists, inferred countries, unified keywords, knowledge fields, SDG tags, impact profiles, and standardized document types. Seventh, descriptive, relational, and thematic metrics are calculated, including co-authorship networks, institutional networks, co-occurrence matrices, thematic evolution, and relationships between fields and SDGs. Eighth, thematic clusters are induced using a vector representation of textual content and an unsupervised clustering procedure. Ninth, the problem of budgeted semantic coverage is formulated and solved using a greedy strategy and a multi-criteria evaluation of coverage and diversity. Tenth, the proposed governance-validated ontology-plus-optimization strategy is compared against three baseline strategies under the same budget levels. The first baseline is an ontology-only strategy, in which candidate anchors are retained after semantic and provenance validation but are not selected through maximum-coverage optimization. This baseline is treated only as an ablation of the optimization layer. The second baseline is a random-selection strategy, in which governance-valid anchors are randomly selected under the same budget constraint; this process is repeated over multiple runs and reported using mean values to reduce sampling variability. The third baseline is a frequency-based strategy, in which governance-valid anchors are ranked according to the number of associated records or semantic states and the highest-ranked anchors are selected until the budget is exhausted. Random selection and frequency-based selection are included as independent reference baselines, so that the evaluation does not rely only on comparing the proposed method with a reduced version of itself. The results and scientific figures are exported to PDF, along with interpretive notes oriented toward reproducibility and transparent evaluation.

The structure of the pseudocode corresponds directly to the Equations (1)–(17). In particular, the construction of the bibliometric graph is derived from (5), the budgeted coverage phase corresponds to (14)–(16), and the multi-criteria score is summarized in (17).

The final algorithm does not perform pairwise entity resolution or exhaustive search over all possible semantic configurations. Instead, it implements a hybrid strategy that integrates data cleaning, ontological modeling, thematic clustering, network analysis, provenance control, and budgeted maximum-coverage optimization. The ontological phase organizes the system's meaning; the thematic phase structures it into interpretable clusters; and the optimization phase selects the most informative subset of validated semantic anchors under explicit budget constraints. This combination is appropriate because it directly exploits the richness of the real-world dataset, avoids reducing the analysis to isolated counts, and preserves a clear translation between the semantic representation, the objective function, and the computational procedure.

From a methodological standpoint, the model's utility lies not solely in describing production, citations, or collaboration. The proposed solution is designed to maximize semantic coverage, structural diversity, and institutional utility. Therefore, in addition to classical descriptive metrics, the system must be evaluated using incidence graphs, thematic heatmaps, multi-curves of temporal evolution, diversity metrics, and ontology-assisted performance profiles for optimization. This methodological decision is consistent with the nature of (14)–(17), where the goal is not merely to represent the corpus, but to construct an analytical foundation that can be reused in research, institutional evaluation, thematic prioritization, and decision-making under explicit budget constraints.

4. Analysis of Results

The results confirm that the proposed architecture does not function as an isolated descriptive framework, but rather as a hybrid infrastructure that integrates ontological representation, bibliometric structuring, thematic clustering, and budget optimization. In this formulation, the quality of the system does not depend solely on counting documents or citations, but on the ability to organize scientific knowledge into a structure that is interpretable, coherent, and useful for decision-making. Under this criterion, the algorithm fulfills the study's central objective by transforming a real bibliographic corpus into a semantic database structured around knowledge areas, sustainable development goals, thematic relationships, institutional links, and optimized coverage profiles.

The empirical validation was conducted using a Scopus-derived institutional corpus comprising 3558 bibliographic records from a single university. The final analytical dataset, identified as `dataset_v4.csv`, preserves the metadata required to reproduce the semantic and optimization workflow, including record identifiers, titles, first authors, publication years, journals, publishers, DOIs, open-access status, source URLs, thematic labels, source names, abstracts, keywords, affiliations, Scopus EIDs, document types, conference names, ISSNs, ISBNs, and citation counts. The true entity identifier was retained exclusively as a controlled reference field for internal validation and was not used to disclose personal or institutional identities in the reported visualizations. These variables support bibliometric normalization, thematic clustering, knowledge-area and SDG mapping, provenance construction, scholarly knowledge graph generation, semantic-anchor definition, and budgeted maximum-coverage optimization. Because the evidence is derived from one institutional corpus, the experiment is intended to demonstrate methodological feasibility, internal reproducibility, and decision-support value within the analyzed university context rather than unrestricted generalizability across higher education systems. The computational workflow was executed in Python 3.12.7 using the reproducible environment and dependency specifications provided with the implementation package.

The effect of governance validation on the candidate semantic anchors is summarized in Table 3.

Table 3. Governance validation effect on candidate semantic anchors before optimization.

Anchor Category	Candidate	Gov.-Valid	Filtered
Thematic clusters	8	8	0
Knowledge areas	8	8	0
SDG labels	13	13	0
Document types	11	11	0
Countries/collaboration states	42	42	0
Publication venues	1708	1704	4
Normalized keyword groups	23,659	4443	19,216
Author/name entities	2093	2091	2
Total	27,542	8320	19,222

Note: Gov.-valid = governance-valid. Governance-valid anchors satisfy provenance, normalization, and auditability requirements before optimization. Publication venues were validated using source traceability and bibliographic identifiers. Normalized keyword groups were retained when they met the minimum recurrence and semantic-consistency criteria required for reproducible anchor selection. The governance layer reduced the optimization search space from 27,542 candidate anchors to 8320 validated anchors, filtering 19,222 anchors, equivalent to 69.8% of the initial candidate set.

The governance-validation layer filtered 19,222 of the 27,542 candidate semantic anchors, corresponding to 69.8% of the initial anchor set. This result provides the experimental evidence for the governance-aware component of the framework. Most exclusions were concentrated in normalized keyword groups, where low-frequency, weakly supported, or semantically unstable terms did not satisfy the minimum recurrence and semantic-consistency requirements for reproducible anchor selection. Publication venues showed four exclusions associated with incomplete bibliographic traceability, and author/name entities showed two exclusions due to non-canonical or placeholder-like forms. In contrast, thematic clusters, knowledge areas, SDG labels, document types, and countries or collaboration states were fully retained because they were derived from complete source fields and reproducible semantic-mapping rules. Therefore, governance awareness is not only described mathematically through the provenance graph and validation constraint; it also changes the experiment by reducing the feasible anchor space before optimization. The optimization model is consequently prevented from selecting anchors lacking sufficient provenance, normalization, or auditability support.

Figure 2 shows the joint evolution of the coverage ratio, composite objective score, and ontology-assisted gain across the evaluated budget levels. Coverage increases rapidly and approaches saturation, whereas both the objective score and ontology-assisted gain reach their highest values within an intermediate budget range before declining moderately at larger budgets. Therefore, the three quantities do not increase simultaneously across the entire budget domain. Instead, the result shows that the ontology structures the semantic decision space and enables the optimization layer to identify informative anchor configurations while preserving a favorable balance between coverage and diversity. This behavior provides direct empirical support for the integration of semantic organization, governance validation, and budgeted optimization.

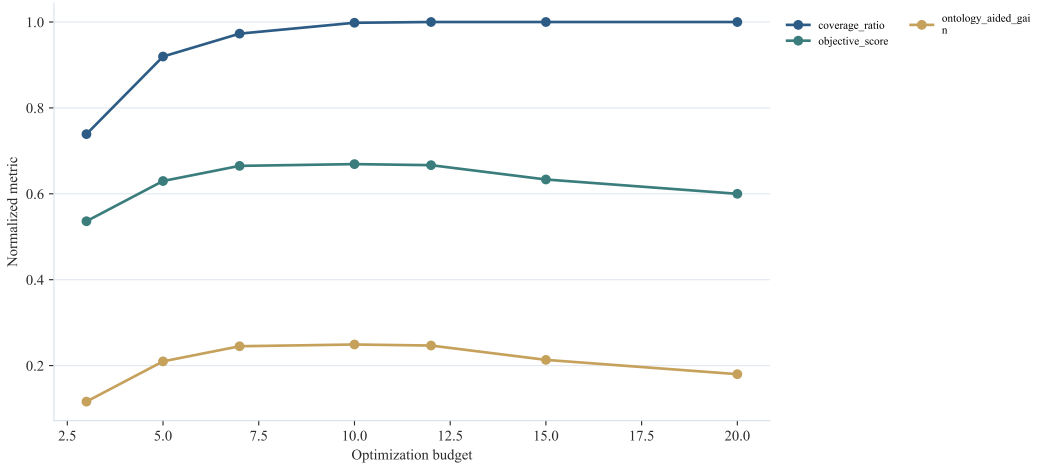


Figure 2. Joint curve of coverage, objective score, and ontology-assisted gain as a function of the optimization budget.

A broader baseline comparison is reported in Table 4. To address the risk of evaluating the method only against an ablated version of itself, the comparison separates the ontology-only configuration from two independent reference baselines. Random selection tests whether the observed gain exceeds arbitrary governance-valid anchor selection, whereas frequency-based selection tests whether the proposed model improves over a simple bibliometric heuristic that prioritizes the most frequent anchors.

Table 4. Baseline comparison under the same constrained anchor-selection budget.

Strategy	Coverage	Topic Div.	Area Div.	SDG Div.	Composite Score
Random selection	0.016	0.463	0.373	0.364	0.304
Frequency-based selection	0.799	1.000	1.000	1.000	0.950
Ontology-only selection	0.913	1.000	1.000	1.000	0.978
Proposed GAGER	0.955	1.000	1.000	1.000	0.989

Note: Div. = diversity. All strategies were evaluated under the same constrained selection budget. Random selection reports the mean value over repeated random trials. Proposed GAGER corresponds to the governance-validated ontology-plus-optimization strategy. It selects governance-valid anchors according to their marginal contribution to the semantic-state universe, while frequency-based selection prioritizes anchors with the largest number of associated records and ontology-only selection retains semantic anchors without marginal-coverage optimization.

Table 4 directly addresses the baseline concern by comparing the proposed strategy with both an ablated configuration and independent reference methods under the same constrained budget. The ontology-only strategy is an internal ablation, whereas random selection and frequency-based selection are external decision rules that do not use marginal-coverage optimization. Proposed GAGER achieved the highest semantic coverage and composite score. The gain over random selection shows that the result is not explained by arbitrary anchor choice. The gain over frequency-based selection shows that prioritizing marginal contribution to the semantic-state universe is more effective than selecting only the most frequent anchors. The gain over ontology-only selection shows that the optimization layer adds decision value beyond semantic representation alone. Therefore, the quantitative evaluation no longer rests on a comparison between the proposed method and only a reduced version of itself.

Figure 3 is retained as a complementary ablation analysis between the ontology-only configuration and the ontology-plus-optimization configuration. However, the main

baseline evidence is provided by Table 4, because it includes independent random-selection and frequency-based references under the same constrained budget.

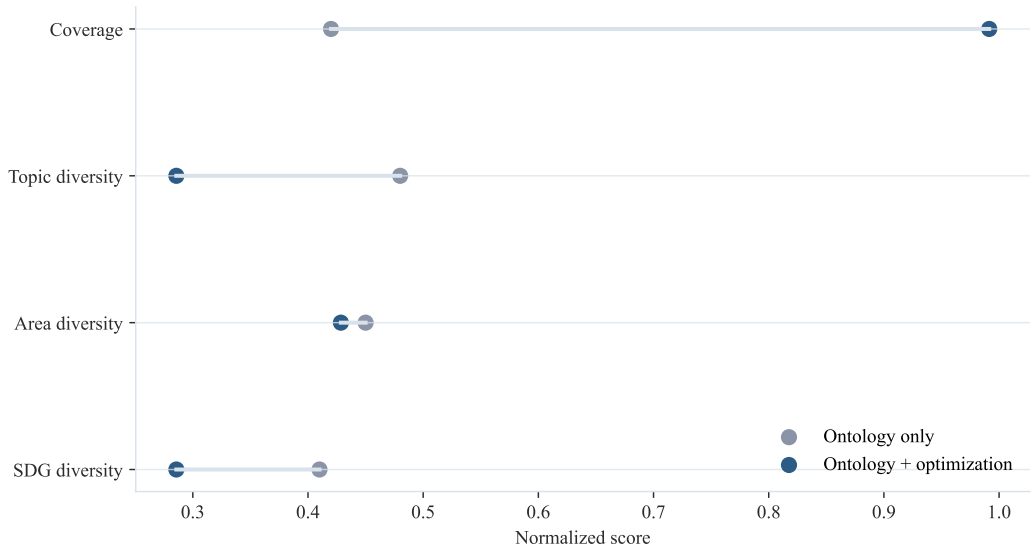


Figure 3. Comparison between the ontology-only configuration and the ontology-plus-optimization configuration for coverage and structural diversity.

The optimizer's response to different budgets is detailed in Figure 4. Coverage increases sharply at low budget levels and then approaches saturation, whereas the diversity-related metrics follow non-monotonic trajectories. Topic diversity decreases as the budget increases, area diversity reaches an intermediate peak and then declines, and SDG diversity improves at early budget levels before stabilizing or slightly decreasing. These results indicate that larger budgets improve semantic coverage, but do not produce uniform gains across all diversity dimensions.

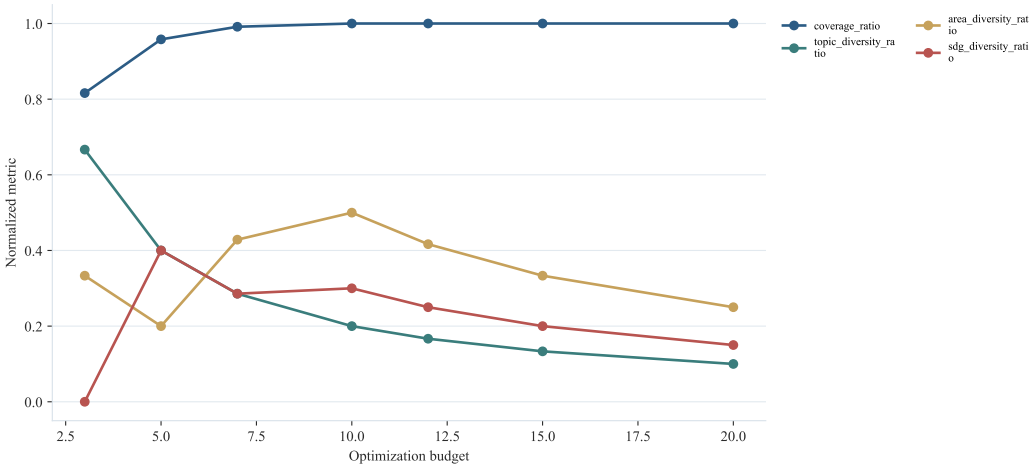


Figure 4. Evolution of coverage and semantic diversity as a function of the optimization budget.

Figure 5 shows that the composite objective score does not improve monotonically with the optimization budget. Instead, it rises sharply from the lowest budget levels, reaches its highest value within an intermediate operating range, and then gradually declines as additional semantic anchors are incorporated. This behavior indicates that the framework does not simply reward larger selections. Rather, it identifies a moderate budget regime in which semantic coverage and structural diversity achieve their most effective balance. From a methodological standpoint, the result confirms that the optimization layer

contributes genuine decision value by regulating the trade-off between representativeness and structural coherence.

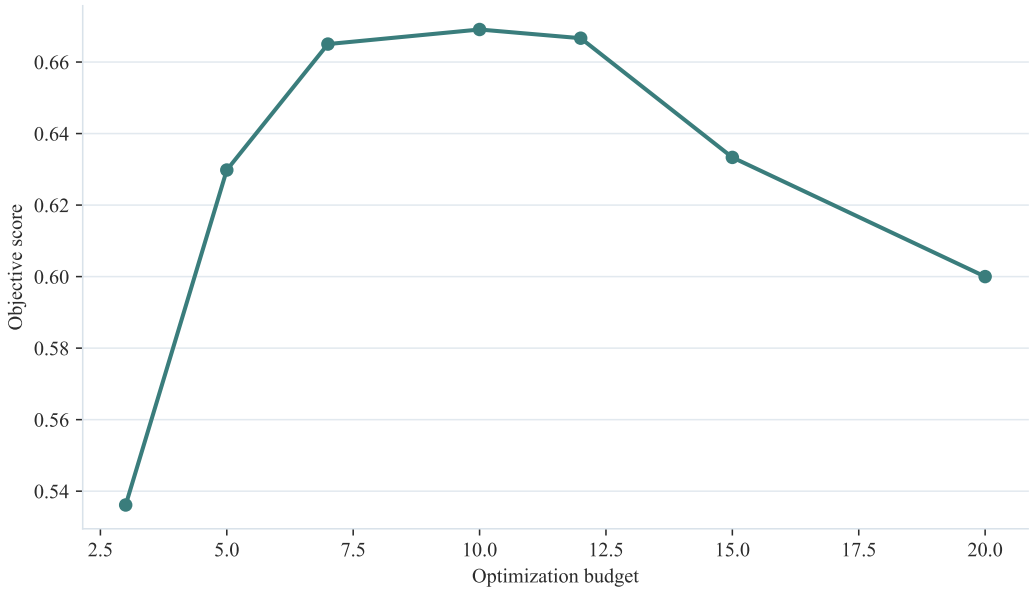


Figure 5. Evolution of the composite objective function as a function of the optimization budget.

From a temporal perspective, Figure 6 reveals that the corpus's evolution is not homogeneous across thematic clusters. The highest intensity is concentrated in a dominant cluster during the recent period, accompanied by a second thematic group showing sustained growth and several secondary clusters with a more moderate presence. This pattern indicates that recent scientific output tends to concentrate on a few high-density thematic axes, while others remain as more limited lines of specialization. The figure is relevant because it connects the structural analysis of the corpus with the logic of strategic selection; a coverage policy that ignores this thematic concentration would lose explanatory power regarding the system's actual dynamics.

Figure 7 shows that the proposed framework captures differentiated relationships between institutional knowledge areas and Sustainable Development Goal tags. The heat map indicates that Education is strongly concentrated in SDG 4, Engineering and ICT exhibit higher incidence in SDG 7 and the unassigned SDG band, Health and Wellbeing is more visibly associated with SDG 3, and Interdisciplinary output reaches its highest intensity in the unassigned SDG region. Additional but lower-intensity links are also visible for areas such as Business and Economics, Environment and Sustainability, and Architecture and Urban Studies across several SDG categories. This result demonstrates that the ontology layer does not merely annotate isolated records, but organizes the corpus into interpretable semantic structures connecting disciplinary organization with development-oriented impact dimensions. The resulting scholarly knowledge graph is therefore both computationally structured and analytically meaningful for institutional research governance.

The evolution across knowledge areas is shown in Figure 8. Engineering and ICT clearly lead the recent growth of the corpus, followed by Interdisciplinary and Education, while the other areas exhibit more moderate trajectories. This pattern is methodologically valuable because it confirms that production is not distributed symmetrically and that a few high-intensity axes dominate the system's thematic structure. The figure also helps explain why certain areas have a greater impact on the optimization process; not all contribute equally to the coverage of the semantic universe.

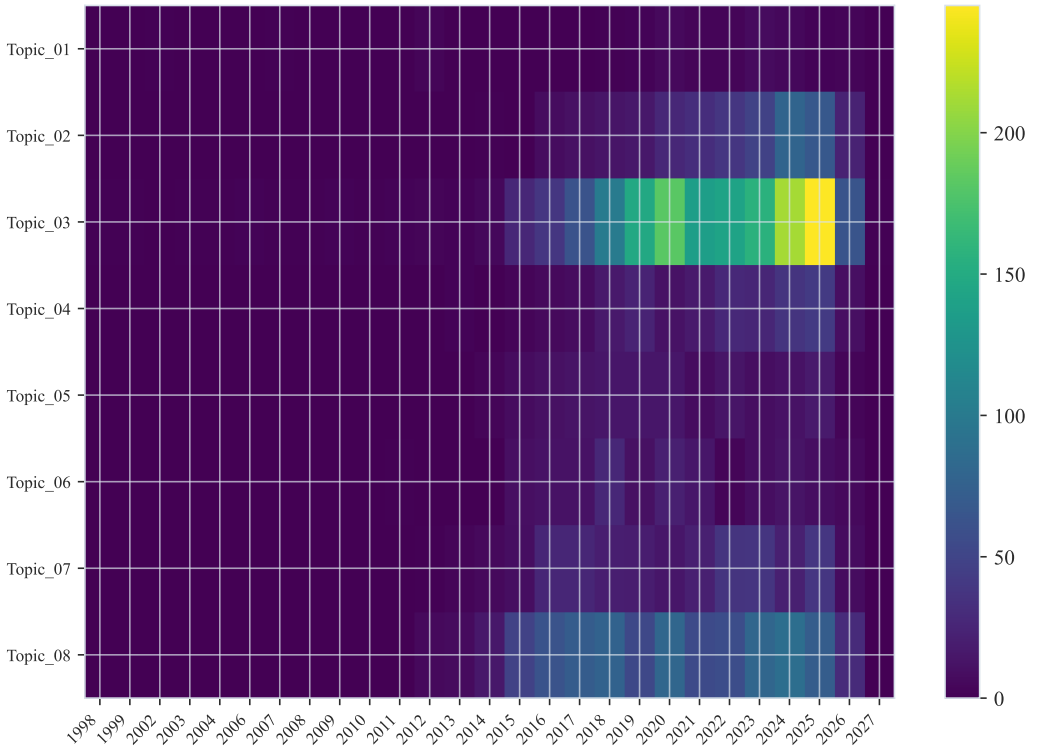


Figure 6. Heat map of the temporal evolution of the corpus's thematic clusters.

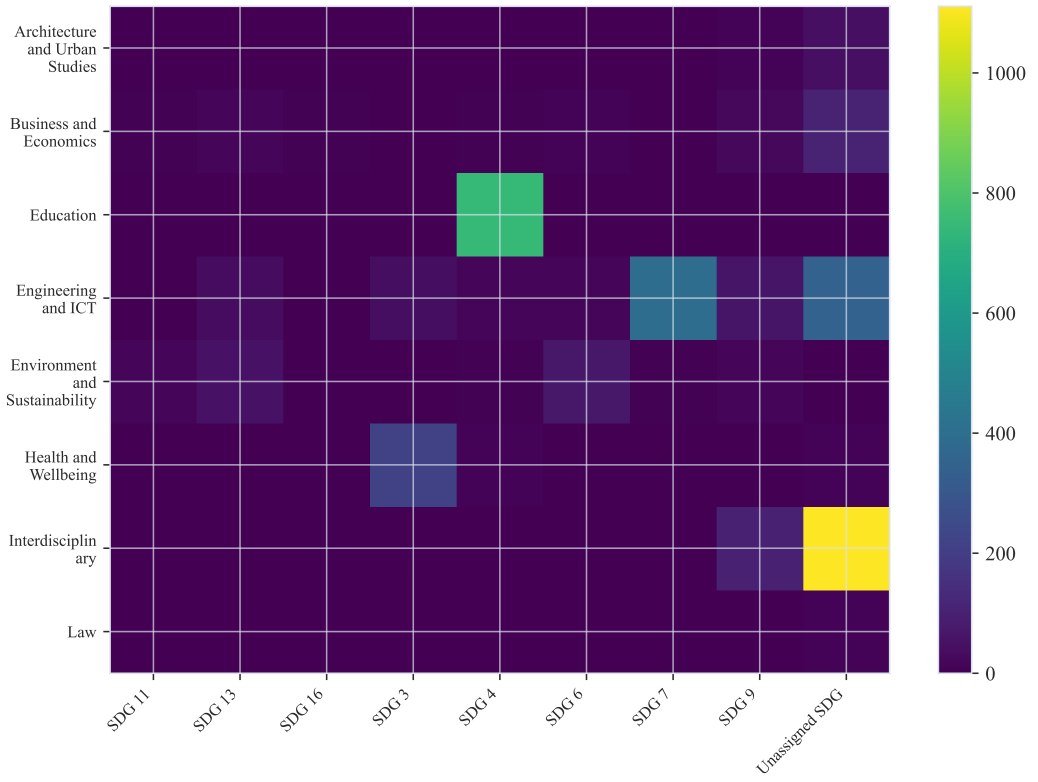


Figure 7. Heat map of the relationship between fields of knowledge according to the knowledge criteria and SDG tags.

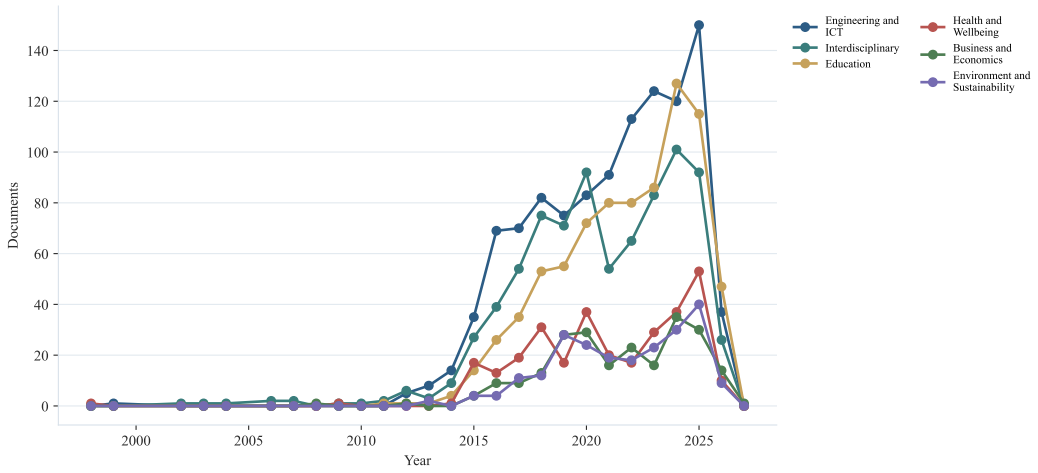


Figure 8. Temporal evolution of scholarly output by knowledge areas.

The temporal impact dimension is reinforced in Figure 9. The citation curve shows that SDG 4 accounts for the largest recent historical volume, followed by the unassigned category and, at a greater distance, by SDG 7, SDG 3, and SDG 9. This distribution reveals that citation intensity does not exactly replicate production intensity but rather generates a differentiated impact map. Therefore, the system's contribution is not limited to showing what is published, but also which lines achieve greater academic resonance under the SDG filter.

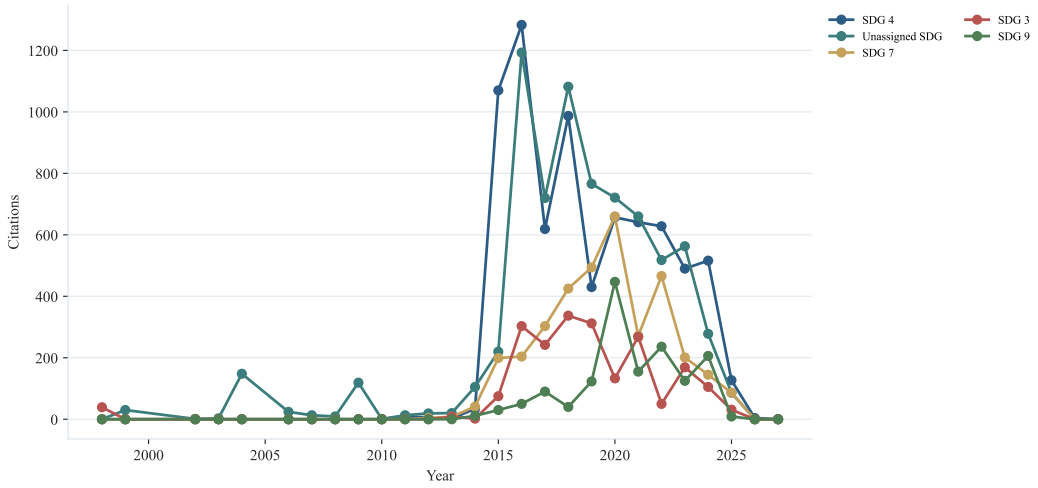


Figure 9. Temporal evolution of citations aggregated by SDG tags.

The relational structure of the corpus is made visible in Figure 10. The ontological semantic graph integrates knowledge domains, SDGs, document types, and high-impact concepts into a single representation. The central position of Article and Conference Paper, along with Education, Engineering, and ICT, and various SDG tags, indicates that the system successfully articulates the documentary, thematic, and impact dimensions simultaneously. This figure is particularly relevant because it synthesizes, in a single structure, the rationale behind the proposed approach, then transforms the original dataset into a navigable, structurally interpretable knowledge space.

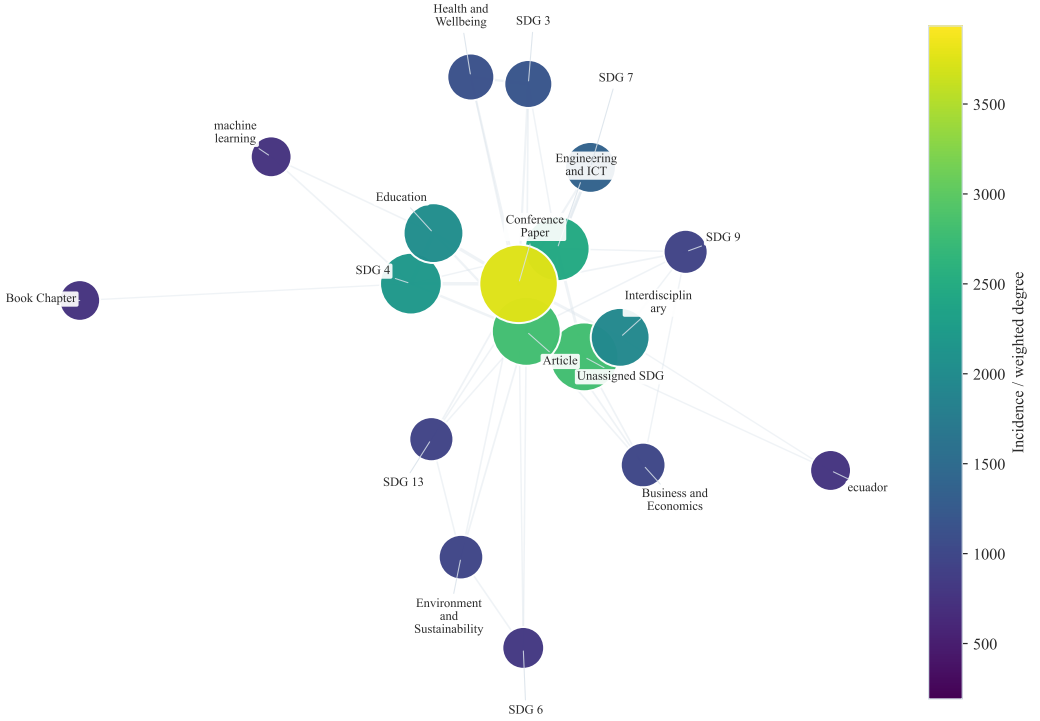


Figure 10. Ontological semantic graph integrating knowledge areas, SDG tags, document types, and high-frequency concepts.

The term co-occurrence network, presented in Figure 11, reinforces the semantic consistency of the corpus. Nodes such as artificial intelligence, machine learning, education, engineering education, optimization, IoT, and fault detection appear as relevant hubs within the graph. This structure confirms that the thematic clusters are not arbitrary artifacts of the algorithm but rather emerge from a technical vocabulary with recognizable, stable relationships. From a methodological standpoint, this strengthens the validity of the ontological layer and justifies its use as a basis for subsequent optimization.

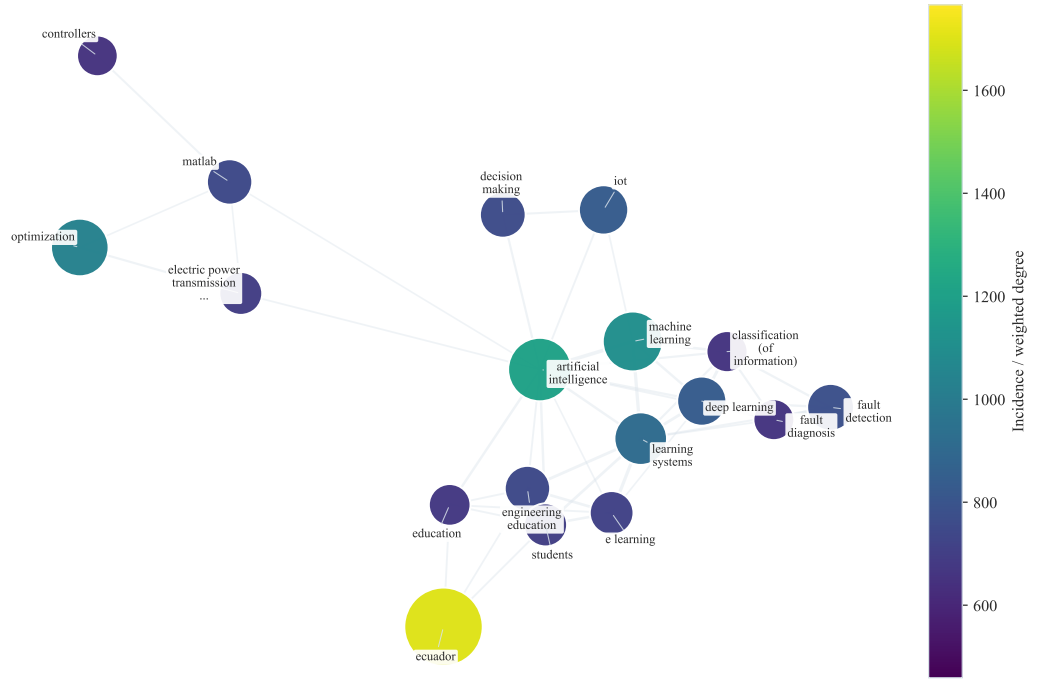


Figure 11. Co-occurrence network of high-frequency keywords in the corpus.

The connection between dissemination spaces and thematic structures is shown in Figure 12. Conferences do not appear as isolated nodes, but rather as channels linked to specific thematic clusters. It indicates that output at scientific events is not distributed randomly, but rather is articulated along relatively consistent thematic lines. This property is important because it makes conferences an analytical component of the graph, rather than mere editorial metadata.

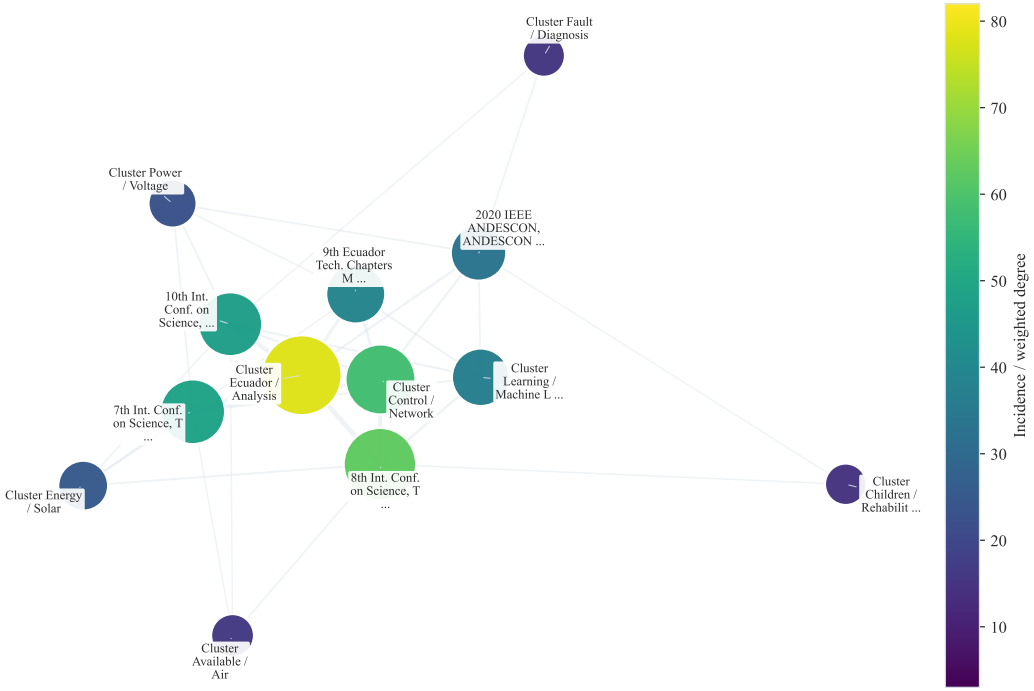


Figure 12. Network of relationships between conferences and inferred thematic clusters.

The network visualizations were generated directly in Python using NetworkX and Matplotlib. VOSviewer software was neither used nor embedded in the computational workflow; the figures only adopt a VOSviewer-like analytical presentation style.

Figure 13 offers an additional overview of the system's macrostructural organization by depicting the interaction between knowledge areas and high-frequency SDG labels. The network confirms that Engineering and ICT, Education, and Interdisciplinary occupy a central position within the corpus topology, while other areas play a more peripheral role. This distribution is consistent with the temporal curves and previous heat maps, and reinforces the interpretation that the analyzed scientific system is organized around a few nuclei of high semantic connectivity.

The geographical dimension is shown in Figure 14, where Ecuador is the main node in the country-based collaboration network, followed by Spain, Colombia, Mexico, and other countries with lower relative incidence. This figure shows that the resulting ontological infrastructure preserves the capacity for territorial interpretation and international collaboration.

Figures 15 and 16 provide a complementary view of the system. The first shows the recent dominance of Conference Papers and Articles over other document types, while the second demonstrates that the trajectories of single-institution, inter-institutional, and international collaboration grow in parallel, with varying intensities across periods. These figures complement the semantic and optimization results by characterizing the documentary and collaboration structure of the corpus on which the proposed framework operates.

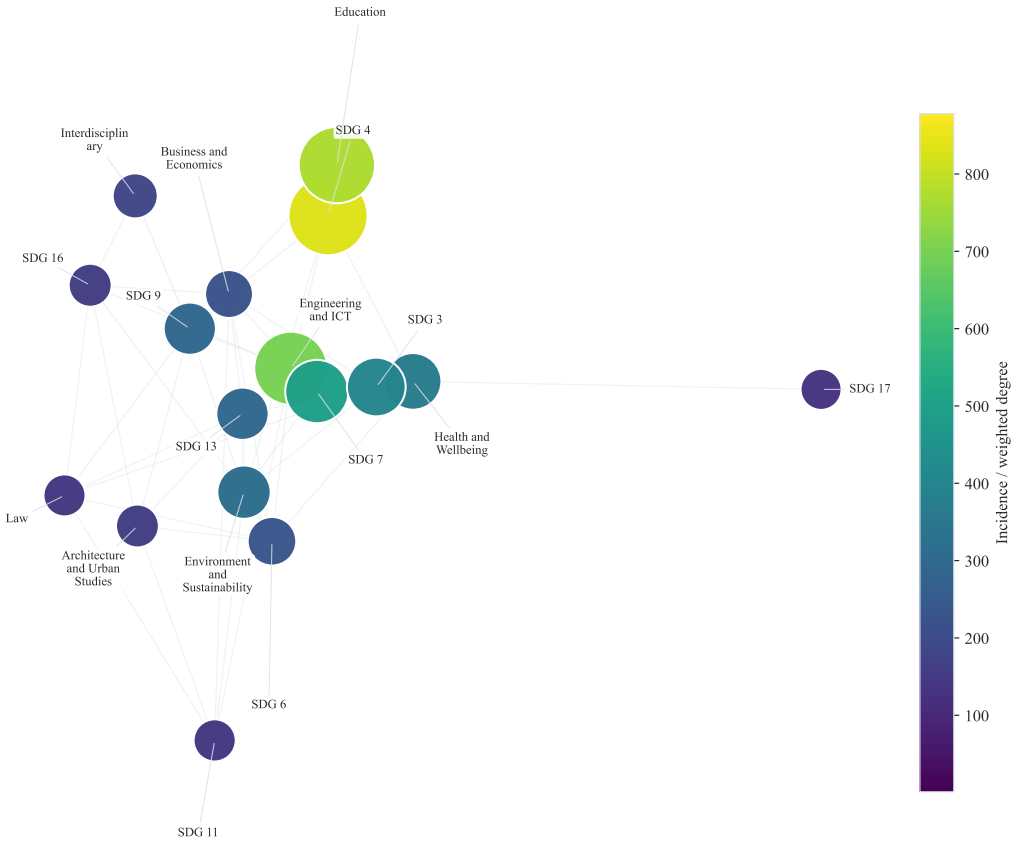


Figure 13. Network Python-generated network of interaction between knowledge areas and high-incidence SDG tags.

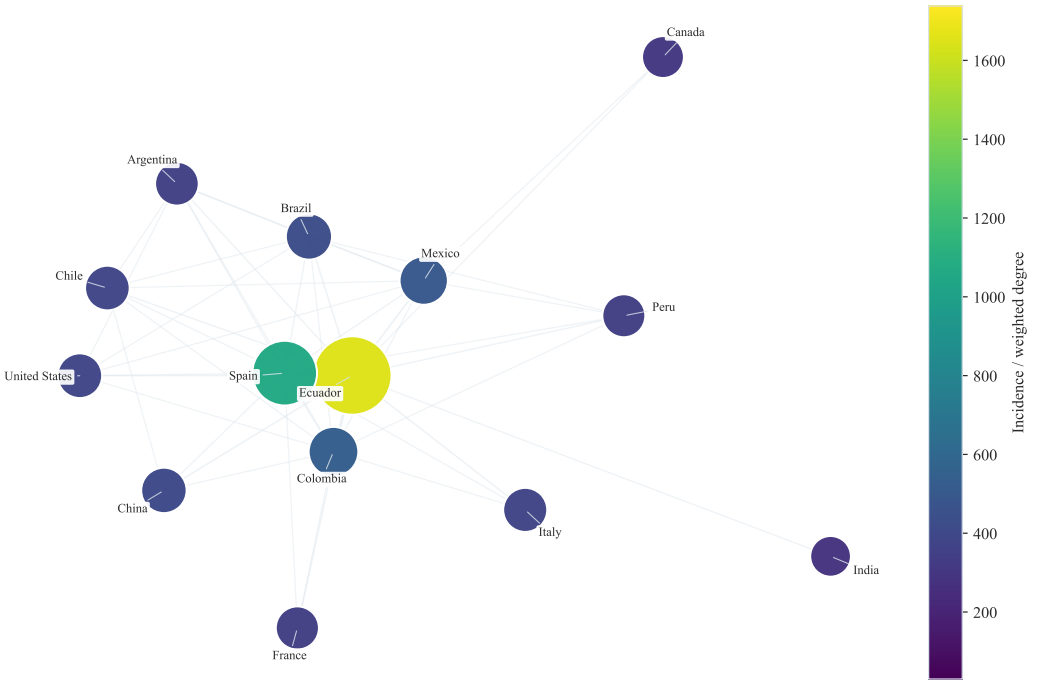


Figure 14. Aggregated international collaboration network by country.

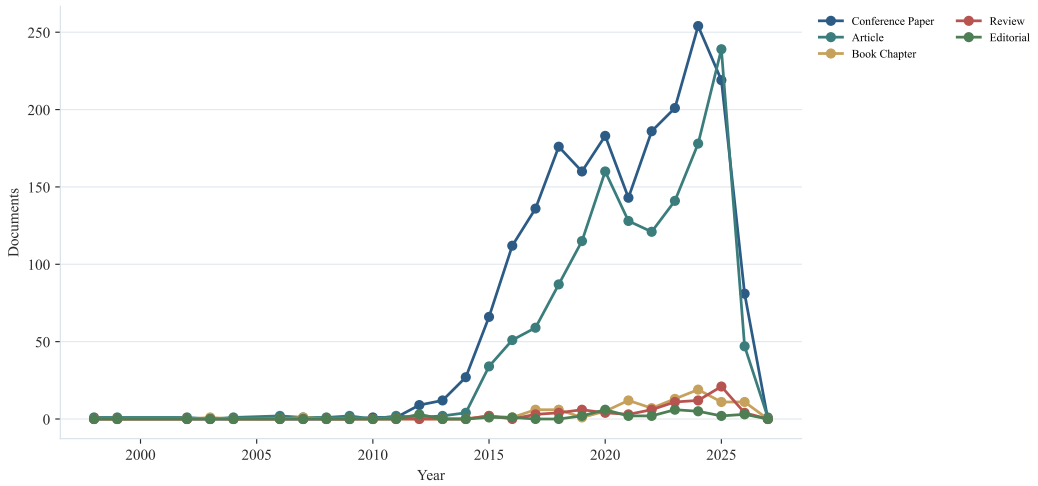


Figure 15. Temporal evolution of production by document type.

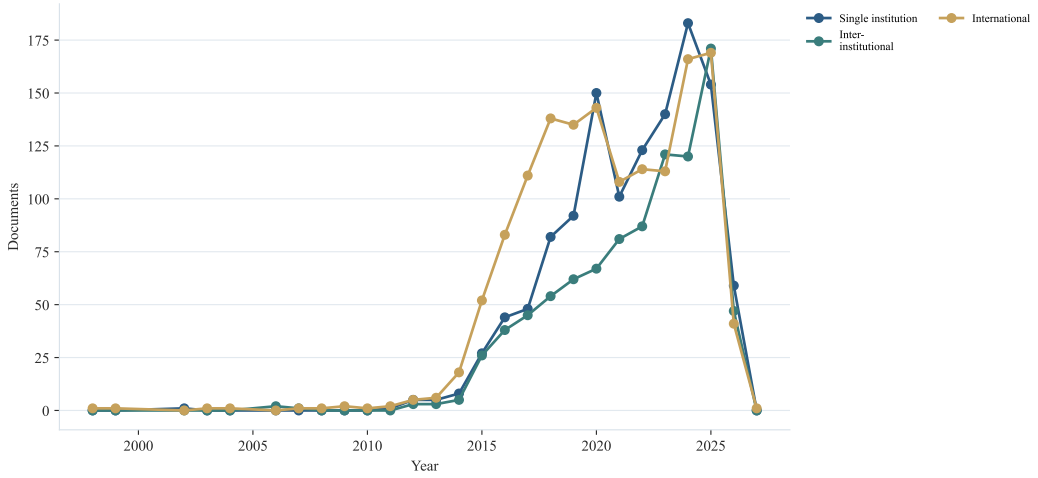


Figure 16. Temporal evolution of the scope of institutional and international collaboration.

Figures 17 and 18 serve a complementary function in characterizing the corpus. The first relates productivity and impact by source, showing that document volume does not always coincide with higher citation intensity. The second summarizes the distribution of documents by conference and confirms the relative concentration in a small set of events. Both figures support the descriptive validation of the corpus and provide additional evidence about the editorial and citation structure on which the ontological layer is constructed.

Overall, the evidence shows that the proposal supports the study's hypothesis across semantic organization, thematic alignment, structural coverage, and utility for knowledge governance. The model's main strength does not lie in a single isolated metric, but rather in the articulation between ontological representation, thematic clustering, relational analysis, and budgeted optimization. This combination transforms the bibliometric corpus into a knowledge infrastructure that is more interpretable, more traceable, and more useful for institutional analytics and strategic decision-making.

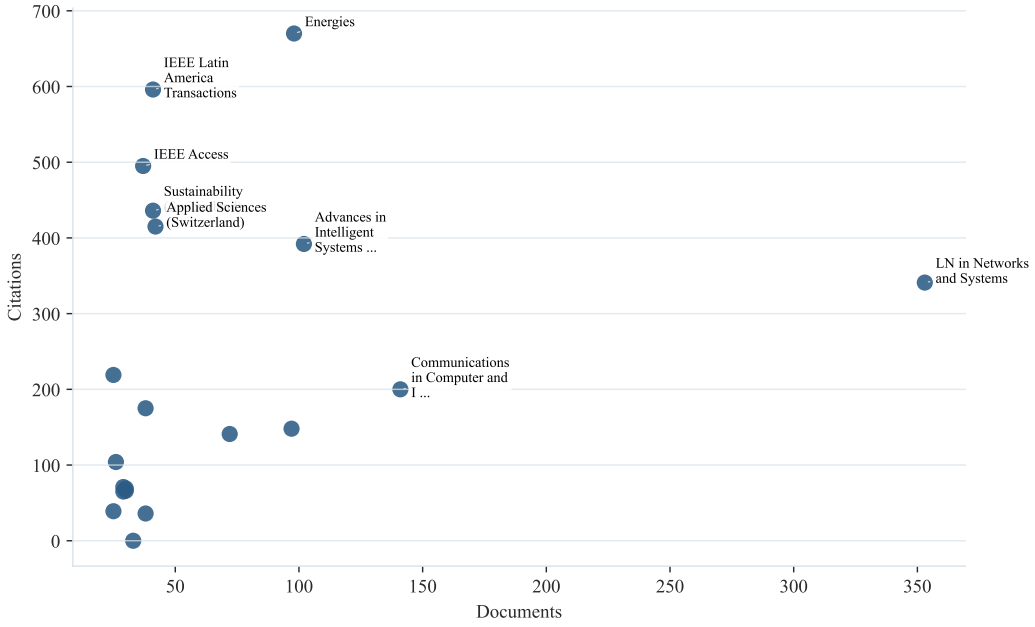


Figure 17. Relationship between document productivity and citation impact in the main sources of the corpus.

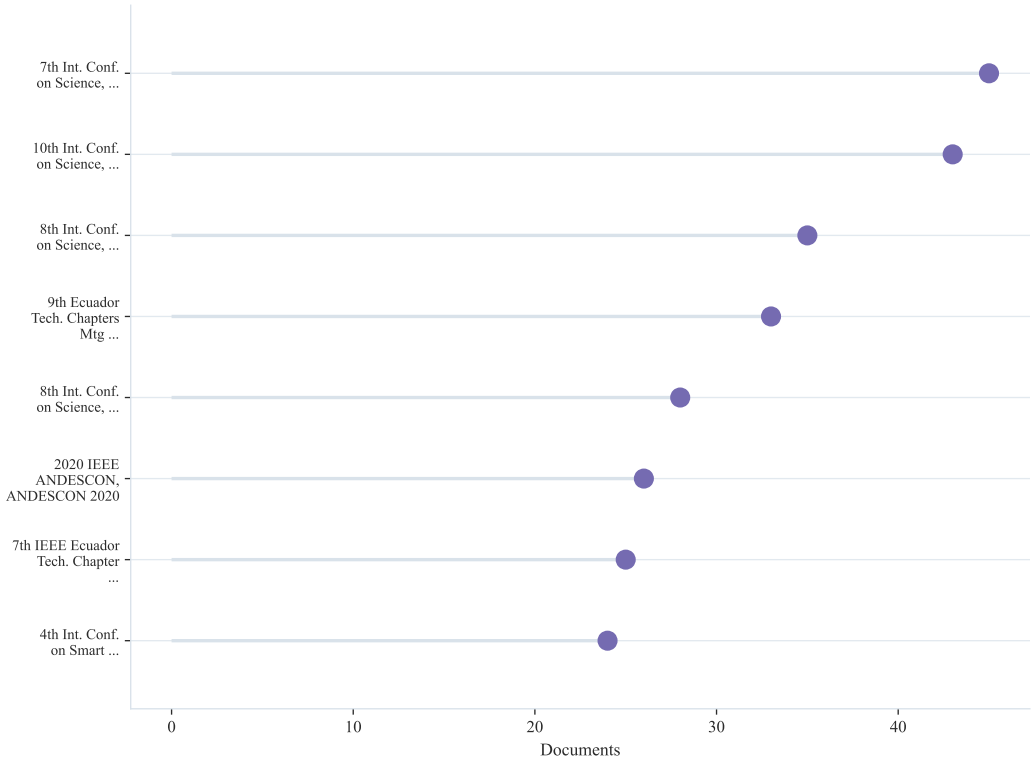


Figure 18. Distributionof documents across the conferences with the greatest impact on the corpus.

Table 5 clarifies the main methodological boundaries of the proposed framework and makes explicit the interpretive scope within which the reported results should be understood. The table is intended to strengthen the transparency of the revised manuscript by distinguishing between what is already demonstrated by the current institutional case study and what remains outside the empirical scope of the present contribution. In particular, it states that the greedy optimization layer should be interpreted as an approximate yet computationally viable strategy for the budgeted semantic coverage problem, rather than as an exact global solver. It also makes explicit that the evidence is restricted to a

single-university Scopus-based corpus and therefore supports a reproducible institutional proof of concept, not an unrestricted claim of generalizability across all higher education systems. In addition, the table reports that objective-weight sensitivity was examined through 20 alternative weighting configurations, clarifies that broader performance benchmarking against external bibliometric platforms remains beyond the current scope, and acknowledges that the temporary restriction on full code release affects immediate external replication even though the methodological pipeline and optimization logic are described in detail.

Table 5. Summary of the methodological concerns explicitly delimited in the revised manuscript.

Methodological Aspect	Interpretive Issue	Clarification Incorporated into the Manuscript
Greedy optimization strategy	The optimization layer relies on a greedy decision rule and therefore does not guarantee an exact global optimum.	The text now states explicitly that the greedy procedure provides an approximate yet computationally viable solution to the budgeted semantic coverage problem, and this limitation is discussed more clearly in the Discussion and Conclusions sections.
Sensitivity to objective-function weights	The relative influence of coverage and diversity terms may substantially affect the preferred operating configuration and the resulting anchor set.	A systematic sensitivity analysis was conducted over 20 alternative weight vectors. Table 6 reports how changes in the relative importance of coverage and diversity affect the preferred budget, achieved coverage, diversity profile, and stability of the selected semantic-anchor set.
Restricted empirical scope	The validation is based on a single-university corpus and therefore cannot sustain unrestricted cross-institutional claims.	The revised manuscript explicitly delimits the scope of applicability and presents the study as a reproducible institutional case study designed to demonstrate methodological feasibility and analytical value within the analyzed university context.
External benchmarking	The framework is not benchmarked against external bibliometric platforms or large-scale institutional analytics systems.	The text now clarifies that the empirical comparison is performed against constrained internal baselines under the same budget regime, while broader external benchmarking is left for future research.
The ongoing intellectual property protection procedure temporarily restricts immediate full replication	Full code release.	The manuscript states more explicitly that this restriction affects immediate full replication, while the methodological pipeline, variables, and optimization logic are reported in sufficient detail to support structured future replication once release becomes possible.

To assess the robustness of the preferred operating configuration, a sensitivity analysis was conducted over 20 alternative weight vectors for the composite objective function. The analysis varied the relative importance assigned to semantic coverage, thematic diversity, knowledge-area diversity, and SDG diversity. For each weight vector, all budget levels from 1 to 20 were evaluated, and the budget maximizing the corresponding composite objective was retained. Table 6 presents the four predefined configurations and selected representative scenarios from the additional weight vectors. The analysis evaluates how

the weighting assumptions affect the preferred budget, achieved coverage, diversity profile, and stability of the resulting semantic-anchor set.

Table 6. Sensitivity of the preferred operating configuration to alternative objective-function weights.

Scenario	β_1	β_2	β_3	β_4	Best Budget	Coverage	Topic Div.	Area Div.	SDG Div.	Jaccard Stability
Baseline	0.500	0.200	0.150	0.150	10	0.998	0.400	0.300	0.300	0.833
Equal weights	0.250	0.250	0.250	0.250	12	1.000	0.333	0.417	0.250	1.000
Coverage-oriented	0.700	0.100	0.100	0.100	12	1.000	0.333	0.417	0.250	1.000
Topic-oriented	0.400	0.300	0.150	0.150	7	0.973	0.571	0.143	0.286	0.583
Area-oriented	0.325	0.131	0.373	0.171	12	1.000	0.333	0.417	0.250	1.000
Balanced area–SDG	0.209	0.078	0.349	0.364	2	0.582	0.000	0.500	0.500	0.167
Mixed diversity-oriented	0.273	0.036	0.304	0.387	4	0.850	0.250	0.250	0.500	0.333
SDG-dominant	0.125	0.012	0.229	0.634	1	0.370	0.000	0.000	1.000	0.083

Note: β_1, β_2, β_3, and β_4 correspond to semantic coverage, thematic diversity, knowledge-area diversity, and SDG diversity, respectively. Jaccard stability measures the similarity between each selected-anchor set and the reference anchor configuration used in the sensitivity analysis. Values are rounded to three decimal places.

The sensitivity results show that the weighting assumptions influence both the preferred operating budget and the semantic composition of the selected solution. The baseline configuration achieves a coverage ratio of 0.998 with a budget of 10, whereas the equal-weight, coverage-oriented, and area-oriented configurations select a budget of 12 and achieve complete semantic coverage. The topic-oriented scenario favors a smaller budget of 7 and increases the relative participation of thematic anchors. In contrast, strongly specialized area–SDG and SDG-dominant configurations select substantially smaller budgets and sacrifice overall coverage to emphasize a specific diversity dimension. The Jaccard results further show that balanced and coverage-oriented settings retain highly similar anchor configurations, whereas strongly specialized weighting preferences generate substantially different selections. These findings demonstrate that the model is stable under balanced and coverage-oriented preferences but becomes deliberately more specialized when one criterion dominates the objective. Consequently, the baseline weights should be interpreted as explicit institutional decision preferences rather than as universally optimal coefficients.

5. Discussion

The main contribution of this work is to shift the analysis of academic knowledge from a predominantly descriptive or representational approach to a hybrid approach that combines semantic organization with structurally optimized decision-making. Ontology-based proposals have demonstrated undeniable value by incorporating explicit semantics, provenance, evidence, traceability mechanisms, and auditable structures to model entities, attributes, and relationships within academic graphs. In this sense, such approaches have helped transform knowledge management from an opaque operation into an interpretable architecture compatible with institutional control requirements.

The comparison with established bibliometric tools requires a clear distinction between performance benchmarking and functional scope. VOSviewer and Bibliometrix provide mature capabilities for bibliometric mapping, network analysis, clustering, thematic exploration, and descriptive evaluation; however, they do not formulate the governance-constrained selection of semantic anchors as a budgeted maximum-coverage problem. Therefore, they cannot be treated as directly interchangeable optimization baselines. As

summarized in Table 7, their inclusion serves to delimit the analytical novelty and scope of GAGER, whereas the quantitative evaluation of the proposed method is performed against random selection, frequency-based selection, and ontology-only selection under the same budget conditions. This distinction avoids an inappropriate performance comparison between systems designed for different purposes while demonstrating that GAGER adds provenance control, governance validation, explicit resource constraints, and semantic-anchor optimization to conventional bibliometric analysis.

Table 7. Functional comparison of GAGER, experimental reference strategies, and established bibliometric-analysis tools.

Tool or Strategy	Primary Analytical Function	Ontology-Based SKG	Provenance and Governance	Explicit Budget	Semantic-Anchor Optimization
VOSviewer	Construction, clustering, visualization, and exploration of bibliometric networks, including co-authorship, co-citation, bibliographic coupling, and term co-occurrence	No	No	No	No
Bibliometrix	Descriptive bibliometric analysis, performance indicators, co-citation, co-authorship, keyword co-occurrence, thematic mapping, and collaboration analysis	No	No	No	No
Random selection	Selection of governance-valid anchors without frequency ranking or marginal-coverage information	No	Yes	Yes	No
Frequency-based selection	Ranking and selection of governance-valid anchors according to their number of associated records or semantic states	No	Yes	Yes	No
Ontology-only selection	Retention of ontology-derived and governance-valid anchors without marginal-coverage optimization	Yes	Yes	Yes	No
Proposed GAGER	Construction of a governance-aware scholarly knowledge graph and selection of governance-valid semantic anchors by maximizing coverage and diversity under explicit resource constraints	Yes	Yes	Yes	Yes

Note: SKG = scholarly knowledge graph. The comparison with VOSviewer and Bibliometrix is functional rather than performance-based because these tools do not solve the same governance-constrained budgeted semantic-anchor selection problem. Random, frequency-based, and ontology-only selection are the experimental reference strategies evaluated under the same budget conditions as GAGER.

However, the results of this research show that, while conceptually robust, a purely ontological approach does not, on its own, resolve the strategic dimension of the problem when the corpus grows in scale, heterogeneity, and thematic complexity. Ontology allows for the precise representation of what a thematic relationship, an alignment with the SDGs, an association between a knowledge area and a document type, or a connection between institutional actors means. However, it does not in itself determine which regions of that semantic space should be prioritized when analytical resources are finite. In real-world academic analytics contexts, the question is no longer merely how to represent

knowledge, but which subset of the system to select to maximize coverage, diversity, and interpretability.

The innovation of this proposal is to maintain the semantic and governance layers while adding a combinatorial optimization layer that can make decisions across the entire space of candidate semantic anchors. This difference is substantial. While an ontological architecture organizes, documents, and connects scientific information, the model proposed here also selects those structures under explicit constraints of budget, coverage, and diversity. In other words, it not only represents the semantics of the corpus but also optimizes the analytical use of that semantics. This transition changes the nature of the problem. The question is no longer merely how to describe the scientific ecosystem but rather which configuration of the ecosystem should be prioritized so that the resulting system is more coherent, more interpretable, and more useful for academic analytics and institutional decision-making.

Within this framework, the contrast with related approaches is particularly revealing. Ontological proposals provide a robust mechanism for organizing entities, attributes, and relationships, and enable the construction of auditable processes for knowledge representation. Their strength lies in explainability and traceability. The strength of this work lies on another level; therefore, the ability to coordinate these semantic structures with an explicit logic of strategic selection. The results show that this coordination produces visible improvements in coverage, thematic diversity, field diversity, and SDG-related diversity, while maintaining an interpretable semantic foundation. This observation is important because it avoids a simplistic comparison. The advance does not consist in replacing the ontology with a more aggressive heuristic, but rather in overcoming the inadequacy of a purely descriptive semantics through a decision-making formulation that internalizes the corpus's global structure.

From a methodological standpoint, this proposal introduces a significant innovation in the analysis of academic knowledge graphs. Purely ontological works assume that the quality of the system depends on the semantic richness of the model and the correct structuring of entities and relationships. The present study demonstrates that such richness remains necessary but is not sufficient in itself. The analytical quality of integrated knowledge also depends on the ability to resolve tensions between concentration and diversity, between intensive coverage and thematic balance, and between exhaustive representation and strategic utility. The combinatorial formulation employed in this work makes this difficulty explicit and turns it into the core of the solution. Therefore, the innovation is not merely algorithmic; it is also epistemological, because it redefines the academic infrastructure as an optimizable space rather than merely a semantically annotated repository.

This distinction has direct implications for research governance in higher education. An ontological approach allows for the identification of knowledge areas, thematic clusters, connections to the SDGs, collaboration networks, and production patterns. The proposed model adds greater utility by enabling the determination of which global configuration of semantic anchors offers the best balance between coverage, diversity, and analytical cost. This property is particularly relevant in university settings where academic graphs feed into scientific output dashboards, collaboration maps, profile systems, capability analytics, accreditation processes, and internationalization strategies. In such contexts, improvements in the system's structural quality can have a considerable impact on the quality of organizational inferences, the identification of scientific strengths, and evidence-based resource allocation.

It should also be emphasized that this work does not invalidate previous ontological developments, but rather re-situates them within a broader architecture. Ontology remains indispensable for defining entities, attributes, provenance, relationships, areas, SDGs, and

thematic clusters. The difference is that it is no longer the sole basis for decision-making. In this proposal, ontology provides the language of organization and explainability, while optimization provides the prioritization mechanism. This articulation is precisely one of the study's greatest contributions, as it avoids a false dichotomy between semantics and optimization. Evidence suggests that academic analytics reaches a higher level of maturity when both dimensions are integrated, as semantics ensures the interpretive legitimacy of the system; optimization ensures its structural and operational utility.

Compared to strictly descriptive or ontological works, this proposal offers four substantive advances. First, it introduces a global decision logic that internalizes the complete structure of the corpus, rather than treating each dimension as an isolated phenomenon. Second, it transforms the representation of relationships into a strategic selection of configurations; that is, it manages not only semantic connections but also the graph's systemic balance. Third, it improves the analytical utility of materialized knowledge by favoring configurations that are more coherent in terms of coverage and diversity, which is where academic consultation and evaluation systems actually operate. Fourth, it offers a more solid foundation for scaling the analysis to complex institutional contexts, where the quality of knowledge cannot depend exclusively on local counts or fragmented manual interpretations.

The discussion of these findings leads to a more bounded interpretation. Ontologies have represented a fundamental advance by incorporating structure, traceability, and explainability into academic knowledge graphs. However, in the analyzed single-university corpus, semantic representation alone was not sufficient to determine which anchors should be prioritized under constrained analytical conditions. The proposal developed in this work responds to that institutional decision problem by combining semantic representation with governance validation and budgeted maximum-coverage optimization. Thus, the study's distinctive contribution lies at the intersection of representation, optimization, and governance, but its empirical support should be interpreted within the analyzed university case rather than as evidence of general performance across all higher education systems.

6. Limitations

This study has five main limitations. First, the empirical validation is based on a single-university Scopus corpus. Therefore, the results demonstrate methodological feasibility and institutional analytical value, but they should not be interpreted as universal evidence for all higher education systems. Second, the assignment of knowledge areas and SDG tags depends on semantic rules derived from titles, abstracts, and keywords; although this procedure is reproducible, expert validation is still required to refine borderline cases. Third, the framework performs semantic normalization and anchor selection, but it does not implement full pairwise entity resolution or duplicate-record matching. Fourth, the comparison with prior works is operationalized through reproducible internal baselines rather than a complete reimplementation of all external algorithms. Fifth, although the sensitivity analysis evaluates 20 alternative weighting configurations, the objective-function weights still represent explicit institutional decision preferences rather than universally valid coefficients. Future multi-institutional studies should validate or estimate these weights through expert elicitation, stakeholder consultation, or data-driven preference-learning procedures.

Both manuscript-level documentation supports the reproducibility of the computational experiment and automatically generated execution artifacts. Table 8 distinguishes the elements explicitly described in the article from those exported by the implementation during each run. The mathematical formulation, dataset schema, governance constraints, algorithmic sequence, baseline definitions, and optimization logic are documented in the

manuscript. In addition, the computational workflow records the random seed, evaluated budgets, dataset dimensions, execution platform, Python version, and versions of the principal analytical libraries. It also generates configuration, dataset-profile, sensitivity-analysis, and reproducibility-manifest files. These resources support methodological reproducibility and controlled internal repetition of the experiment. Nevertheless, unrestricted external replication remains temporarily limited because the complete source code has not yet been publicly released during the intellectual property protection procedure.

Table 8. Reproducibility resources documented and generated by the GAGER implementation.

Reproducibility Component	Current Status	Evidence Provided
Analytical dataset schema	Documented	The final corpus contains 3558 bibliographic records, and the input fields used for bibliometric normalization, thematic clustering, ontology construction, provenance control, and optimization are described in the manuscript.
Mathematical formulation	Documented	The semantic states, decision variables, provenance constraints, governance-validation variable, coverage function, diversity measures, budget constraint, and composite objective function are explicitly defined.
Algorithmic workflow	Documented	Algorithm 1 reports the computational sequence from data validation and normalization to thematic clustering, provenance construction, governance validation, semantic-anchor selection, and baseline evaluation.
Random-seed control	Generated automatically	The implementation uses a fixed random seed of 42 for stochastic procedures, random-selection baselines, clustering initialization, and sensitivity-related computations.
Budget configuration	Generated automatically	The default optimization budgets and the maximum value used in the dense budget analysis are recorded in the execution configuration.
Software environment	Generated automatically	The computational workflow was executed using Python 3.12.7 on a 64-bit operating-system platform, with pandas 2.2.3, NumPy 2.1.3, NetworkX 3.4.2, Matplotlib 3.9.2, and scikit-learn. The implementation automatically records the exact scikit-learn version and the complete software environment in `run_config.json` for each execution.
Execution configuration	Generated automatically	The workflow exports `run_config.json`, containing the principal execution parameters, dataset characteristics, random seed, budget configuration, and software-environment information.
Dataset profile	Generated automatically	The workflow exports `dataset_profile.csv`, including the number of loaded records, number of columns, and relevant execution metadata.
Reproducibility manifest	Generated automatically	The workflow exports `reproducibility_manifest.txt` as a human-readable record of the configuration, computational environment, and principal parameters used in the experiment.
Weight-sensitivity results	Generated automatically	The implementation exports `table_weight_sensitivity_summary.csv`, containing the evaluated weight configuration, preferred budget, coverage, thematic diversity, knowledge-area diversity, SDG diversity, objective score, selected anchors, and Jaccard stability relative to the reference anchor configuration used in the sensitivity analysis.
Intermediate and final results	Generated automatically	The implementation exports thematic assignments, ontology mappings, optimization profiles, baseline comparisons, sensitivity summaries, tables, and publication-ready figures.
Complete source code	Temporarily restricted	Public availability is pending completion of the intellectual property protection procedure; therefore, unrestricted external replication is not yet possible.

Note: "Documented" indicates information explicitly reported in the manuscript. "Generated automatically" indicates artifacts created during execution of the computational workflow. The temporary restriction on the complete source code limits immediate external replication and should not be interpreted as full public reproducibility.

Future work should extend the validation to multi-institutional and cross-country datasets, incorporate expert-reviewed semantic assignments, integrate stronger entity-disambiguation modules, and compare the proposed framework with external algorithms under shared benchmark conditions.

7. Conclusions

This study proposed GAGER, a governance-aware, ontology-based framework for scholarly knowledge graph analytics evaluated as a Scopus-based single-university case study. The framework integrates bibliometric normalization, semantic mapping, provenance control, governance validation, thematic clustering, graph construction, semantic-anchor selection, sensitivity analysis, and baseline-based evaluation within a single documented and methodologically reproducible architecture.

The results show that combining ontology-based representation with governance-validated budgeted maximum-coverage optimization improves the analytical value of the scholarly knowledge graph within the analyzed institutional corpus. The proposed strategy achieved higher semantic coverage and a stronger composite score than ontology-only selection, random selection, and frequency-based selection baselines evaluated under the same budget constraints. Therefore, the baseline-based analysis indicates that the observed gain is not only a consequence of semantic annotation, but also of the optimization layer's ability to prioritize anchors according to their marginal contribution to the semantic-state universe.

The governance-aware component also produced a measurable experimental effect. Before optimization, the governance-validation layer reduced the initial candidate-anchor space from 27,542 anchors to 8320 validated anchors by filtering out anchors with insufficient provenance, weak semantic consistency, incomplete traceability, or non-canonical forms. This result shows that governance awareness is not only a conceptual or mathematical addition, but an operational filtering mechanism that constrains the feasible decision space used by the optimizer.

The main contribution of this work is not the replacement of ontology with optimization, but the integration of ontology, provenance, governance validation, and budgeted semantic selection. The ontology organizes the meaning of the scholarly corpus, the provenance layer preserves traceability, the governance layer restricts selection to auditable anchors, and the optimization layer selects the most informative semantic anchors under explicit budget constraints. Within the limits of the analyzed single-university case study, this combination supports institutional research analytics, thematic prioritization, and evidence-based decision-making.

Because the present evidence is derived from one institutional corpus, the findings should not be interpreted as universal evidence for higher education systems or science-policy environments in general. Future research should validate the framework using multi-institutional and cross-country datasets, strengthen expert-based validation of semantic assignments, incorporate advanced entity-disambiguation procedures, validate the objective-function weights through multi-stakeholder consultation, and compare the proposed model with external benchmark algorithms under controlled experimental conditions.

Author Contributions: E.I., V.R., D.L. and D.P. contributed to the conceptualization of the study and the design of the methodological framework. E.I. and V.R. developed the mathematical formulation and implemented the computational optimization approach. V.R. led the data processing and analytical evaluation and contributed to the formulation of the semantic coverage optimization model. E.I. contributed to integrating governance criteria, scholarly knowledge graph structures, and institutional research management requirements. D.L. supported the interpretation of the optimization logic and its structural consistency, and their alignment with knowledge governance in higher education. D.P. contributed to the validation of the methodological workflow, the review of the analytical structure, and the refinement of the manuscript's technical consistency. E.I., V.R., D.L. and D.P. participated in the drafting and revision of the manuscript. E.I. coordinated the overall research process and managed communication with the journal. All authors have read and agreed to the published version of the manuscript.

Funding: This research received institutional support from Universidad Politécnica Salesiana and the Telecommunications Research Group (GIREI-GIIATA) under the project "Integration of Interactive and AI-Assisted Technologies for Active and Personalized Learning in STEM Educational Environments to Enhance Critical Thinking and Problem-Solving Skills".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author (einga@ups.edu.ec) upon reasonable request. Due to an ongoing intellectual property protection process, the full dataset and complete Python implementation are currently not publicly accessible. However, reproducibility is supported through the availability of non-confidential materials, including scripts for figure and table generation, parameter configurations, and anonymized input data, which may be shared under appropriate confidentiality agreements, such as a non-disclosure agreement. The experimental protocol, model formulation, and evaluation procedures are fully described in the manuscript to ensure methodological transparency. A public repository containing the complete dataset and source code will be released once the intellectual property protection process is complete.

Acknowledgments: This research was conducted within the framework of the Universidad Politécnica Salesiana through the Research Vice-Rectorate and the Smart Grid Research Group (GIREI). The authors acknowledge the academic collaboration and intellectual exchange supported by the IUS Network through the RECI initiative on Smart Grids and Smart Cities, which contributed to the conceptual strengthening of this work. The use of generative artificial intelligence was strictly limited to minor grammatical and stylistic refinement. All conceptual, methodological, and analytical developments were carried out exclusively by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Verma, S.; Bhatia, R.; Harit, S.; Batish, S. Scholarly Knowledge Graphs Through Structuring Scholarly Communication: A Review. *Complex Intell. Syst.* **2023**, *9*, 1059–1095. [CrossRef] [PubMed]
2. Liu, F. Knowledge Graph-Based Intelligent Data Management and Information Innovation Service Model for University Library Systems. *PLoS ONE* **2026**, *21*, e0341307. [CrossRef] [PubMed]
3. Rajabi, E.; Midha, R.; de Souza, J.F. Constructing a Knowledge Graph for Open Government Data: The Case of Nova Scotia Disease Datasets. *J. Biomed. Semant.* **2023**, *14*, 4. [CrossRef] [PubMed]
4. Santini, C.; Gesese, G.A.; Peroni, S.; Gangemi, A.; Sack, H.; Alam, M. A Knowledge Graph Embeddings Based Approach for Author Name Disambiguation Using Literals. *Scientometrics* **2022**, *127*, 4887–4912. [CrossRef]
5. Rodrigues, N.S.; Mariano, A.M.; Ralha, C.G. Author Name Disambiguation Literature Review with Consolidated Meta-Analytic Approach. *Int. J. Digit. Libr.* **2024**, *25*, 765–785. [CrossRef]
6. Zhou, Q.; Chen, W.; Zhao, P.; Wang, X.; Zhang, H.; Lei, Z. Towards Effective Author Name Disambiguation by Hybrid Attention. *J. Comput. Sci. Technol.* **2024**, *39*, 929–950. [CrossRef]
7. Choi, D.; Jang, J.; Song, S.; Lee, H.; Lim, J.; Bok, K.; Yoo, J. Name Disambiguation Scheme Based on Heterogeneous Academic Sites. *Appl. Sci.* **2024**, *14*, 192. [CrossRef]
8. Jia, Z.; Fang, Z.; Zhang, H. Normalization of Web of Science Institution Names Based on Deep Learning. *Algorithms* **2024**, *17*, 312. [CrossRef]
9. Azeroual, O.; Jha, M.; Nikiforova, A.; Sha, K.; Alsmirat, M.; Jha, S. A Record Linkage-Based Data Deduplication Framework with DataCleaner Extension. *Multimodal Technol. Interact.* **2022**, *6*, 27. [CrossRef]
10. Ali Omar, Z.; Zamzuri, Z.H.; Mohd Ariff, N.; Abu Bakar, M.A. Training Data Selection for Record Linkage Classification. *Symmetry* **2023**, *15*, 1060. [CrossRef]
11. Shah, N.; Patiyara, S.; Basak, J.; Sahni, S.; Mathur, A.; Park, K.; Rajasekaran, S. Efficient Record Linkage in the Age of Large Language Models: The Critical Role of Blocking. *Algorithms* **2025**, *18*, 723. [CrossRef]
12. Karapiperis, D.; Verykios, V.S. Scaling Entity Resolution with K-Means: A Review of Partitioning Techniques. *Electronics* **2025**, *14*, 3605. [CrossRef]
13. Nananukul, N.; Sisaengsuwanchai, K.; Kejriwal, M. Cost-Efficient Prompt Engineering for Unsupervised Entity Resolution in the Product Matching Domain. *Discov. Artif. Intell.* **2024**, *4*, 56. [CrossRef]

14. Shu, Y.; Zhang, J.; Huang, G.; Chi, C.H.; He, J. Entity Alignment via Graph Neural Networks: A Component-Level Study. *World Wide Web* **2023**, *26*, 4069–4092. [CrossRef]
15. Zhu, B.; Wang, R.; Wang, J.; Shao, F.; Wang, K. A Survey: Knowledge Graph Entity Alignment Research Based on Graph Embedding. *Artif. Intell. Rev.* **2024**, *57*, 229. [CrossRef]
16. Zhou, Y.; Yan, X.; Huang, H.; Hao, Z.; Zhu, H.; Liu, F. Knowledge Graph-Based Entity Alignment with Unified Representation for Auditing. *Complex Intell. Syst.* **2025**, *11*, 229. [CrossRef]
17. Li, H.; Zhang, C.; Chen, X.; Wang, C.; Yue, Y. A Multi-Modal Entity Alignment Method Based on Image Enhancement and Credibility Iteration. *Complex Intell. Syst.* **2025**, *11*, 466. [CrossRef]
18. Li, J.; Ma, Y.; Meng, L. QLGAN: A Quantum-Lineage Graph Attention Network for Temporal Knowledge Graph Entity Alignment. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 66. [CrossRef]
19. Fang, J.; Yan, X. MDSEA: Knowledge Graph Entity Alignment Based on Multimodal Data Supervision. *Appl. Sci.* **2024**, *14*, 3648. [CrossRef]
20. Cai, W.; Zhou, R.; Ma, W. Efficient and Effective Unsupervised Entity Alignment in Large Knowledge Graphs. *Appl. Sci.* **2025**, *15*, 1976. [CrossRef]
21. Wang, C.; Wang, C.; Weng, Y.; Liu, Z. Fusion-Optimized Multimodal Entity Alignment with Textual Descriptions. *Information* **2025**, *16*, 534. [CrossRef]
22. Cai, W.; Ma, W. Learnable Convolutional Attention Network for Unsupervised Knowledge Graph Entity Alignment. *Entropy* **2025**, *27*, 924. [CrossRef] [PubMed]
23. Li, H.; Hou, Y.; Liu, J.; Zhang, P.; Wang, C.; Liu, K. Multi-Modal Entity Alignment Based on Enhanced Relationship Learning and Multi-Layer Feature Fusion. *Symmetry* **2025**, *17*, 990. [CrossRef]
24. Zhu, S.; Tai, Q.; Wang, J.; Tang, M.; Wang, L.; Li, N.; Hou, S.; Liu, X. UMEAD: Unsupervised Multimodal Entity Alignment for Equipment Knowledge Graphs via Dual-Space Embedding. *Symmetry* **2025**, *17*, 1869. [CrossRef]
25. Qu, J.; Wang, J.; Zhao, Z.; Chen, X. MBJELEL: An End-to-End Knowledge Graph Entity Linking Method Applied to Civil Aviation Emergencies. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 237. [CrossRef]
26. Karpov, I.; Kirillovich, A.; Goncharova, E.; Parinov, A.; Chernyavskiy, A.; Ilvovsky, D. SynEL: A Synthetic Benchmark for Entity Linking. *PLoS ONE* **2026**, *21*, e0339468. [CrossRef] [PubMed]
27. Gu, J.; Liu, W.; Yang, X. An Attribute-Enhanced Relationship-Aware Neighborhood Matching Model with Dual Attention. *PLoS ONE* **2025**, *20*, e0324290. [CrossRef] [PubMed]
28. Galán-Mena, J.; López-Nores, M.; Pulla-Sánchez, D.; Guerrero-Vásquez, L.F.; Salgado-Guerrero, J.P. OntoDup: Governance-Aware Entity Matching for Scholarly Knowledge Graph Deduplication. *Information* **2026**, *17*, 325. [CrossRef]
29. Mavridis, A.; Tegos, S.; Anastasiou, C.; Papoutsoglou, M.; Meditskos, G. Large Language Models for Intelligent RDF Knowledge Graph Construction: Results from Medical Ontology Mapping. *Front. Artif. Intell.* **2025**, *8*, 1546179. [CrossRef] [PubMed]
30. Ghanem, H.; Cruz, C. Fine-Tuning or Prompting on LLMs: Evaluating Knowledge Graph Construction Task. *Front. Big Data* **2025**, *8*, 1505877. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.