

Nov 13, 2025

Integrating Bibliometrics and Topic Modeling: A PRISMA-Based Protocol for Reviewing Grey Clustering Research

DOI

<https://dx.doi.org/10.17504/protocols.io.ewov11wdyvr2/v1>

Camelia Delcea¹

¹Bucharest University of Economic Studies, Bucharest, Romania

Grey Clustering



Camelia Delcea

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.ewov11wdyvr2/v1>

Protocol Citation: Camelia Delcea 2025. Integrating Bibliometrics and Topic Modeling: A PRISMA-Based Protocol for Reviewing Grey Clustering Research. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.ewov11wdyvr2/v1>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: November 13, 2025

Last Modified: May 18, 2026

Protocol Integer ID: 232321

Keywords: reviewing grey clustering research, mining approaches such as latent dirichlet allocation, systematic bibliometric review, integrating bibliometric, latent dirichlet allocation, broader framework of grey systems theory, grey clustering method, grey clustering research this protocol, topic modeling, grey systems theory, applications across research domain, research domain, mining approach, web of science, methodological framework, topic, publication, prisma 2020 principle

Abstract

This protocol defines the methodological framework for a systematic bibliometric review focused on *grey clustering methods and their applications across research domains*. The study aims to map the evolution, thematic directions, and areas of application of grey clustering within the broader framework of Grey Systems Theory. Publications were retrieved from the Web of Science (WoS) Core Collection and analyzed using bibliometric and topic-modeling tools. The research integrates PRISMA 2020 principles with advanced text-mining approaches such as Latent Dirichlet Allocation (LDA) and BERTopic, providing a comprehensive overview of the field's conceptual and thematic development.

Guidelines

Eligibility Criteria

Inclusion criteria:

- Language: English
- Document type: Article (including conference papers classified as articles)
- Indexing: Must be listed in WoS Core Collection
- Topic relevance: Must explicitly address *grey clustering* methods or applications

Exclusion criteria:

- Non-English publications (9 excluded)
- Non-article document types (324 excluded)
- Retracted papers (7 excluded)
- Papers published in 2025 (8 excluded due to incomplete year coverage)
- Irrelevant papers (5 excluded after manual screening)

Final dataset: 318 articles.

Screening Process

- The **PRISMA 2020 methodology** was followed for systematic selection.
- Articles were filtered using keyword queries, then screened manually for topical relevance.
- Duplicates, non-relevant, and retracted records were excluded.
- The final selection is illustrated in **Figure 1 of the manuscript (PRISMA flowchart)**.

Materials

- **Biblioshiny (R package Bibliometrix 5.0)** – descriptive and network bibliometric analyses
- **VOSviewer 1.6.20 and CiteSpace 6.2.R4** – visualization of co-occurrence and citation networks
- **Python 3.11** environment for NLP analyses:
 - **Gensim** – Latent Dirichlet Allocation (LDA)
 - **BERTopic** – transformer-based topic modeling (using HDBSCAN clustering)
- Grid search used for **LDA alpha/eta** tuning; BERTopic parameters (minimum cluster size, sample size) optimized for balance between coherence and interpretability.

Before start

Data Preprocessing for NLP

- Text converted to lowercase
- Punctuation and special characters removed
- Token normalization applied (e.g., "GST" → "grey systems theory")
- Domain-specific stopwords and selection terms removed (e.g., "greycluster", "graycluster")
- Term lemmatization performed for semantic consistency

Abstract

- 1 This protocol defines the methodological framework for a systematic bibliometric review focused on *grey clustering methods and their applications across research domains*. The study aims to map the evolution, thematic directions, and areas of application of grey clustering within the broader framework of Grey Systems Theory. Publications were retrieved from the Web of Science (WoS) Core Collection and analyzed using bibliometric and topic-modeling tools. The research integrates PRISMA 2020 principles with advanced text-mining approaches such as Latent Dirichlet Allocation (LDA) and BERTopic, providing a comprehensive overview of the field's conceptual and thematic development.

Databases and Indexes Used

- 2 All records were extracted from the *Web of Science (WoS) Core Collection*, using an institutional subscription. The following ten indexes were included:
 - Science Citation Index Expanded (SCIE)
 - Social Sciences Citation Index (SSCI)
 - Arts & Humanities Citation Index (A&HCI)
 - Emerging Sources Citation Index (ESCI)
 - Conference Proceedings Citation Index—Science (CPCI-S)
 - Conference Proceedings Citation Index—Social Sciences and Humanities (CPCI-SSH)
 - Book Citation Index—Science (BKCI-S)
 - Book Citation Index—Social Sciences and Humanities (BKCI-SSH)
 - Current Chemical Reactions (CCR-Expanded)
 - Index Chemicus (IC)

Search Strategy

- 3 The search was performed in the Title (TI), Abstract (AB), and Author Keywords (AK) fields using the following keyword queries:
 - "grey cluster*"
 - "gray cluster"Special search operators:
 - * (asterisk) was used as a wildcard to capture plural and lexical variations.To ensure inclusiveness, both British ("grey") and American ("gray") spellings were applied. The search was conducted on all records available in the selected indexes up to December 2024.

Eligibility Criteria

4 Inclusion criteria:

- Language: English
- Document type: Article (including conference papers classified as articles)
- Indexing: Must be listed in WoS Core Collection
- Topic relevance: Must explicitly address *grey clustering* methods or applications

Exclusion criteria:

- Non-English publications (9 excluded)
- Non-article document types (324 excluded)
- Retracted papers (7 excluded)
- Papers published in 2025 (8 excluded due to incomplete year coverage)
- Irrelevant papers (5 excluded after manual screening)

Final dataset: 318 articles.

Screening Process

- 5
- The **PRISMA 2020 methodology** was followed for systematic selection.
 - Articles were filtered using keyword queries, then screened manually for topical relevance.
 - Duplicates, non-relevant, and retracted records were excluded.
 - The final selection is illustrated in **Figure 1 of the manuscript (PRISMA flowchart)**.

Tools and Software

- 6
- **Biblioshiny (R package Bibliometrix 5.0)** – descriptive and network bibliometric analyses
 - **VOSviewer 1.6.20 and CiteSpace 6.2.R4** – visualization of co-occurrence and citation networks
 - **Python 3.11** environment for NLP analyses:
 - **Gensim** – Latent Dirichlet Allocation (LDA)
 - **BERTopic** – transformer-based topic modeling (using HDBSCAN clustering)
 - Grid search used for **LDA alpha/eta** tuning; BERTopic parameters (minimum cluster size, sample size) optimized for balance between coherence and interpretability.

Data Preprocessing for NLP

- 7
- Text converted to lowercase
 - Punctuation and special characters removed
 - Token normalization applied (e.g., "GST" → "grey systems theory")



- Domain-specific stopwords and selection terms removed (e.g., "greycluster", "graycluster")
- Term lemmatization performed for semantic consistency

Key Outcomes Measured

- 8
 - Annual scientific production and growth rate
 - Most influential sources, authors, and documents (citation and co-citation metrics)
 - Keyword Plus and Authors' Keywords co-occurrence networks
 - Thematic maps (Biblioshiny) and evolution across time periods
 - Topic discovery via LDA and BERTopic, with cross-validation
 - Identification of research gaps, challenges, and future directions

Retrospective Registration Note

- 9

This protocol was registered *retrospectively after data extraction* began, in accordance with **PRISMA 2020 transparency guidelines**. No amendments have been made after registration.