Thiruppugazh-KG Dataset: A Manually Annotated Resource for Computational Analysis of Tamil Devotional Literature

Garthigan Kumarasamy, Jubeerathan Thevakumar, Sathurgini Uthayakumar, Disne Kajanath, Narthana Sivalingam, Uthayasanker Thayasivam
Department of Computer Science and Engineering,
University of Moratuwa, Sri Lanka
{garthigank.25,jubeerathan.25,sathurgini.25}
{disne.25,narthanas,rtuthaya}@cse.mrt.ac.lk

Abstract

This paper introduces Thiruppugazh-KG, a semantically annotated dataset and knowledge graph derived from the Thiruppugazh corpus, a 14th-century collection of 1,335 Tamil devotional hymns composed by Arunagirinathar. The dataset includes annotations for entities, devotional themes, mythological events, philosophical concepts, imagery, and sacred locations mentioned in each hymn. Using these annotations, we construct a Neo4j-based knowledge graph that models relationships between hymns and their associated cultural and narrative elements. Graph analytics, including PageRank, are applied to identify prominent entities and sacred locations within the corpus. The resulting resource provides a structured representation of Tamil devotional literature and supports computational analysis of cultural texts in low-resource languages.

1 Introduction

Tamil devotional literature forms a significant part of South Asian religious and literary culture. Thiruppugazh[1], composed by the 14th-century poet-saint Arunagirinathar, is one of the most celebrated devotional works in Tamil tradition. The corpus contains 1,335 hymns dedicated primarily to Lord Murugan and is renowned for its intricate poetic style, complex rhythmic structure, and vivid devotional symbolism (Zvelebil, 1973) . These hymns combine mythology, spiritual philosophy, and descriptions of sacred temples, creating a rich literary and religious landscape that has been studied extensively within Tamil scholarship.

Recent developments in digital humanities have introduced computational approaches for analyzing large literary corpora. Methods such as distant reading and macroanalysis allow scholars to identify large-scale thematic and structural patterns across texts rather than focusing only on individual works (Moretti, 2013) (Jockers, 2013). Within this context, knowledge graphs provide a powerful framework for representing semantic relationships extracted from textual data. In a knowledge graph, entities and concepts are modeled as nodes while relationships between them are represented as edges, enabling network analysis, semantic querying, and discovery of hidden patterns within complex datasets. Knowledge graphs model entities and relationships as nodes and edges, enabling network analysis and semantic querying of complex datasets (Hogan et al., 2021) (Robinson et al., 2015).

Knowledge graph techniques have been successfully applied in domains such as web knowledge bases, digital libraries, and cultural heritage datasets. Projects such as DBpedia demonstrate how structured knowledge can be extracted from textual sources and organized into machine-readable semantic networks (Auer et al., 2007) (Bizer et al., 2023). Despite these advances, similar computational approaches remain relatively rare in the study of Tamil devotional literature and Bhakti poetry.

This study addresses this gap by representing the Thiruppugazh corpus as a knowledge graph. A manually annotated dataset of 1,335 songs was created containing entities, mythological events, devotional themes, philosophical concepts, symbolic imagery, and temple locations mentioned in each hymn. Using these annotations, a Neo4j-based knowledge graph was constructed and analyzed using graph algorithms to identify dominant themes, frequently occurring mythological narratives, and prominent temples referenced within the corpus.

The main contributions of this work are as follows:

1. We introduce Thiruppugazh-KG, the first semantically annotated knowledge graph dataset

[1] https://en.wikipedia.org/wiki/Tiruppukal

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 62–70

for Tamil devotional literature.

2. We provide manual annotations for 1,335 hymns covering entities, themes, mythology, philosophy, imagery, and sacred geography.

3. We construct a Neo4j-based devotional knowledge graph supporting semantic querying and graph analytics.

4. We demonstrate the utility of the resource through graph-based retrieval and computational literary analysis.

2 Related Work

Computational approaches in digital humanities have enabled new ways of analyzing large literary corpora using semantic and data-driven methods. Techniques such as knowledge graphs, network analysis, and probabilistic text modeling help represent and analyze relationships among entities in cultural and literary datasets. This section reviews previous research related to knowledge graphs and computational approaches used in literary analysis.

2.1 Knowledge Graphs

Knowledge graphs are structured representations of entities and relationships that enable semantic reasoning and data integration across complex domains. Hogan et al. define knowledge graphs as networks of real-world entities connected by semantic relationships that enable knowledge discovery and reasoning across heterogeneous data sources (Hogan et al., 2021). Such graphs have become fundamental infrastructure in modern data management and artificial intelligence systems.

Large-scale knowledge graphs such as DBpedia extract structured information from Wikipedia and represent it in RDF format, enabling semantic web applications and knowledge discovery (Auer et al., 2007). These systems demonstrate the potential of graph-based representations for organizing large bodies of information. Graph databases such as Neo4j provide efficient implementations of the property graph model and support complex graph queries and algorithms. Robinson, Webber, and Eifrem note that graph databases are particularly well suited for domains involving highly connected data structures (Robinson et al., 2015).

2.2 Knowledge Graphs in Digital Humanities

Digital humanities research increasingly uses knowledge graphs to represent cultural heritage data, historical texts, and literary networks. Romanello and colleagues demonstrate that knowledge graphs allow scholars to represent relationships among texts, authors, places, and historical events in a machine-readable structure that supports computational analysis (Kokash et al., 2023).

Knowledge graph models have also been applied to historical archives, museum collections, and literary corpora to capture relationships between characters, events, and locations. These approaches enable researchers to analyze narrative structures and cultural networks using computational techniques (Crane et al., 2006; McCarty, 2005).

2.3 Computational Literary Analysis

Computational literary studies use data-driven techniques to analyze narrative structures, themes, and character networks in literary texts. Newman's work on network science provides theoretical foundations for analyzing relationships within complex networks, including social and narrative networks (Newman, 2001).

Network analysis has been applied to classical literature, including Shakespeare's plays and epic narratives, to examine relationships among characters and thematic patterns. Similarly, distant reading approaches proposed by Moretti emphasize analyzing large collections of texts using quantitative methods rather than focusing solely on individual works (Moretti, 2013).

Topic modeling techniques and probabilistic text models have also been used to identify thematic structures within literary corpora (Blei, 2012). These computational methods demonstrate the value of large-scale textual analysis for literary scholarship.

2.4 Tamil Literature Studies

Tamil literature has been extensively studied by scholars such as Zvelebil, who documented its historical development and cultural significance, particularly within the context of devotional poetry and religious traditions (Zvelebil, 1973). Murugan devotion occupies a central place in Tamil religious literature and continues to influence contemporary cultural practices.

However, computational approaches to Tamil literary analysis remain relatively limited. Most existing research focuses on linguistic processing, digitization, and corpus creation rather than semantic knowledge modeling. The present study contributes to this emerging field by introducing

a knowledge graph representation for Tamil devotional literature.

3 Dataset Description

The dataset used in this study is derived from the *Thiruppugazh* corpus, a classical Tamil devotional literary work composed by the 14th-century saint-poet Arunagirinathar. The corpus contains 1,335 devotional hymns dedicated primarily to Lord Murugan, a central deity in Tamil Hindu spiritual traditions. The hymns are known for their complex poetic structure, mythological references, symbolic imagery, and philosophical teachings.

To enable computational analysis, the Thiruppugazh corpus was transformed into a structured dataset with semantic annotations capturing mythological entities, devotional themes, philosophical ideas, and narrative events. This structured representation supports tasks such as knowledge extraction, semantic search, and digital humanities analysis.

3.1 Data Source

The original Tamil lyrics and explanatory interpretations were collected from the Kaumaram Thiruppugazh digital archive[2], which preserves the works of Arunagirinathar. Digital repositories of classical literature are increasingly used in digital humanities research as they enable computational analysis of historical texts and cultural narratives (Crane et al., 2006).

Each hymn entry contains the original Tamil text, Tamil explanatory notes, and an English interpretation. Additional semantic annotations were created to capture key literary and mythological concepts present in the hymns.

3.2 Dataset Structure

Each hymn is represented as a structured record containing linguistic content and semantic annotations. These fields allow systematic analysis of devotional themes, mythological narratives, and philosophical concepts present in the Thiruppugazh corpus. The dataset schema is summarized in Table 1.The final dataset contains semantic annotations for all hymns in the Thiruppugazh corpus. Each hymn has been structured to capture both linguistic information and semantic metadata.The overall dataset statistics are summarized in Table 2.

[2]https://www.kaumaram.com/thiru/

Field Name	Description
song_number	Unique identifier assigned to each hymn.
song_name	Title or commonly recognized name of the hymn.
abode	Sacred location associated with Murugan referenced in the hymn.
song_tamil	Original Tamil lyrics of the hymn.
definition_tamil	Explanatory interpretation of the hymn in Tamil.
definition_english	English explanation describing the meaning of the hymn.
themes	Major devotional or spiritual themes present in the hymn.
entities	Named entities such as deities, mythological characters, or places.
myth_events	Mythological narratives or divine events described in the hymn.
philosophy	Spiritual or philosophical concepts conveyed in the text.
imagery	Symbolic or poetic imagery used within the hymn.
relationships	Interactions between divine figures or characters.
summary	Concise description summarizing the hymn.
keywords	Key terms representing major concepts in the hymn.
languages	Languages available in the record (Tamil and English).

Table 1: Schema of the Thiruppugazh dataset

The annotated Thiruppugazh-KG dataset used in this study is publicly available to support reproducibility and future research in Tamil digital humanities and low-resource NLP. The repository[3] includes the structured JSON dataset, annotation guidelines, Neo4j knowledge graph resources, example Cypher queries, and documentation for dataset construction and usage.

3.3 Annotation Methodology

The dataset annotations were created through a structured manual annotation process carried out

[3]https://github.com/Garthigan/Thiruppugazh-KG-Dataset

Attribute	Value
Corpus Name	Thiruppugazh
Author	Arunagirinathar
Number of Hymns	1,335
Languages	Tamil , English
Annotation Types	Themes, Entities, Myth Events, Imagery, Philosophy, Relationships
Data Source	Kaumaram Thiruppugazh Archive
Data Format	JSON structured records
Domain	Tamil devotional literature

Table 2: Statistics of the Thiruppugazh dataset

by two annotators with background in Tamil literature and computational linguistics. Both annotators were familiar with the cultural and religious context of the Thiruppugazh corpus. In cases where disagreements occurred, a third reviewer with expertise in Tamil devotional literature was consulted to resolve conflicts.

Each hymn in the Thiruppugazh corpus, along with its accompanying interpretation from the Kaumaram digital archive, was carefully examined to identify important semantic components. The annotation focused on extracting key elements such as named entities, devotional themes, mythological events, philosophical concepts, and poetic imagery.

The annotation process followed a predefined workflow. First, the Tamil hymn and its interpretation were reviewed to understand the narrative and symbolic meaning. Next, semantic elements such as divine figures, mythological references, and spiritual teachings were identified. These elements were then categorized into predefined annotation fields including *themes*, *entities*, *imagery*, *mythological events*, *philosophical concepts*, and *relationships*.

To ensure consistency across annotations, a set of annotation guidelines was developed describing the definition and scope of each category. For example, *entities* correspond to divine figures or mythological characters referenced in the hymn, while *themes* represent the primary devotional or spiritual ideas expressed in the verse. *Mythological events* capture references to well-known narratives from Hindu mythology, and *imagery* represents symbolic metaphors used in the poetic language of the hymn.

To evaluate annotation consistency, Whole hymns was independently annotated by both annotators. Inter-annotator agreement was measured using Cohen's Kappa (Cohen, 1960) coefficient across the major annotation categories. The resulting agreement score was $\kappa = 0.78$, indicating substantial agreement between annotators.

This annotation approach enables a structured representation of literary knowledge while preserving the religious and cultural context of the original hymns. The resulting dataset supports computational analysis of Tamil devotional literature and facilitates tasks such as semantic search, knowledge graph construction, and knowledge-based retrieval for large language models.

The overall workflow used to construct the dataset is illustrated in Figure 1. The process begins with collecting the hymns and interpretations from the Kaumaram digital archive, followed by manual textual analysis and semantic annotation. The annotated information is then organized into a structured JSON format to create the final Thiruppugazh dataset.

4 Knowledge Graph Construction

The knowledge graph for the Thiruppugazh corpus was constructed using the Neo4j graph database, which is a widely adopted platform for implementing property graph–based knowledge systems (Robinson et al., 2015). Neo4j enables efficient storage and querying of complex relationships among entities, making it suitable for representing the semantic structure of devotional literature(Neo4j Inc., 2023). In this work, the knowledge graph models relationships between songs, mythological references, philosophical ideas, and geographic locations mentioned in the Thiruppugazh text.

4.1 Node Relationship Types

The graph schema consists of several node types representing the key semantic components extracted from the Thiruppugazh corpus. Each node represents a distinct concept or entity within the text. Those are Song, Entity, Theme, MythEvent, Philosophy, Imagery and Place. These node types enable structured representation of the literary and spiritual knowledge embedded within the Thiruppugazh corpus.

The semantic connections between nodes are modeled through directed relationships. These re-

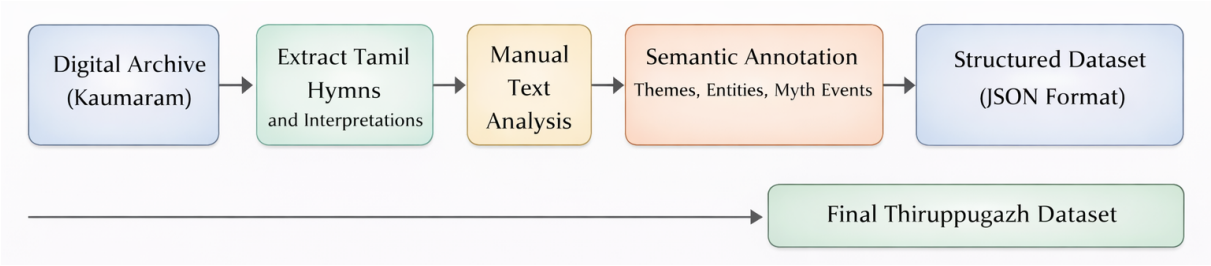

Figure 1: Annotation workflow used to construct the Thiruppugazh dataset

Figure 2: Architecture of the Thiruppugazh Knowledge Graph showing the relationships between Song nodes

Figure 3: Example subgraph for Song 1183 "Porudha kayalvizhi"

lationships capture the contextual associations between songs and their corresponding thematic or mythological elements.

The overall structure of these node types and relationships is illustrated in Figure 2, which presents the conceptual architecture of the Thiruppugazh knowledge graph.

4.2 Example Subgraph

An example of a subgraph representing a Thiruppugazh hymn is illustrated below Figure 3

This representation captures the semantic relationships between the hymn, the deity Murugan, the mythological event referenced in the text, and the sacred location associated with the song.

4.3 Murugan Devotional Network

To further analyze the devotional structure of the Thiruppugazh corpus, a Murugan-centered subgraph was extracted from the knowledge graph. This subgraph highlights how multiple hymns are connected through references to the deity Murugan and their associated thematic and mythological elements.

The following Cypher query was used to retrieve songs that mention Murugan and their related semantic nodes:

Listing 1: Cypher query used to extract the Murugan devotional network

```
MATCH (e:Entity {name:"Murugan"})<-[:
    MENTIONS_ENTITY]-(s:Song)
MATCH (s)-[r]->(n)
RETURN e, s, r, n
LIMIT 100
```

The resulting network places Murugan as the central entity, connected to multiple Song nodes through the MENTIONS_ENTITY relationship. Each song node is further linked to additional semantic nodes such as Theme, MythEvent, and Place, which represent the thematic context, mythological references, and sacred locations associated with the hymn.

As illustrated in Figure 4, this structure forms a devotional knowledge network where Murugan acts as the central hub connecting various hymns and their corresponding semantic attributes within the Thiruppugazh corpus.

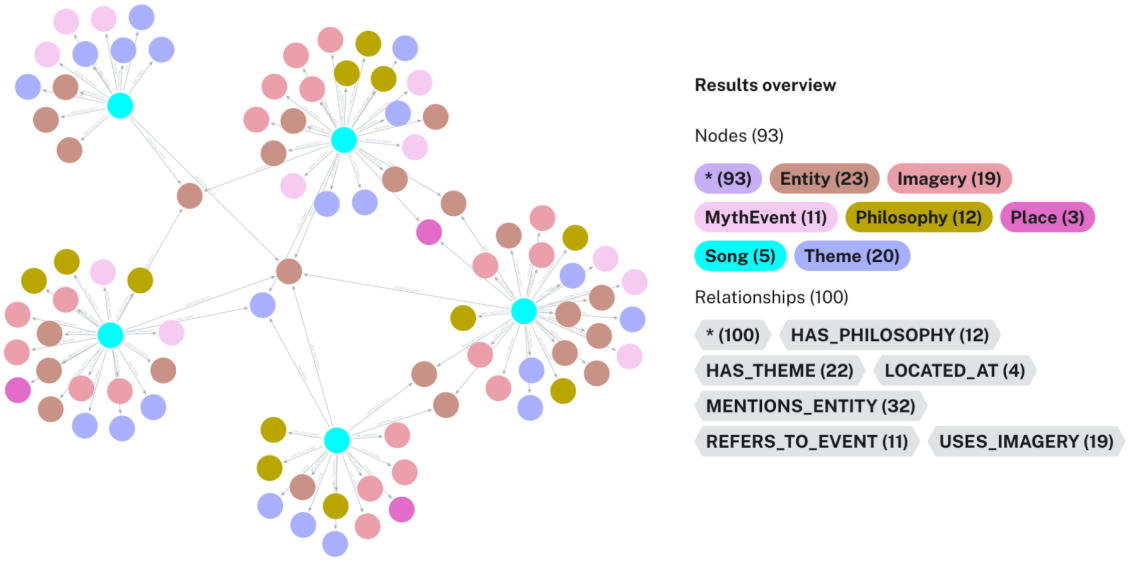

Figure 4: Murugan-centered devotional network extracted from the Thiruppugazh knowledge graph

5 Graph Analytics

To provide an overview of the constructed knowledge graph, Table 3 summarizes key structural statistics including the number of nodes, edges, average degree, and detected communities.

Metric	Value
Nodes	11992
Edges	28228
Average Degree	4.70
Communities	9758

Table 3: Summary statistics of the Thiruppugazh knowledge graph

Graph analytics techniques were applied to the constructed knowledge graph to explore the structural and semantic relationships present within the Thiruppugazh corpus. By leveraging the graph-based representation implemented in Neo4j, it becomes possible to analyze how devotional themes, mythological events, and sacred locations are interconnected through individual hymns.

The PageRank algorithm was applied to identify the most central entities within the Thiruppugazh knowledge graph.The computation was performed using the Neo4j Graph Data Science (GDS) library with a damping factor of 0.85.To facilitate interpretation and comparison, the resulting PageRank scores were normalized using Min–Max scaling.

Entity	PageRank Score
Murugan	1.0
Siva	0.32
Valli	0.30
Brahma	0.19

Table 4: Top entities ranked by normalized PageRank scores in the Thiruppugazh knowledge graph

Table 4 presents the top ranked entities based on the normalized PageRank scores. While Murugan naturally appears as the highest ranked entity due to the devotional focus of the corpus, the relative prominence of other entities provides additional insight into the structure of the narrative network. Figures such as Siva and Valli appear as highly influential nodes, reflecting their strong mythological and relational connections within the hymns. Brahma also emerges as a notable entity, indicating the presence of broader Hindu cosmological references within the corpus.

These rankings highlight how the knowledge graph captures relationships among divine figures and mythological elements across the hymns. In particular, the prominence of entities such as Siva and Valli suggests recurring narrative associations with Murugan that contribute to the interconnected structure of the devotional network.

Community detection analysis was further performed to examine the structural organization of

the devotional network. The detected communities revealed clusters corresponding to battle mythology, sacred geography, and devotional philosophy. These clusters indicate that semantically related hymns form densely connected subgraphs within the knowledge graph, demonstrating the effectiveness of the proposed representation in capturing thematic and narrative relationships across the corpus.

6 Results and Discussion

The constructed knowledge graph enables quantitative analysis of the Thiruppugazh corpus by examining the frequency of themes, mythological events, sacred locations, and relationship types extracted from the hymns. Through graph-based representation and querying, the dataset allows systematic exploration of devotional patterns and narrative structures embedded within the corpus.

The analysis of sacred places shows that several temple locations frequently appear in the hymns. Among these, Pazhani (98 occurrences), Thiruchchendhur (83), Thiruvaruṇai (81), Chidhambaram (67), and Thiruththanigai (64) are prominently referenced. These locations correspond to historically significant Murugan temples and pilgrimage centers in Tamil religious tradition. The frequency of these places within the graph highlights how devotional poetry is strongly associated with sacred geography, where hymns not only express spiritual devotion but also situate that devotion within specific temple sites and pilgrimage contexts.

The analysis of mythological references reveals recurring narrative motifs across the corpus. Several events associated with Murugan's battles against demonic forces appear frequently, including Murugan defeating Suran (45 occurrences), Murugan's victory over demons (41), and destruction of Suran (33). Other mythological narratives such as Murugan's marriage to Vaḷḷi, the destruction of Thiripuram, and the churning of the ocean are also represented in the graph. These results demonstrate that Thiruppugazh hymns incorporate a wide range of mythological episodes drawn from broader Hindu narrative traditions, which are interwoven with devotional expression.

Thematic analysis further reveals the distribution of spiritual concepts within the corpus. The theme devotion to Murugan appears most frequently (969 occurrences), followed by spiritual longing (234), spiritual enlightenment (168), divine intervention (164), spiritual awakening (149), and divine protection (131). While the devotional focus of the corpus is already recognized in traditional literary studies, the knowledge graph provides a structured and quantitative representation of these themes, enabling computational examination of how devotional ideas recur and connect across multiple hymns.

The distribution of relationship types within the graph also provides insights into the structural composition of the dataset. The most common relationship is MENTIONS_ENTITY (8,564 occurrences), indicating that references to divine figures and mythological characters form a central component of the corpus. This is followed by USES_IMAGERY (6,166), HAS_THEME (5,544), and HAS_PHILOSOPHY (4,104), which highlight the extensive use of poetic imagery, thematic expression, and philosophical reflection in the hymns. Relationships such as REFERS_TO_EVENT (2,515) and LOCATED_AT (1,335) further link songs to mythological narratives and sacred places. Together, these structural patterns demonstrate how the knowledge graph captures multiple layers of semantic information, including narrative, symbolic, and geographical dimensions of the Thiruppugazh corpus.

Table 5 summarizes the two most frequent occurrences in each category extracted from the knowledge graph. These results provide a compact overview of the most prominent themes, mythological narratives, sacred locations, and relationship types represented in the graph.

6.1 Limitations and Future Work

Although the proposed knowledge graph successfully represents important semantic relationships within the Thiruppugazh corpus, several limitations remain.

First, the current system relies on manual or semi-automatic extraction of entities and themes, which may limit scalability when processing larger textual datasets. Automated natural language processing techniques could be incorporated in future work to improve the extraction of entities, relationships, and semantic concepts from the corpus.

Second, the present graph focuses primarily on a limited set of node types such as songs, themes, entities, mythological events, and places. Future research could extend the schema to include additional semantic layers such as poetic structure, lin-

Category	Rank 1 (Frequency)	Rank 2 (Frequency)
Place	Pazhani (98)	Thiruchchendhur (83)
Mythological Event	Murugan defeating Suran (45)	Murugan's victory over demons (41)
Theme	Devotion to Murugan (969)	Spiritual longing (234)
Relationship Type	MENTIONS_ENTITY (8564)	USES_IMAGERY (6166)

Table 5: Top Two Frequent Concepts Extracted from the Thiruppugazh Knowledge Graph

guistic features, historical context, and inter-textual references.

Future work may also integrate advanced graph analytics techniques, including community detection, centrality measures, and knowledge graph embeddings, to further analyze the devotional structure and thematic relationships present in the corpus. Additionally, interactive visualization tools could be developed to allow scholars and researchers to explore the knowledge graph more effectively.

7 Conclusion

We present a manually annotated knowledge graph dataset derived from the Thiruppugazh corpus to model semantic relationships between songs, themes, mythological events, entities, imagery, philosophical concepts, and sacred locations. The annotations were used to construct a Neo4j-based knowledge graph that enables structured exploration of devotional literature. Our analysis shows that devotion to Murugan is the most dominant theme, while sacred locations such as Pazhani and Thiruchchendhūr frequently appear in the hymns. These findings demonstrate that knowledge graph–based representations can effectively capture the semantic structure of the Thiruppugazh corpus. This work provides a foundation for future computational analysis and knowledge-driven exploration of Tamil devotional literature.

Acknowledgments

The author sincerely thank the volunteers who assisted in manually annotating the Thiruppugazh dataset.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2023. Linked data-the story so far. In *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web*, pages 115–143.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. 2006. Beyond digital incunabula: Modeling the next generation of digital libraries. In *International Conference on Theory and Practice of Digital Libraries*, pages 353–366. Springer.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, and 1 others. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Natallia Kokash, Matteo Romanello, Ernest Suyver, and Giovanni Colavizza. 2023. From books to knowledge graphs. *Journal of Data Mining & Digital Humanities*, 2023.

Willard McCarty. 2005. *Humanities computing*. Springer.

Franco Moretti. 2013. *Distant reading*. Verso Books.

Neo4j Inc. 2023. *Neo4j Graph Data Science Documentation*. Accessed: 2026-03-01.

Mark EJ Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409.

Ian Robinson, Jim Webber, and Emil Eifrem. 2015. *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.".

Kamil Zvelebil. 1973. *The Smile of Murugan: On Tamil Literature of South India*. Brill.

A Thiruppugazh-KG Dataset Example

Examples in Thiruppugazh-KG Dataset are shown in Table 6.

Field	Example Value
song_number	0001
song_name	Muththaiththaru
abode	thiruvarunai
entities	Murugan, Siva, Brahma, Vishnu, Ravana, Arjuna
themes	devotion to Murugan; divine protection; spiritual teachings
myth_events	preaching of OM; destruction of Ravana's heads; churning of the ocean
philosophy	Karma Yoga; spiritual enlightenment
imagery	pearls and teeth; powerful spear; milky ocean
relationships	Murugan and Siva; Murugan and Arjuna
keywords	Murugan, devotion, protection, spirituality

Table 6: Example annotated record from the Thiruppugazh dataset