Constructing Knowledge Graphs from Text Using Large Language Models: Scoping Review

Giovanna Borges Bottino[1], **José de Jesus Pérez Alcazár**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
SP – Brazil

{giovanna.bottino,jperez}@usp.br

Abstract. *The use of Large Language Models (LLMs) for Knowledge Graph (KG) construction has gained significant traction, yet the field lacks methodological standardization. This study conducts a scope review following the PRISMA framework to map existing techniques and categorize them into four main approaches: (i) RDF-Based; (ii) Prompt-Based; (iii) RAG-Based; (iv) Hybrid Pipelines. Analyzing 126 primary studies, we identify key benefits such as scalability and automation, alongside challenges like low precision and manual curation. Our findings highlight research art-state.*

1. Introduction

Organizations generate large volumes of documentation throughout their projects, whether in financial reports, employee performance records, design and engineering documents, and more. The majority of this data is unstructured, accounting for 90% of the data generated by organizations [Muscolino et al. 2023]. However, most organizations have limited visibility into the content of these documents [Negro et al. 2022]. Additionally, approximately 58% of unstructured data is reused more than once after its initial use [Muscolino et al. 2023].

This data represents a valuable source of information, and organizations that learn to process, manage, and analyze it can gain a competitive advantage over those that do not [Negro et al. 2022]. Leveraging unstructured data creates opportunities to enhance productivity and manage the complexity of data [Muscolino et al. 2023]. If a sufficiently rich semantic model could be expressed in a machine-readable format, it could derive sophisticated insights and even perform causal reasoning [Kejriwal et al. 2021].

Transforming unstructured data into knowledge is a complex process that involves multiple stages [Negro et al. 2022]. The challenges associated with enabling machines to understand natural language have driven the adoption of Knowledge Graphs (KGs). A Knowledge Graph is a machine-readable and practical way of representing information about the world. It consists of an interconnected set of facts describing real-world entities and their relationships, making them understandable both by humans and computational systems [Kejriwal et al. 2021].

KGs have proven to be effective tools for enriching business processes with contextualized content and personalized responses [Zhong et al. 2023]. However, building knowledge graphs remains a challenge, and a systematic solution capable of automatically constructing a knowledge graph from unstructured data would represent a significant advancement over manual efforts [Zhong et al. 2023].

Automatically creating a knowledge graph requires advanced techniques to represent entities and relationships as a network [Negro et al. 2022]. This process includes entity recognition, coreference resolution, and relation extraction. Entity recognition identifies mentions of entities within the data. Coreference resolution identifies pairs of mentions that refer to the same entity. Relation extraction establishes semantic connections between entities [Zhong et al. 2023].

Leveraging LLMs offers a faster and more efficient alternative than manually creating [Negro et al. 2022]. These models handle spelling errors, variations in entity names, and gaps in incomplete information while also understanding complex linguistic contexts [Alammar and Grootendorst 2024].

Understanding the methods for constructing Knowledge Graphs from text using large language models is crucial due to their growing role in automating knowledge extraction and structuring. A scope review of this topic is essential to map existing techniques, compare their effectiveness, and identify gaps in the literature. This helps researchers and practitioners select appropriate methodologies, refine current approaches, and advance the field by addressing limitations and optimizing KG construction.

This article is structured into several sections. The second section, Methodology, outlines the research protocol, detailing the review conducted and reporting process. Finally, the article analyzes the KG construction methods identified in the review and concludes with a discussion of key findings.

2. Methodology

This study conducted a scoping review to identify key concepts, available types of evidence, and gaps in existing research, thereby mapping the field of study. The review followed the structured methodology outlined by [Institute 2015], which included defining the objective and guiding research question, establishing inclusion and exclusion criteria, selecting data sources, collecting and organizing results, and presenting the findings.

The review process was divided into three phases: planning, which involved developing the research protocol; execution, which entailed conducting the review; and reporting, which consisted of summarizing and presenting the results. Despite the process, this research field is rapidly evolving, which may affect the completeness of the review.

2.1. Research Protocol

The research protocol for this state-of-the-art review aimed to outline the process for identifying and analyzing the existing body of knowledge in the literature. Below, we describe the research questions, steps, and criteria applied in conducting the review.

2.1.1. Research Questions

The study aimed to identify primary research and examine existing approaches in the literature that utilize the construction of Knowledge Graphs from Text Sources using large language models. The key research question was framed using the PCC acronym (P = Population; C = Concept; C = Context) [Institute 2015]. Where the Population is Knowledge Graphs, the Concept is Large Language Models, and the Context is Raw Text as a Data Source:

"What techniques are associated with the use of LLMs in constructing knowledge graphs from Text Sources?"

To guide this review, the following research questions were formulated, each with specific motivations, objectives, and methodological strategies:

Q1. Which LLMs have been most commonly used for this purpose? Q2. In which domains have LLMs been most frequently applied for this purpose?

Q3. What are the key advantages of using LLMs for this purpose?

Q4. What are the key disadvantages of using LLMs for this purpose?

2.1.2. Eligibility Criteria

The study selection strategy was based on and adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, which consists of four main phases: Identification, Screening, Eligibility, and Inclusion. The adoption of PRISMA ensured a standardized, transparent, and rigorous selection process [Page et al. 2021].

The research topic guided the formulation of the eligibility criteria definition. Only studies that met all inclusion criteria and did not meet exclusion criteria proceeded to the final stage of data extraction and analysis.

Exclusion Criteria

- **EC1**: The article was not written in English.
- **EC2**: The article was not peer-reviewed (e.g., reports, editorials).
- **EC3**: The article was a secondary or tertiary study.

Inclusion Criteria

- **IC1**: The article addressed knowledge graph construction.
- **IC2**: The article focused on knowledge graph construction from textual data.
- **IC3**: The article proposed a method utilizing LLMs for knowledge graph construction from text.
- **IC4**: The article explains the technique or method developed.
- **IC5**: The article included an evaluation of the generated knowledge graphs.

2.1.3. Information Sources

The choice of database was based on the following criteria: it had to provide a web-based search mechanism, contain relevant sources, and support searches by title and abstract. Based on these requirements, Scopus, IEEE Xplore, ScienceDirect, and ACM were selected as the databases due to their extensive collections of abstracts and citations from peer-reviewed sources.

2.1.4. Search Strategy

The search string was formulated based on the research topic and guiding questions. *"Knowledge Graph"AND ("Large Language Model"OR "LLM"OR "Generative Pre-trained Transformers") AND text*.

2.1.5. Data Extraction and Synthesis Strategy

Data extraction was carried out in three stages: (i) compiling a reading form to organize information obtained from full-text reading; (ii) conducting a detailed analysis of concepts relevant to the research questions; and (iii) synthesizing the data for integration into the systematic review.

The reading form included bibliographic details, such as author, title, publisher, publication date, and country, as well as content-related information, including the general theme, problem or hypothesis, methodology, results, and key contributions. After completing the reading form, a comprehensive report was prepared to address the research questions and consolidate findings in the review synthesis phase.

Figure 1 illustrates the PRISMA flowchart, which represents the study selection process in a systematic review. The process is structured into three main phases: Identification, Screening, and Inclusion. The flowchart follows a linear sequence, depicting the selection process step by step. On the left, vertical labels identify each phase, while the boxes on the right outline the applied inclusion and exclusion criteria.

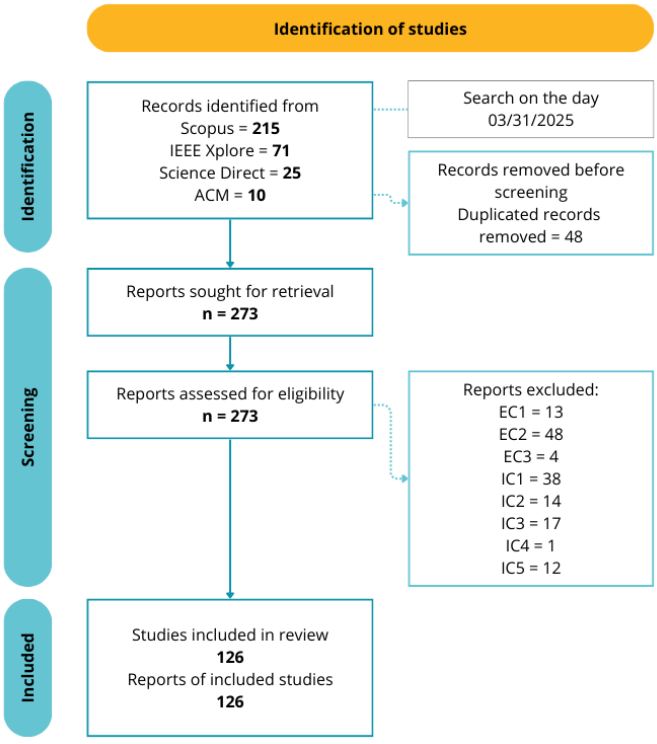

Figura 1. Review flowchart

In the Identification phase, a search was conducted on Scopus, IEEE Xplore, Science Direct, and ACM on March 31, 2025, retrieving 321 studies, with 48 duplicates removed. After filtering by language (EC1), peer-review status (EC2), and study type (EC3), 208 studies remained.

Abstract screening further reduced the set to 140. Among these, 170 addressed knowledge graph construction (IC1), 156 focused on textual data (IC2), and 139 utilized large language models (IC3). After full-text screening, 138 studies explained the developed technique (IC4), and 126 included an evaluation of the generated knowledge graphs (IC5).

2.2. Reporting the review

This section aims to present and analyze research on. Reporting and answering the research questions. The data can be accessed through the Google Drive [1].

2.2.1. Which LLMs have been most commonly used for this purpose?

Among the 126 studies analyzed, 36 compared multiple models in their approaches. The most frequently used LLM was OpenAI GPT, followed by LLAMA and Chat GLM. The OpenAI GPT, popularly known as ChatGPT, had the most commonly used versions, GPT-4 and its branches. This high adoption of ChatGPT-4 is probably because of the easy access to it. As it is online, it doesn't require installation, and some branches, such as GPT -4o, are free with usage limits.

2.2.2. In which domains have LLMs been most frequently applied for this purpose?

Overall, the analyzed articles are not limited to a single application domain. When specified, each study addressed various themes. General knowledge stood out the most, followed by Health and Manufacturing Processes. This indicates that aside from the Health domain, the Knowledge Graph is multi-domain. Used and experimented in many domains.

2.2.3. What are the key advantages of using LLMs for this purpose?

The studies identified three main advantages of using LLMs for knowledge graph construction from text: automation, scalability, and precision. Automation reduces the need for manual effort in extracting triples, entities, and relationships. Scalability refers to the ability to handle large datasets through adaptable and reusable frameworks across domains. Precision is reflected in the improved accuracy of entity and relation identification, as well as in the generation of consistent triples.

[1] https://drive.google.com/drive/folders/1GVBULi4RgmpVqri2MreDfsoAfTEELe6A?usp=sharing

2.2.4. What are the main limitations of using LLMs for this purpose?

Some studies reported both advantages and limitations in the same areas, reinforcing that the field is still evolving and that no single approach fully addresses all challenges.

Regarding limitations, the studies highlighted three recurring issues: manual curation, limited generalization, and low precision. Manual curation highlights the ongoing requirement for human intervention, whether in system configuration, ontology development, or rule-based setups. Limited generalization reflects the difficulty of adapting techniques to new domains, often due to rigid models or domain-specific dependencies. Low precision relates to challenges such as false positives, semantic ambiguity, and systemic errors, especially in tasks like named entity recognition and triple generation.

3. Constructing KG from text source using LLM

To understand the methods, excerpts of texts describing the methods applied in the study were extracted from each article. The excerpts were analyzed and grouped by similarity, as presented in Figure 2, resulting in four groups: RDF-Based, Prompt-Based, RAG-Based, and Hybrid Pipelines.

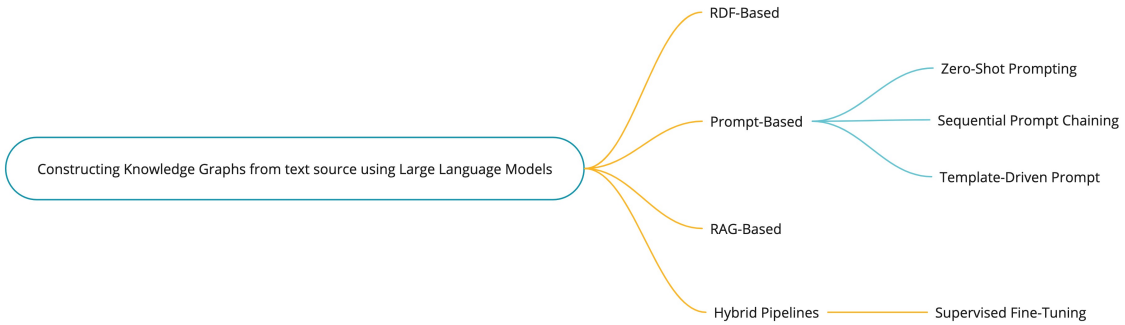

Figura 2. Categories and sub categories

3.1. RDF-Based

RDF-Based Graph Construction is the most adopted category, with 50 studies. This approach aims to produce knowledge graphs with RDF syntax and tools. LLMs help convert raw data into valid RDF triples. This approach seeks to build graphs that are semantically compatible with existing formal ontologies. It is ideal for scenarios where interoperability, semantic consistency, and connection to Linked Open Data are desired.

After defining the domain and selecting existing ontologies, the user generates a prompt that sends the existing ontologies as text, a link, or a description, along with the text. In this category, the most used LLM was OpenAI GPT, primarily GPT-3 and GPT-3.5.

Following this, each extracted element is aligned with established vocabularies. The data is serialized in RDF, which may be supported by LLMs. Finally, the resulting triples are stored in a triple store, where they can be queried using SPARQL. They are evaluated using Recall, variants, and F1-score. The outputs often do not match the quality typically achieved by human experts, since it cannot resolve contradictions or conflicting knowledge within RDF.

3.2. Prompt-Based

This category comprises 35 studies, emphasizing the use of prompt engineering to directly extract structured triples. With this method, no supervised fine-tuning is required; it just emphasizes prompt design without relying on predefined ontologies.

The process also begins with a corpus of text. General-purpose documents predominate in this category, with some studies focusing on policy and education. A prompt is designed, it can be a template, zero-shot, or few-shot, and passed to an LLM, primarily OpenAI GPT. The model outputs structured triples, which may be cleaned and stored in a graph database or serialized as RDF.

This method requires no training and is easy to adapt to multiple domains. However, hallucinations persist even with prompt refinement and have poor support for complex relations.

It was divided into subcategories: Zero-Shot Prompting, Sequential Prompt Chaining, and Template-Driven Prompt.

3.2.1. Zero-Shot Prompting

This subcategory comprises 19 studies and is most commonly used in healthcare, including biomedical research and clinical documentation.

Zero-shot prompting refers to a prompting approach that doesn't specify reference examples. It is the standard approach to prompting, where the LLM is provided with an instruction and optionally some input text, but no examples or demonstrations are provided on how to solve the task [Berryman and Ziegler 2024].

The LLM generalizes based on a generic prompt to guide extraction, without examples or additional context. It relies entirely on the model to predict the correct behavior. This method is beneficial when annotated data is scarce and rapid adaptation to new domains is needed.

3.2.2. Sequential Prompt Chaining

This subcategory comprises 12 studies and was utilized by General domains, including legal and educational content, as well as cross-modal content such as text-image pairs, where OpenAI GPT remains dominant, followed by LLaMA and ChatGLM.

Sequential Prompt Chaining is a technique used to break down complex tasks into smaller, sequential steps. This allows the model to focus more on each individual question and improving overall performance [Alammar and Grootendorst 2024]. Each stage serves a specific function, such as extraction, validation, enrichment, or contextual expansion. This modularity allows better control and improved consistency in generating structured knowledge.

3.2.3. Template-Driven Prompt

This subcategory comprises four studies that utilized Baidu ERNIE 4.0 in healthcare settings. Template-Driven Prompt is a prompt engineering technique that generates dynamic prompts with variable substitution. This approach is particularly useful in scenarios where the content of the prompt needs to change based on the context or input provided by the user [Berryman and Ziegler 2024]. It uses well-defined prompt structures (e.g., "extract (subject, predicate, object)"). It is beneficial in clinical or regulatory settings where reproducibility is crucial.

This use of prompts gives the researcher the impression that they are guiding the LLM. However, the model is probabilistic, not deterministic, and results vary with small changes in the prompt.

3.3. RAG-Based

This category comprises 22 studies, primarily used in manufacturing processes, including quality control and industrial diagnostics. It is based on Retrieval-Augmented Generation (RAG). It is a technique in artificial intelligence that combines information retrieval with text generation. RAG systems enhance LLMS by first searching external information to find relevant information in response to a user query [Li et al. 2024].

The goal is to enhance LLM with external knowledge sources, such as document bases, wikis, databases, or existing graphs, before or during the triple generation process.

The pipeline begins with a user-provided input, such as a document or a short text. This input is preprocessed and passed to an RAG module, which performs a semantic search over an indexed corpus to retrieve relevant documents. These documents, along with the original input, are combined into a single enriched prompt that is then fed into the LLM.

The LLM generates subject-predicate-object triples based on this contextual information. These are then evaluated and integrated into the existing graph. Then, it is usually assessed by precision and Recall.

In most cases, the LLM generates new triples without access to the existing graph, which prevents it from verifying whether the information already exists or conflicts with prior assertions. Unless a post-processing module performs explicit checks, overlapping triples may be added to the graph.

This category has a high accuracy in triple generation, but does not incorporate symbolic reasoning or graph inference, and has no automated validation or integration with RDF-based ontologies.

3.4. Hybrid Pipelines

This category encompasses 19 studies that combine traditional Natural Language Processing (NLP) techniques, such as Named Entity Recognition and Relation Extraction, with LLM-based reasoning. NLP involves computational techniques for the automatic analysis and representation of human language, aiming to process and understand its structure, meaning, and context [Nandigam et al. 2023].

It employs LLMs for post-processing, semantic validation, or contextual augmentation to enhance accuracy. The workflow begins with raw documents, typically from the healthcare industry. Then, NLP tools (e.g., spaCy, BERT, dependency parsers) extract initial entities and relationships.

The output is then passed to an LLM, the most used of which was Owen, which refines, validates, or augments it with semantic or contextual depth. Final triples are aligned with ontologies and stored. Recall that SimKGC and WN18RR were the metrics used for evaluation.

This method, like the others, drastically reduces time and cost. But it requires expert validation for higher precision, since some models presented only 68% initial accuracy.

3.4.1. Supervised Fine-Tuning

This subcategory encompasses 13 studies in which models are fine-tuned on labeled datasets to enhance the extraction of entities and relationships for knowledge graph construction. Fine-tuning is a process in machine learning where a pre-trained model is further trained to adapt it for a particular application [Ma et al. 2021]. These models may include transformer-based architectures, such as BERT or LLaMA, which are adapted for domain-specific tasks, typically in the biomedical field.

The fine-tuning approach utilizes a pre-trained model, which is then further trained on a labeled dataset specific to the new task [Ma et al. 2021]. The process of creating a KG from text begins with this fine-tuned model, which does not rely on the LLM for inference. The LLM is a semantic refiner with a secondary role. The resulting triples are verified, cleaned, and inserted into a structured graph.

Although the flow differs from the structure described in the parent category, this subcategory falls under Hybrid Pipelines due to its integration of supervised model training, typically using large language models with domain-specific entity and relation extraction tasks.

Despite being easier to create high-confidence knowledge graphs, this approach requires large annotated datasets, which may not always be readily accessible.

3.5. Summary

The analysis identified four distinct groups of LLM-based methods for knowledge graph construction, each varying in complexity and autonomy. These categories differ in their level of semantic formalism, reliance on prompting, architectural complexity, and intended use cases.

One of the fundamental distinctions lies in the degree of alignment with formal ontologies. In contrast to RDF-based approaches that explicitly rely on vocabularies, the Prompt-based approaches extract triples using open-ended instructions without reference to ontologies. RAG-Based methods sit somewhere in between: while they benefit from contextual richness via document retrieval, they typically do not enforce alignment with any formal schema. Hybrid pipelines may or may not incorporate ontological structures, depending on how the classical NLP components are configured.

Prompting itself is another axis of divergence. Prompt-based methods position the LLM as the core inference engine. This creates significant sensitivity to prompt formulation and introduces variability across runs. RDF-based and hybrid approaches, by contrast, treat the LLM as a component within a controlled framework, using it for specific tasks such as refinement or disambiguation rather than primary extraction. RAG-based inference methods also rely on prompting, but with an additional level of complexity: the prompt is dynamically constructed using retrieved context.

Supervision and structure further differentiate these paradigms. Hybrid pipelines are the only class that typically incorporates supervised learning components such as fine-tuned NER and RE models. This introduces a higher initial cost in terms of data preparation and training, but offers more deterministic behavior and task-specific accuracy. Prompt-based and RAG-based approaches are largely unsupervised and more general-purpose, making them suitable for low-resource domains; however, they are less reliable for high-stakes applications.

Reproducibility and semantic robustness also vary significantly. RDF-based methods and Hybrid methods are the most reproducible. In contrast, prompt-based and rag-based methods are highly sensitive to non-deterministic factors, making them more fragile in applied settings.

Finally, these approaches differ in their capacity for formal evaluation. In essence, RDF-based and hybrid pipelines prioritize semantic control, structural rigor, and interpretability, while prompt-based and rag-based methods offer flexibility, scalability, and accessibility at the cost of reproducibility and formal grounding.

4. Conclusion

This study explores the use of Large Language Models in constructing Knowledge Graphs from unstructured text sources. It provides a thorough analysis of the state of the art, reviewing previous research and identifying gaps. The methodology follows the PRISMA framework, ensuring transparency and reproducibility. By analyzing 126 primary studies, the review identified four main methodological categories: RDF-Based, Prompt-Based, RAG-Based, and Hybrid Pipelines. Each category presents distinct characteristics in terms of semantic formalism, dependency on prompts, architectural complexity, and applicability across domains.

Despite these advances, the field still faces several limitations. The studies reviewed highlighted persistent challenges, including low precision, the need for manual curation, and limited generalization across domains. Moreover, methodological heterogeneity, variability in evaluation metrics, and the absence of standardized benchmarks complicate direct comparison between approaches.

As limitations of this review, the rapid evolution of the field may have led to the exclusion of very recent research that has not yet been indexed in the consulted sources. For future research, there is a clear need to develop standardized metrics for evaluating LLM-generated Knowledge Graphs. Explore new models that mitigate hallucinations and enhance extraction reliability to improve accuracy and computational efficiency.

Referências

Alammar, J. and Grootendorst, M. (2024). *Hands-On Large Language Models*. O'Reilly Media.

Berryman, J. and Ziegler, A. (2024). *Prompt Engineering for LLMs: The Art and Science of Building Large Language Model–Based Applications*. "O'Reilly Media, Inc.".

Institute, J. B. (2015). *The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews*. Adelaide, Australia.

Kejriwal, M., Knoblock, C. A., and Szekely, P. (2021). *Knowledge graphs: Fundamentals, techniques, and applications*. MIT Press.

Li, Z., Xu, T., Wang, W., Wu, H., Xiong, F., Chen, E., Lyu, Y., Niu, S., Liu, H., and Tang, B. (2024). Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43:1 – 32.

Ma, Y., Chen, Z., and Church, K. (2021). Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27:763 – 778.

Muscolino, H., Machado, A., Vesset, D., and Rydning, J. (2023). Untapped value: What every executive needs to know about unstructured data. White paper, IDC, Framingham, MA. Sponsored by Box.

Nandigam, J., Patil, R., Boit, S., and Gudivada, V. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Negro, A., Kus, V., Futia, G., and Montagna, F. (2022). *Knowledge graphs applied*. Manning.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Zhong, L., Wu, J., Li, Q., Peng, H., and Wu, X. (2023). A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.