

Domain Ontology Modeling in the Digital Archive of the Islamic Revolution Document Center: A Hybrid Approach of Text Analysis and Ontology Reuse

Nader Naghshineh ¹ , Alireza Entehaee Saray ²  , Behrouz Minaei-Bidgoli ³ , and Ali Shabani ⁴ 

1. Department of Knowledge and Information Science, Faculty of Management, University of Tehran, Tehran, Iran. E-mail: nnaghsh@ut.ac.ir
2. *Corresponding Author*, Department of Knowledge and Information Science, Faculty of Management, University of Tehran, Tehran, Iran. E-mail: a.entehaee@ut.ac.ir
3. Department of Artificial Intelligence and Robotics, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran. E-mail: b_minaei@iust.ac.ir
4. Department of Knowledge and Information Science, Faculty of Management, University of Tehran, Tehran, Iran. E-mail: ali.shabani@ut.ac.ir

ABSTRACT

Organizing information in digital archives, particularly historical ones such as the Islamic Revolution Document Center of Iran, faces challenges like the inefficiency of traditional information retrieval systems. This center, with over 4.5 million document pages, 31,000 hours of oral history, and millions of news titles and articles, requires innovative approaches such as domain ontology and knowledge graphs to improve semantic access to key entities. The aim of this research is to model a domain ontology for the digital archive of this center using a hybrid approach of text analysis and reuse of existing ontologies. The research method is mixed: semi-automatic and automatic text analysis for entity extraction, model design based on ontology principles, and validation of findings using the nominal group technique. The findings include 535 classes that, after validation with criteria of 1) hierarchical logic, 2) historical/cultural accuracy, 3) alignment and equivalence, and 4) completeness and simplicity, resulted in 457 classes confirmed, 70 classes added, 53 classes relocated in the hierarchy, and 19 classes removed. This ontology provides a foundation for a knowledge graph, enhances semantic retrieval, and serves as a logical basis for artificial intelligence systems in managing Iran's historical documents. This research fills the gap in designing practical ontologies for contemporary Iranian history and is extensible to similar domains.

Keywords: Digital Archive, Domain Ontology, Entity Extraction, Islamic Revolution Document Center, Knowledge Graph, Semantic Information Retrieval, Semantic Search

Cite this Article: Naghshineh, N., Entehaee Saray, A., Minaei-Bidgoli, B., & Shabani, A. (2026). Domain Ontology Modeling in the Digital Archive of the Islamic Revolution Document Center: A Hybrid Approach of Text Analysis and Ontology Reuse. *Knowledge Retrieval and Semantic Systems*, 13(47), 29-54. <https://doi.org/10.22054/jks.2025.88985.1746>



© 2026 by The author(s) retain the copyright

Publisher: Allameh Tabataba'i University Press

DOI: <https://doi.org/10.22054/jks.2025.88985.1746>

EXTENDED ABSTRACT

Introduction

Information organization is a cornerstone of every information-retrieval system, since query effectiveness depends on how information is organized and represented. Digital archives, especially historical ones such as the Islamic Revolution Document Center (IRDC) of Iran, face a persistent challenge: traditional retrieval systems cannot provide comprehensive, meaning-based access to document content. Established in 1981 to compile the history of the Islamic Revolution and collect the documents and memoirs of Imam Khomeini's movement, the IRDC holds more than 4.5 million written-document pages, about 31,000 hours of oral-history interviews, nearly 372,000 photographs, more than 13 million news titles, and around 90,000 domestic and foreign articles. Although much of this material has been digitized through optical character recognition (OCR) and automatic speech recognition (ASR), the center's retrieval system relies mainly on descriptive metadata and full-text search, so that key entities (persons, places, events, and concepts) remain hidden, limiting comprehensive access. This calls for a redesign based on knowledge-graph and ontology approaches that can extract and organize entities semantically. This research models a domain ontology for the IRDC digital archive through a hybrid approach combining text analysis with the reuse of existing ontologies, establishing a foundation for a knowledge graph that supports semantic retrieval and AI systems for managing Iran's historical documents.

Research Questions

1. What are the entities and classes of the subject domain of the IRDC digital-archive documents, and how can they be represented within a domain ontology?
2. What should the hierarchical structure and the relationships among the ontology classes be?
3. Which ontologies and classification systems are most suitable for reuse in building the IRDC domain ontology?

Literature Review

Research on ontology- and knowledge-graph-based knowledge organization falls into several strands: automatic or hybrid extraction of structured information from text (Emami, 2022; Moradi et al., 2015; Gutierrez et al., 2016; Wimalasuriya & Dou, 2010); knowledge-graph development, including a Persian knowledge graph (Sajjadi & Minaei Bidgoli, 2019) and FarsBase (Asgari-Bidhendi et al., 2019); knowledge-graph-based question answering (Yang et al., 2023; Tong et al., 2019); human-driven ontology design in specialized domains (Torabi et al., 2022; Sadeghi Niaraki et al., 2018; Servati et al., 2017); knowledge-graph-based retrieval and search engines (Homayouni et al., 2018; Li, 2019; Wang et al., 2019); ontology-based retrieval optimization (Jafari Pavarsi et al., 2020; Asim et al., 2019); and thesaurus enrichment for ontology construction (Hosseini Beheshti & Ejei, 2015; Wu, 2018). A review of this body of work shows that, although these studies advance meaning-based organization, most rely on relatively structured or general sources, including Wikipedia, web tables, scientific databases, and technical texts, rather than real archival environments marked by heterogeneous sources, diverse formats, and semantic ambiguity, and they emphasize technical extraction methods over practical application in operational systems. The present study therefore occupies a distinct, applied, problem-driven position, designing and implementing, for the first time, a domain ontology for contemporary Iranian history within a real digital-archive context and addressing the gap in the domestic literature.

Methodology

A mixed-method design was adopted. Informational entities were identified through semi-automatic and automatic text analysis, and the ontology was designed following the best practices of Arp et al. (2015) for building domain ontologies on Basic Formal Ontology (BFO): defining the scope, identifying universals from existing ontologies and domain texts, arranging them in a general-to-specific hierarchy, ensuring logical coherence and human-readable definitions, and formalizing the result in a computer-usable language. The scope was limited to contemporary Iranian history, that is, the second Pahlavi period, the Islamic Revolution and its aftermath, and the Islamic Republic to the present, with fine-grained classes for persons, parties, governance periods, organizations, places, and related events.

The research population comprised all digitized resources of the archive, from which a purposive sample was selected in consultation with the center's managers and experts. Two tools supported entity recognition: the open-source annotation software INCEpTION for manual and semi-automatic tagging, and a large language model (Grok v3.0) for automatic named-entity recognition, whose errors were corrected under human supervision. Existing ontologies and classifications were analyzed with Protégé and SPARQL queries over RDF/OWL files, reusing DBpedia, the Persian knowledge-graph ontology (FarsBase), BIBFRAME, and the Library of Congress classification for the history of Iran. The hierarchy was built around an *is_a* backbone with single inheritance, the open-world assumption, and the objectivity principle, using DBpedia as the base. Finally, the classes were validated through the Nominal Group Technique with eleven managers and experts across two three-hour sessions, applying four criteria: hierarchical logic, historical and cultural accuracy, alignment and conceptual equivalence, and completeness and simplicity.

Results

The combination of text analysis with the reuse of DBpedia, BIBFRAME, and the Library of Congress Iranian-history classification produced 535 terms, organized under main classes including agent, person, organization, family, event, work, government period, subject concept, place, architectural structure, and ethnic group. After validation by the expert group and majority consensus, 457 classes were confirmed, 70 were added, 53 were relocated within the hierarchy, 19 were removed, 5 terms were renamed, and 2 classes were both relocated and renamed.

Most of the additions concerned the agent–person and place branches, and a dedicated subject-concept branch was created for intellectual, political, social, and economic currents, schools of thought, and ideologies. The architectural-structure branch was merged into the place branch as spatial subdivisions, and several unconventional or overlapping class labels were corrected or removed. Regarding the third question, no domain ontology for contemporary Iranian history was found domestically; the Library of Congress expansion was preferred over the Dewey-based one because it subsumes the Dewey classes, DBpedia was confirmed as the core general ontology (the Persian knowledge graph differing in only ten classes), and BIBFRAME served as the reference for the work-related classes.

Discussion

The study yielded a validated, extensible base ontology for contemporary Iranian history grounded in the IRDC archive. Its results agreed with earlier Persian-language extraction studies and with hybrid ontology-based extraction approaches, but its data derive from authentic, semi-structured historical documents rather than structured or general texts, with

emphasis on a native conceptual structure suited to the Iranian historical domain rather than on extraction algorithms. The scale of the expert-driven revision showed that general structures such as DBpedia or the Persian knowledge graph cannot accurately cover native Iranian-history concepts without extensive localization, consistent with Sajjadi and Minaei Bidgoli (2019) and Asgari-Bidhendi et al. (2019). Unlike purely conceptual ontology-design studies, this work was validated through practical feedback from archival experts in a real institutional setting, making its output directly applicable within the digital-archive system.

Conclusion

By combining text analysis with the reuse of established ontologies and rigorous expert validation, this research produced a base domain ontology for contemporary Iranian history within the real context of the IRDC digital archive. The ontology provided a foundation for a historical knowledge graph, enhanced semantic retrieval, and served as a logical basis for artificial-intelligence systems that manage the country's historical documents, filling a gap in practical ontology design for this domain. Future research should develop interoperable domain ontologies in other subject areas, implement and evaluate an operational knowledge graph, advance more automated extraction of concepts and relations with large language models, and study relation-weighting algorithms.

Ethical Considerations

Author Contributions

This article is derived from the doctoral dissertation of Alireza Entehaee Saray. Nader Naghshineh (supervisor): study design, supervision of the research process, verification and validation of the results, review and finalization of the manuscript. Alireza Entehaee Saray: conceptualization, data collection, text analysis and entity extraction, ontology modelling, formal analysis and preparation of the original draft. Behrouz Minaei-Bidgoli (supervisor) and Ali Shabani (advisor): participation in study design, research supervision, and review and revision of the manuscript.

Data Availability Statement

Data available on request from the authors.

Acknowledgements

The authors would like to thank the Head and the Deputy of the Islamic Revolution Document Center for supporting this research and providing the data and infrastructure required, and the expert panel for the time and knowledge they contributed to this study.

Ethical Considerations

The authors avoided data fabrication, falsification, plagiarism, and misconduct.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

“The authors declare no conflict of interest.”

مدل‌سازی هستی‌شناسی دامنه در آرشیو دیجیتال مرکز اسناد انقلاب اسلامی: رویکرد ترکیبی تحلیل متن و بازاستفاده از هستی‌شناسی‌ها

نادر نقشینه^۱، علیرضا انتهای سرای^۲، بهروز مینایی بیدگلی^۳، و علی شعبانی^۴

۱. گروه علم اطلاعات و دانش‌شناسی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: nnaghsh@ut.ac.ir

۲. نویسنده مسئول، گروه علم اطلاعات و دانش‌شناسی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران. رایانامه:

a.entehaee@ut.ac.ir

۳. گروه هوش مصنوعی و رباتیک، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران. رایانامه:

b_minaei@iust.ac.ir

۴. گروه علم اطلاعات و دانش‌شناسی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: ali.shabani@ut.ac.ir

چکیده

سازمان‌دهی اطلاعات در آرشیوهای دیجیتال، به‌ویژه آرشیوهای تاریخی مانند مرکز اسناد انقلاب اسلامی ایران، با چالش‌هایی مانند ناکارآمدی نظام‌های سنتی بازیابی اطلاعات مواجه است. این مرکز، با بیش از ۴/۵ میلیون برگ سند، ۳۱ هزار ساعت تاریخ شفاهی، و میلیون‌ها عنوان خبری و مقاله، نیاز به رویکردهای نوین مانند هستی‌شناسی دامنه و گراف دانش دارد تا دسترسی معنایی به موجودیت‌های کلیدی را بهبود بخشد. هدف این پژوهش، مدل‌سازی هستی‌شناسی دامنه برای آرشیو دیجیتال این مرکز با رویکرد ترکیبی تحلیل متن و بازاستفاده از هستی‌شناسی‌های موجود است. روش پژوهش آمیخته است: از تحلیل متن به شکل نیمه‌خودکار و خودکار برای استخراج موجودیت‌ها، طراحی مدل بر اساس اصول هستی‌شناسی و اعتبارسنجی یافته‌ها با تکنیک گروه اسمی استفاده شده است. یافته‌ها شامل ۵۳۵ رده است که پس از اعتبارسنجی با معیارهای (۱) منطق سلسله‌مراتبی، (۲) صحت تاریخی/فرهنگی، (۳) هم‌ترازی و معادل‌سازی و (۴) کامل بودن و سادگی، ۴۵۷ رده تأیید، ۷۰ رده اضافه‌شده، ۵۳ رده در سلسله‌مراتب جابجا و ۱۹ رده حذف شدند. این هستی‌شناسی مبنایی را برای گراف دانش فراهم می‌کند، بازیابی معنایی را ارتقا می‌دهد، و پایه‌ای منطقی برای سیستم‌های هوش مصنوعی در مدیریت اسناد تاریخی ایران است. این پژوهش خلأ طراحی هستی‌شناسی کاربردی در تاریخ معاصر ایران را پر کرده و قابل توسعه برای دامنه‌های مشابه است.

کلیدواژه‌ها: آرشیو دیجیتال، استخراج موجودیت، بازیابی اطلاعات معنایی، جست‌وجوی معنایی، گراف دانش، مرکز اسناد انقلاب اسلامی، هستی‌شناسی دامنه

استناد به این مقاله: نقشینه، نادر، انتهای سرای، علیرضا، مینایی بیدگلی، بهروز، و شعبانی، علی. (۱۴۰۵). مدل‌سازی هستی‌شناسی دامنه در

آرشیو دیجیتال مرکز اسناد انقلاب اسلامی: رویکرد ترکیبی تحلیل متن و بازاستفاده از هستی‌شناسی‌ها. *بازیابی دانش و*

نظام‌های معنایی، ۱۳ (۴۷)، ۲۹-۵۴. <https://doi.org/10.22054/jks.2025.88985.1746>

© ۲۰۲۶ نویسنده(گان)

ناشر: دانشگاه علامه طباطبائی



مقدمه

سازمان‌دهی اطلاعات و دانش یکی از ارکان هر نظام بازیابی اطلاعات است. درواقع، پرس‌وجوی کاربر با نمایه‌های نظام بازیابی به مدارک و اطلاعات هدایت می‌شود و میزان اثربخشی آن به نوع و روش سازمان‌دهی اطلاعات و کارکرد صحیح آن مرتبط می‌شود. تاکنون روش‌های مختلفی برای سازمان‌دهی اطلاعات و منابع اطلاعاتی تعریف شده و به کار گرفته شده‌اند که ازجمله آن‌ها می‌توان به نظام‌های رده‌بندی، نمایه‌سازی، فهرست‌نویسی، نظام‌های کدگذاری و جست‌وجوی متن آزاد اشاره کرد. برای پشتیبانی از این نظام‌ها منابع مختلفی نیز به‌عنوان طرح‌های رده‌بندی، سرعنوان‌های موضوعی، اصطلاح‌نامه‌ها، مستندات نام‌های اشخاص، مکان‌های جغرافیایی و سازمان‌ها توسعه پیدا کرده‌اند. همچنین استانداردهای مختلفی برای توصیف، کدگذاری و مدیریت ساختار منابع تدوین شده‌اند تا امکان ذخیره‌سازی و بازیابی اطلاعات به شکل یکدست را فراهم کرده و امکان تعامل‌پذیری بین سامانه‌های اطلاعاتی را فراهم کنند. با ظهور عصر دیجیتال و توسعه سریع استانداردهای حوزه داده و اطلاعات و فناوری‌های جدید کتابخانه‌ها و آرشیوهای دیجیتالی گروه‌های مختلف به دنبال راهکارهای نوینی برای سازمان‌دهی بهینه حجم بالایی از اطلاعات دیجیتال هستند. رویکردهای دهه اخیر در این حوزه متمرکز بر وب معنایی و داده‌های پیوندی، نسل جدید فهرست‌های کتابخانه‌ای، چارچوب دسترسی و توصیف منبع^۱، سازمان‌دهی و به اشتراک‌گذاری تصاویر دیجیتال، جست‌وجو و مرور چهریزه‌ای و استانداردهای کیفیت فراداده‌ها بوده است (Park & Howarth, 2013).

یکی از رویکردهای نوین برای سازمان‌دهی اطلاعات و دانش استفاده از هستی‌شناسی (یا آنتولوژی) و در مفهوم دیگر توسعه گراف‌های دانش است. در هستی‌شناسی تمرکز تنها بر شناسایی مفاهیم به‌عنوان نقاط دسترسی نیست، بلکه جایگاه مفاهیم در دنیای هستی و نحوه ارتباط و اثرگذاری آنها نیز موردتوجه قرار می‌گیرد؛ بنابراین آنچه مورد پرس‌وجو قرار می‌گیرد صرفاً یک نمایه مرتبط با مدارک نیست، بلکه شبکه یا گرافی از دانش مورد کاوش و بازنمایی آن است. گراف دانش^۲ که به‌عنوان یک حوزه اصلی در بازیابی اطلاعات و پشتیبانی از هوش مصنوعی ظاهر شده است، اولین بار توسط گوگل برای بازنمون دانش از حجم وسیعی از اطلاعات مورد استفاده قرار گرفت (Kejriwal et al., 2021). گراف‌های دانش مجموعه‌ای بزرگ از موجودیت‌های مرتبط به هم هستند که به‌وسیله برچسب‌های معنایی غنی شده‌اند. در اینجا منظور از موجودیت انواع موجودیت‌های نامدار و غیر نامدار مانند اسامی اشخاص، مکان، سازمان، رویداد، زمان، مفاهیم و... است (سجادی و مینایی بیدگلی، ۱۳۹۸).

یکی از چالش‌های کلیدی آرشیوهای دیجیتال، به‌ویژه آرشیوهای تاریخی مانند آرشیو دیجیتال مرکز اسناد انقلاب اسلامی، ناکارآمدی نظام‌های سنتی بازیابی اطلاعات در ارائه دسترسی جامع و معنایی به محتوای اسناد است. این مرکز که در سال ۱۳۶۰ با هدف تدوین تاریخ انقلاب اسلامی و جمع‌آوری اسناد، مدارک و خاطرات مرتبط با نهضت امام خمینی تأسیس شد، آرشیوی غنی شامل ۴/۵ میلیون برگ سند مکتوب، ۳۱ هزار ساعت مصاحبه تاریخی شفاهی، ۳۷۱،۹۰۶ عکس، ۱۳ میلیون عنوان خبری و ۹۰،۰۰۰ مقاله داخلی و خارجی را در اختیار دارد. بخش قابل توجهی از این اسناد با استفاده از

فناوری‌های تشخیص کاراکتر نوری و تبدیل گفتار به متن به فرمت دیجیتال تبدیل شده‌اند، اما نظام فعلی بازیابی اطلاعات این مرکز عمدتاً بر فراداده‌های توصیفی اسناد و جستجوی درون‌متنی متمرکز است. این امر سبب شده تا بسیاری از موجودیت‌های اطلاعاتی کلیدی (مانند اسامی اشخاص، مکان‌ها، رویدادها و مفاهیم) در جستجوهای کاربران مخفی بمانند و دسترسی پژوهشگران به اطلاعات جامع محدود شود؛ بنابراین، ضرورت بازطراحی نظام بازیابی اطلاعات با استفاده از رویکردهای مبتنی بر گراف دانش و هستی‌شناسی به منظور استخراج و سازمان‌دهی معنایی موجودیت‌ها و بهبود دسترسی به اطلاعات احساس می‌شود. هدف این پژوهش، مدل‌سازی هستی‌شناسی دامنه برای آرشیو دیجیتال مرکز اسناد انقلاب اسلامی با استفاده از رویکرد ترکیبی تحلیل متن و بازاستفاده از هستی‌شناسی‌های موجود است. این پژوهش با تمرکز بر استخراج خودکار و نیمه‌خودکار موجودیت‌های اطلاعاتی (مانند اشخاص، سازمان‌ها، رویدادها و مفاهیم) و طراحی ساختار سلسله‌مراتبی، به دنبال ایجاد پایه‌ای برای گراف دانش است که بتواند از بازیابی اطلاعات معنایی پشتیبانی کند. این رویکرد ترکیبی، با بهره‌گیری از فناوری‌های پردازش زبان طبیعی و مدل‌های زبانی بزرگ و یکپارچه‌سازی هستی‌شناسی‌ها و رده‌بندی‌های دیگر مانند دی‌بی‌پدیا^۱، هستی‌شناسی گراف دانش فارسی^۲ و رده‌بندی رده DSR تاریخ ایران، امکان سازمان‌دهی بهینه و بازنمایی دانش موجود در آرشیو را فراهم می‌کند.

در این راستا، پژوهش با استفاده از فناوری‌های پردازش زبان طبیعی، مدل‌های زبانی بزرگ و تلفیق هستی‌شناسی‌ها و رده‌بندی‌های معتبر، به دنبال فراهم‌سازی بستر سازمان‌دهی و بازنمایی معنایی اطلاعات آرشیوی است. اهداف فرعی پژوهش عبارت‌اند از:

۱. شناسایی و استخراج خودکار و نیمه‌خودکار موجودیت‌های اطلاعاتی (ازجمله اشخاص، سازمان‌ها، رویدادها و مفاهیم) از اسناد آرشیو دیجیتال مرکز اسناد انقلاب اسلامی؛
 ۲. طراحی ساختار سلسله‌مراتبی رده‌ها و روابط میان موجودیت‌ها در چارچوب یک هستی‌شناسی دامنه؛
 ۳. شناسایی و بازاستفاده از هستی‌شناسی‌ها و رده‌بندی‌های موجود مناسب برای یکپارچه‌سازی با مدل پیشنهادی.
- بدین ترتیب، نتایج این پژوهش می‌تواند به توسعه یک گراف دانش تاریخی برای ایران و بهبود نظام‌های بازیابی اطلاعات معنایی در محیط‌های آرشیوی کمک کند.

پیشینه پژوهش

پژوهش‌های انجام گرفته در این حوزه به مقوله‌های استخراج خودکار یا ترکیبی اطلاعات ساختاریافته مبتنی بر مدل هستی‌شناسی، توسعه گراف دانش، پاسخگویی به پرس‌وجوها بر اساس گراف دانش، طراحی هستی‌شناسی به صورت انسانی، بهینه‌سازی بازیابی اطلاعات با هستی‌شناسی، بازیابی اطلاعات و موتورهای جست‌وجوی مبتنی بر پایه گراف دانش، بهینه‌سازی پرس‌وجو بر مبنای گراف دانش، طراحی سیستم‌های اطلاعاتی مبتنی بر هستی‌شناسی و غنی‌سازی اصطلاح‌نامه‌ها برای ساخت هستی‌شناسی توجه کرده‌اند.

امامی (۱۴۰۱)؛ مرادی و همکاران (۱۳۹۴)؛ گوتیرز و همکاران^۱ (۲۰۱۶) و ویمالاسوریا و دوو^۲ (۲۰۱۰) به حوزه استخراج خودکار یا ترکیبی اطلاعات به صورت ساختاریافته از درون متن پرداخته‌اند.

امامی (۱۴۰۱) به دنبال روش جدیدی برای استخراج خودکار اطلاعات ساختاریافته محدود به دامنه از متون فارسی صفحات وب که نیازی به دانش پیش‌زمینه در مورد موجودیت‌ها و ویژگی‌های آن‌ها ندارد، پرداخته و بر اساس آزمایشی کارایی آن را بهتر از روش‌های دیگر یافته است.

مرادی و همکاران (۱۳۹۴) با هدف ساخت هستی‌شناسی برای دانش عرفی زبان فارسی به کمک سه روش مبتنی بر الگو، ترجمه ماشینی و استفاده از منابع ساخت‌یافته به کمک سه پیکره زبان فارسی هفت نوع رابطه و در مجموع هفتاد هزار اظهار (رابطه) استخراج کرده و در قالب یک هستی‌شناسی ارائه کرده‌اند. میانگین دقت روابط استخراج‌شده این هستی‌شناسی برابر با ۷۵ درصد برای روش مبتنی بر الگو، ۷۰ درصد برای ترجمه ماشینی و ۱۰۰ درصد برای اطلاعات استخراج‌شده از جعبه اطلاعات بود.

گوتیرز و همکاران (۲۰۱۶) یک سیستم هیبرید را برای توسعه دو افزونه برای افزودن دقت استخراج ماشینی موجودیت‌ها و روابط آنها از متن پیشنهاد کرده‌اند.

ویمالاسوریا و دوو (۲۰۱۰) مقدمه‌ای بر استخراج خودکار اطلاعات به کمک مدل هستی‌شناسی پرداخته و جزئیات سامانه‌هایی که تاکنون برای این منظور توسعه یافته‌اند را بررسی کرده‌اند.

در مقوله توسعه گراف دانش، مینایی و سجادی (۱۳۹۸)، رجبی و اطمینانی^۳ (۲۰۲۲)؛ چویی^۴ (۲۰۲۲)، تان و همکاران^۵ (۲۰۲۱)، عسگری و همکاران^۶ (۲۰۱۹)، زکی و همکاران^۷ (۲۰۱۹) و رحمان و تاکاسو^۸ (۲۰۱۷) به شکل متمرکز به ساختن گراف دانش، توسعه آن یا بررسی ادبیات این مقوله پرداخته‌اند.

مینایی و سجادی (۱۳۹۸) به دنبال استخراج گراف دانش زبان فارسی از ویکی‌پدیای فارسی برای تهیه جامع‌ترین پایگاه دانش زبان فارسی به صورت عمومی و چند دامنه‌ای بوده که گرافی مشتمل بر ۵۰۰ هزار موجودیت و ۷ میلیون رابطه میان آن‌ها ارائه کرده‌اند.

رجبی و اطمینانی (۲۰۲۲) به مرور نظام‌مند پرداخته و در ۴۰ مقاله موضوع نحوه به‌کارگیری گراف‌های دانش در هوشمندی درونی سیستم‌های هوش مصنوعی را بررسی کرده‌اند و نتیجه گرفته‌اند که عمده این گراف‌ها به شناسایی و استخراج موجودیت‌ها و روابط آنها و همچنین تفسیر و استنتاج پرداخته‌اند.

چویی (۲۰۲۲) در مطالعه خود از خلاصه ادبیات پژوهش یک روش گسترش گراف دانش با درونی‌سازی گراف پیشنهاد کرده است.

-
1. Gutierrez et al.
 2. Wimalasuriya & Dou
 3. Rajabi & Etminani
 4. Choi
 5. Tan et al.
 6. Asgari-Bidhendi et al.
 7. Zaki et al.
 8. Rahman & Takasu

تان و همکاران (۲۰۲۱) به ساخت مدل گراف دانش و استنتاج دانش در حوزه ترافیک شهری پرداخته و نتایج این تحقیق نشان داده است که این مدل در شناسایی روابط پنهان میان موجودیت‌های ترافیک شهری کارآمد هستند.

عسگری و همکاران (۲۰۱۹) از فنون متنوع و ترکیبی برای استخراج و یکپارچه‌سازی دانش از منابع مختلف مانند ویکی‌پدیا، جداول وب و متون ساختاریافته استفاده کرده‌اند تا فارس‌بیس را که یک گراف دانش باز فارسی است را ایجاد کنند.

زکی و همکارانش (۲۰۱۹) به بررسی روش‌هایی پرداخته‌اند که پایگاه‌هایی اطلاعاتی زیست‌پزشکی را به هم پیوند داده و اثربخشی رویکردهای موجود را که می‌توانند پیوندهای ناشناخته بین موجودیت‌ها را در پایگاه‌های داده به‌منظور بهینه‌سازی گراف دانش پیش‌بینی کنند را ارزیابی کرده‌اند.

رحمان و تاکاسو (۲۰۱۷) دو مدل را ارائه کرده‌اند که روابط موجودیت‌ها برای تولید گراف دانش را پیشنهاد می‌کند. یانگ و همکاران^۱ (۲۰۲۳) و تانگ و همکاران^۲ (۲۰۱۹) به حوزه پاسخگویی به پرس‌وجوها بر اساس گراف دانش پرداخته‌اند.

یانگ و همکاران (۲۰۲۳) یک مدل جدید برای تجزیه سؤال را بکار گرفتند که از مکانیسم‌های توجه و رمزگذارهای دوگانه استفاده می‌کند تا بازنمایی سؤال را بهبود بخشد. نتایج آزمایشی روی مجموعه داده ComplexWebQuestions کارآمدی مدل پیشنهادی را نشان می‌دهد، شاخص F1 به امتیاز ۷۲ درصد دست می‌یابد و اثربخشی این رویکرد را در تجزیه سؤالات پیچیده به زیرمجموعه‌های ساده و بهبود دقت منطق سؤال را نشان می‌دهد.

تانگ و همکاران (۲۰۱۹) در مطالعه خود مدلی را ارائه می‌کنند که به سؤال‌ها بر اساس گراف دانش موضوعی پاسخ می‌دهد.

در مقوله طراحی هستی‌شناسی به‌صورت انسانی، ترابی و همکاران (۱۴۰۱) به طراحی و پیاده‌سازی هستی‌شناسی حوزه پزشکی هسته‌ای؛ نیارکی و همکاران (۱۳۹۷) به طراحی و توسعه هستی‌شناسی تجربیات مکانی در مسیریابی شهری؛ تقی‌زاده نائینی و همکاران (۱۳۹۸) به استخراج هستی‌شناسی هویت دیجیتال مبتنی بر تحلیل حوزه و ثروتی و همکاران (۱۳۹۶) به توسعه یک هستی‌شناسی جغرافیایی برای استفاده در کاربردهای نظامی پرداخته‌اند. همه این پژوهش‌ها بر اساس تجربه انسانی به مدل‌سازی یک هستی‌شناسی مبتنی بر دامنه پرداخته‌اند.

در حوزه بازیابی اطلاعات و موتورهای جست‌وجوی مبتنی بر پایه گراف دانش، همایونی و همکاران^۳ (۲۰۱۸)، لی (۲۰۱۹) و وانگ و همکاران^۴ (۲۰۱۹) پژوهش کرده‌اند.

همایونی و همکاران (۲۰۱۸) مطالعه گراف پزشکی^۵ به‌عنوان یک چارچوب بازیابی اطلاعات زیست‌پزشکی معنایی با استفاده از جاسازی گراف دانش برای پاب‌مد^۶ را انجام دادند. نتایج این مطالعه شواهدی را ارائه می‌کند مبنی بر اینکه

1. Yang et al.
2. Tong et al.
3. Homayouni
4. Wang et al.
5. MedGraph
6. PubMed

رویکردهای معنایی برای جست‌وجو و ارتباط در کتابخانه‌های دیجیتال زیست‌پزشکی که بر مدل‌سازی گراف دانش تکیه دارند، نتایج جست‌وجوی بهتری را در مقایسه با روش‌های سنتی از نظر معیارهای هدف ارائه می‌دهند.

لی (۲۰۱۹) در مقاله خود به بررسی تئوری ساخت و روش پیاده‌سازی این نوع موتورهای جست‌وجو در یک حوزه خاص می‌پردازد، به این امید که از طریق این موتور جست‌وجو بتوان پرس‌وجوی موجودیت را در یک حوزه خاص ارائه کرد، یعنی از طریق درک عمیق نیازهای کاربر، موجودیت‌های مربوطه را از پایگاه داده موجودیت بازیابی و به کاربران بازگرداند.

و وانگ و همکاران (۲۰۱۹) در پژوهش خود راه‌حل ساخت یک گراف دانش چندمنبعی در مقیاس بزرگ را از ابتدا برای شرکت سوگو^۱، شامل معماری، پیاده‌سازی فنی و کاربردهای آن به‌منظور پشتیبانی از جست‌وجو در وب و فراتر از آن معرفی می‌کند.

مطالعه شن و همکاران^۲ (۲۰۲۲) بر روی بهینه‌سازی پرس‌وجو بر مبنای گراف دانش بوده است. ایشان یک مولد پرسش پیشنهاد کرده‌اند که از یک گراف دانش بیرونی به‌عنوان راهنما استفاده می‌کند.

در مقوله بهینه‌سازی بازیابی اطلاعات با هستی‌شناسی، جعفری پاورسی و همکاران (۱۳۹۹) به بهینه‌سازی بازیابی معنایی اطلاعات با استفاده از روش‌های برچسب‌گذاری و هستی‌شناسی پرداخته و به این نتیجه رسیده‌اند که دو روش پردازش زبان طبیعی و هستی‌شناسی به ارتقای بازیابی معنایی اطلاعات منجر می‌شود.

علاوه بر این، اسیم و همکاران^۳ (۲۰۱۹) به مرور و ارزیابی روش‌های جدید بازیابی اطلاعات مبتنی بر هستی‌شناسی برای متون و چندرسانه‌ای‌ها و انواع داده‌های بین‌زبانی پرداخته است.

عالیشان کرمی و همکاران (۱۳۹۶) در پژوهشی به بررسی ۵ مقاله پرداخته‌اند که در خصوص کاربردها و روش ساخت هستی‌شناسی‌هایی هستند که در نظام‌های بازیابی اطلاعات زیست‌پزشکی مبتنی بر هستی‌شناسی به کار رفته‌اند. یکی از پژوهش‌های یافت شده به طراحی سیستم اطلاعاتی مبتنی بر هستی‌شناسی در حوزه کشاورزی پرداخته است.

لیائو و همکاران^۴ (۲۰۱۵) یک سامانه اطلاعاتی کشاورزی یکپارچه بر اساس هستی‌شناسی پیشنهاد کرده‌اند تا اطلاعات جامع و دقیقی برای راهنمایی کارآمد کشاورزان و متخصصان کشاورزی ارائه کند.

در مقوله غنی‌سازی اصطلاح‌نامه‌ها برای ساخت هستی‌شناسی، حسینی بهشتی و اژه‌ای (۱۳۹۴) و وو^۵ (۲۰۱۸) به ساخت هستی‌شناسی بر اساس اصطلاح‌نامه موضوعی پرداخته‌اند.

حسینی بهشتی و اژه‌ای (۱۳۹۴) به طراحی و پیاده‌سازی هستی‌شناسی علوم پایه بر اساس مفاهیم و روابط موجود در اصطلاح‌نامه‌های مرتبط تمرکز کرده است.

1. Sogou Inc

2. Shen et al.

3. Asim et al.

4. Liao et al.

5. Wu

وو (۲۰۱۸) روشی را برای غنی‌سازی اصطلاح‌نامه با ایجاد روابط متفاوت بین اصطلاح‌های مرتبط بخشی از اصطلاح‌نامه اریک^۱ ارائه کرده است. نتایج آزمایشی این مطالعه نشان داد که اصطلاح‌نامه غنی‌شده ابزار بهتری برای بازیابی اطلاعات است.

مرور ادبیات موجود در حوزه سازمان‌دهی دانش و اطلاعات بر پایه هستی‌شناسی و گراف‌های دانش نشان می‌دهد که بخش قابل توجهی از پژوهش‌های انجام‌شده در دهه‌های اخیر در چند محور عمده متمرکز بوده‌اند. نخست، پژوهش‌هایی که به استخراج خودکار یا ترکیبی اطلاعات ساختاریافته از متون پرداخته‌اند. دوم، مطالعاتی که به توسعه گراف‌های دانش در دامنه‌های خاص مانند زبان فارسی، زیست‌پزشکی، ترافیک شهری یا سامانه‌های جست‌وجو معطوف بوده‌اند. سوم، پژوهش‌هایی که بر طراحی هستی‌شناسی به‌صورت انسانی یا نیمه‌خودکار در حوزه‌های محدود علمی و تخصصی مانند پزشکی، جغرافیا، کشاورزی و مسیریابی تمرکز داشته‌اند. علاوه بر این، شماری از تحقیقات نیز به بهینه‌سازی بازیابی اطلاعات مبتنی بر هستی‌شناسی و غنی‌سازی اصطلاح‌نامه‌ها برای ساخت هستی‌شناسی‌های تخصصی پرداخته‌اند.

بررسی محتوایی این پژوهش‌ها نشان می‌دهد که هرچند همگی در راستای بهره‌گیری از رویکردهای معنا‌محور برای سازمان‌دهی و بازنمایی دانش حرکت کرده‌اند، اما دامنه داده‌ها و بستر کاربردی آن‌ها محدود و متفاوت از محیط‌های آرشیوی واقعی بوده است. به‌بیان‌دیگر، بیشتر این مطالعات بر روی متون عمومی یا تخصصی نسبتاً ساخت‌یافته مانند ویکی‌پدیا، جداول وب، بانک‌های اطلاعاتی علمی و متون فنی انجام شده‌اند؛ درحالی‌که محیط‌های آرشیوی، به‌ویژه در حوزه اسناد تاریخی، با چالش‌هایی از قبیل ناهمگنی منابع، تنوع قالب‌ها و ابهام معنایی موجودیت‌ها مواجه‌اند. افزون بر این، در اغلب پژوهش‌های پیشین، هدف اصلی ارائه یا ارزیابی روش‌های فنی برای استخراج یا مدل‌سازی هستی‌شناسی بوده و کمتر به کاربرد عملی این مدل‌ها در نظام‌های واقعی بازیابی اطلاعات و مدیریت دانش توجه شده است.

در چنین بستری، پژوهش حاضر جایگاهی متمایز و رویکردی کاربردی دارد. تفاوت بنیادین آن با مطالعات پیشین نه تنها در دامنه داده‌ها، بلکه در رویکرد ترکیبی، مسئله‌محور و عملیاتی آن نهفته است. در این پژوهش، طراحی هستی‌شناسی دامنه با بهره‌گیری از دو مسیر مکمل انجام می‌شود:

۱. تحلیل و استخراج خودکار و نیمه‌خودکار موجودیت‌ها و روابط از اسناد تاریخی آرشیوی؛

۲. بازاستفاده و یکپارچه‌سازی هستی‌شناسی‌ها و رده‌بندی‌های موجود در سطح وب معنایی و منابع ملی.

بدین ترتیب، این پژوهش در پی آن است که برای نخستین بار یک مدل هستی‌شناسی دامنه در حوزه تاریخ ایران را در بستر واقعی آرشیو دیجیتال طراحی و پیاده‌سازی کند. به‌طور کلی، این مطالعه ضمن بهره‌گیری از دستاوردهای نظری پیشین در حوزه وب معنایی، هستی‌شناسی و گراف دانش، خلأ موجود در ادبیات پژوهش داخلی در زمینه به‌کارگیری فناوری‌های معنا‌محور برای سازمان‌دهی و بازیابی معنایی اسناد تاریخی ایرانی را برطرف می‌سازد. دستاورد مورد انتظار آن، ایجاد پایه‌ای برای گراف دانش تاریخی ایران است که می‌تواند زمینه‌ساز توسعه سامانه‌های هوشمند پژوهشی، تصمیم‌یار و مدیریت دانش تاریخی کشور باشد.

روش

برای انجام این پژوهش از روش آمیخته استفاده شده است. برای تهیه هستی‌شناسی دامنه در حوزه تاریخ معاصر ایران، لازم به شناخت موجودیت‌های اطلاعاتی اسناد، طراحی ساختار هستی‌شناسی، استخراج اطلاعات بر اساس مدل هستی‌شناسی و استخراج روابط بین موجودیت‌ها و اعتبارسنجی داده‌های استخراج‌شده و روابط بین آن‌هاست. برای شناخت موجودیت‌های اطلاعاتی از شیوه تحلیل متن استفاده شده است. انواع موجودیت‌های اطلاعاتی استخراج‌شده مقوله‌بندی شده و با نظرسنجی از خبرگان به روش تکنیک گروه اسمی^۱ (متخصصان حوزه تاریخ معاصر و مدیران و کارشناسان مرکز اسناد انقلاب اسلامی) اعتبارسنجی شده‌اند. در طراحی هستی‌شناسی دامنه از اصول بهروش‌های^۲ طراحی هستی‌شناسی دامنه بر پایه هستی‌شناسی شکلی^۳ پایه ارائه‌شده توسط رابرت آرپ و همکارانش^۴ (۲۰۱۵) در مرکز ملی هستی‌شناسی زیست‌پزشکی^۵ الگو گرفته شده است که شامل مراحل زیر است:

۱. تعیین حدود موضوع هستی‌شناسی.
۲. جمع‌آوری اطلاعات: شناسایی اصطلاحات کلی^۶ استفاده شده در هستی‌شناسی‌های موجود و متون جامعه هدف؛ تجزیه و تحلیل برای حذف افزونگی‌ها.
۳. مرتب‌سازی این اصطلاحات در یک سلسله‌مراتب از اصطلاحات عام‌تر و خاص‌تر.
۴. نظم‌دهی به نتیجه برای اطمینان از: الف. انسجام منطقی، فلسفی و علمی، ب. انسجام و سازگاری با هستی‌شناسی‌های مشابه، ج. قابلیت فهم برای انسان، به‌ویژه از طریق تدوین تعاریف قابل‌خواندن برای انسان.
۵. رسمی‌سازی دست‌سازه^۷ بازنمایی سازمان‌یافته در یک زبان قابل استفاده توسط کامپیوتر به گونه‌ای که نتیجه بتواند در یک چارچوب محاسباتی پیاده‌سازی شود.

بر اساس بهروش‌های ارائه‌شده در طراحی هستی‌شناسی دامنه که توسط آرپ و همکارانش ارائه‌شده است، اولین گام در ساخت یک هستی‌شناسی دامنه، تعیین صریح دامنه موردنظر هستی‌شناسی است، یعنی پاسخ به این پرسش: «این هستی‌شناسی، هستی‌شناسی چه بخشی از واقعیت است؟» ارائه یک بیانیه صریح از این دامنه، نشان‌دهنده آن چیزی است که باید در هستی‌شناسی گنجانده شود و آنچه باید از آن حذف شود. بیانیه هستی‌شناسی مرکز اسناد انقلاب اسلامی ایران به شرح زیر است:

هستی‌شناسی مرکز اسناد انقلاب ایران محدود به تاریخ معاصر ایران شامل دوره‌های تاریخی پهلوی دوم، انقلاب اسلامی و پس از آن، دوره جمهوری اسلامی تا به امروز است. این هستی‌شناسی در رده‌های مربوط به اشخاص سیاسی، احزاب و گروه‌ها، دوره‌های حکومت، ساختار حکومت‌ها، خانواده‌ها و گروه‌های قومی، تشکلات، سازمان‌ها و نهادها، مکان‌ها و زیرساخت‌ها و رویدادهای مرتبط با این دوران دارای سطح دانه‌بندی تا جزئی‌ترین رده‌هاست. دانه‌بندی در هستی‌شناسی به

-
1. Nominal Group Technique
 2. Best Practices
 3. Basic Formal Ontology (BFO)
 4. Arp et al.
 5. National Center for Biomedical Ontology
 6. Universals
 7. Artefact

این معنی است که به‌عنوان مثال، زمانی که در مورد اشخاص صحبت می‌شود صرف شخص بودن، یا فقط سیاست‌مدار بودن یا هنرمند بودن کافی نیست و نوع مسئولیت سیاسی فرد به تناسب شغل وی از جمله رئیس‌جمهور، وزیر، سفیر، نماینده مجلس سنا و... مانند آن در هستی‌شناسی مورد توجه قرار گرفته است.

در گام دوم طراحی هستی‌شناسی دامنه، به‌منظور جمع‌آوری اطلاعات از منابع اسنادی هدف و شناسایی اصطلاحات دامنه مورد بررسی، متون تحلیل شدند و استخراج اصطلاح‌های کاندید به دو روش نیمه‌خودکار و خودکار برای استخراج اصطلاح‌ها با توجه به محدوده تعیین شده هستی‌شناسی انجام گرفت. جامعه پژوهش را تمامی منابع دیجیتال شده موجود در آرشیو دیجیتال مرکز اسناد انقلاب اسلامی تشکیل می‌دهد. جامعه اسنادی این مرکز شامل بیش از ۴/۵ میلیون برگ سند مکتوب، حدود ۳۱ هزار ساعت مصاحبه تاریخی شفاهی، نزدیک به ۳۷۲ هزار قطعه عکس، بیش از ۱۳ میلیون عنوان خبری و حدود ۹۰ هزار مقاله داخلی و خارجی است. بخش قابل توجهی از این منابع با استفاده از فناوری‌های تشخیص کاراکتر نوری و تبدیل گفتار به متن به فرمت دیجیتال تبدیل شده‌اند و از طریق سامانه آرشیو دیجیتال مرکز قابل دسترسی هستند. با توجه به گستردگی جامعه پژوهش، نمونه مورد بررسی در این مطالعه بر اساس تحلیل محتوایی و با مشورت مدیران و کارشناسان مرکز، به مجموعه‌ای از منابع منتخب محدود شد که به‌طور مستقیم با محورهای زیر مرتبط بودند:

- منابعی که در مورد اشخاص از جمله رجال سیاسی، رجال مذهبی، افراد نظامی، هنرمندان، مشاهیر و شخصیت‌های نامی علمی و ادبی، خانواده‌ها و اعضای دربار پهلوی دوم گردآوری یا تدوین شده‌اند؛
- منابعی که درباره خانواده‌های معروف و خاندان سلطنتی تهیه شده‌اند،
- منابع مرتبط با گروه‌های قومی و اقلیت‌های مذهبی
- منابع مرتبط با رویدادها و گاه‌شمارها
- منابعی که درباره احزاب، تشکلات و سازمان‌ها هستند،
- منابع مرتبط با اماکن، بناها، زیرساخت‌ها و سازه‌ها

سپس، برای شناخت و استخراج موجودیت‌های اطلاعاتی، از دو ابزار اصلی پژوهش استفاده شد:

نخست، نرم‌افزار متن‌باز اینسپشن^۱ به‌عنوان ابزار برچسب‌گذاری به‌منظور برچسب‌گذاری دستی و نیمه‌خودکار متون هدف به کار گرفته شد (شکل ۱). این نرم‌افزار امکان تعریف انواع موجودیت‌ها، روابط و کنترل کیفی داده‌های زبانی را فراهم می‌سازد.

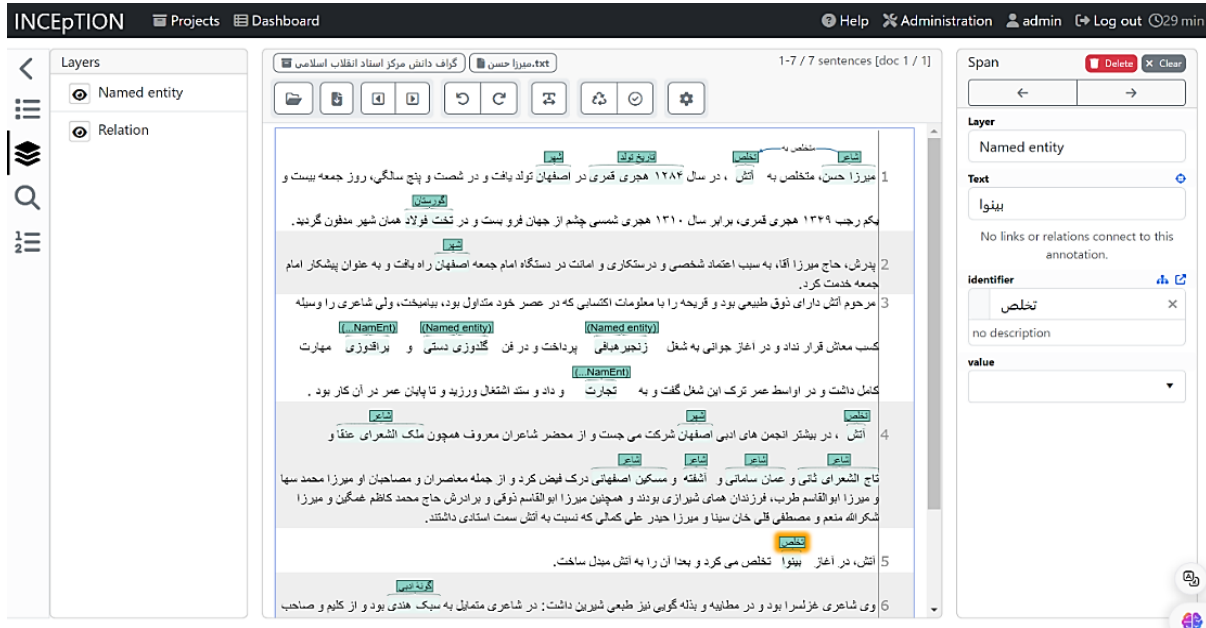
دوم، مدل زبانی بزرگ گروک^۲ نسخه ۳ به‌عنوان ابزار پردازش زبان طبیعی برای استخراج خودکار موجودیت‌ها بر پایه روش تشخیص موجودیت‌های اسمی^۳ مورد استفاده قرار گرفت (شکل ۲).

1. Inception: <https://inception-project.github.io/>

2. Grok

3. Named Entity Recognition

شکل ۱. برچسب‌گذاری با استفاده از نرم‌افزار متن باز اینسپشن



شکل ۲. برچسب‌گذاری متن با استفاده از مدل زبانی بزرگ گروک نسخه ۳

- شخص: آبی
- مکان: سبز
- سازمان: بنفش
- رویداد: نارنجی
- مفهوم: خاکستری

[شاپور بختیار]، متولد ۱۲۹۳، ادعا می‌کرد پدرش، یکی از خوانین [بختیاری]، در سال ۱۳۱۳ توسط [رضاخان] اعدام شده است. تحصیلات ابتدایی را در [شهر کرد] گذراند و در سال ۱۳۰۵ به [اصفهان] رفت، جایی که با محافل [فراماسونری] آشنا شد. در سال ۱۳۰۹ به [بیروت] رفت و دیپلم ریاضی از مدرسه فرانسوی‌ها گرفت. در سال ۱۳۱۳ به [ایران] بازگشت و سپس برای ادامه تحصیل به [فرانسه] رفت. در سال ۱۳۱۸ با یک دختر فرانسوی ازدواج کرد. در [جنگ جهانی دوم]، در ارتش [فرانسه] خدمت کرد. در سال ۱۳۲۴ به [ایران] بازگشت و در سال ۱۳۲۸ مدیرکل اداره کار [خوزستان] شد. در سال ۱۳۲۹ به [کمیته فدائیان شاه] پیوست و از [محمد مصدق] حمایت کرد. او به [حزب ایران]، یکی از احزاب زیرمجموعه [جبهه ملی ایران]، پیوست و به یکی از رهبران آن تبدیل شد. [شاپور بختیار] به جاسوسی برای [ساواک]، [انگلیس]، [فرانسه]، و ارتباط دوستانه با [سفارت آمریکا] شهرت داشت. او در روزهای پایانی [رژیم پهلوی]، [نخست‌وزیری ۱۳۵۷] را پذیرفت تا رژیم را نجات دهد، اما با هوشیاری [امام خمینی] و مردم، پس از ۳۷ روز از [ایران] گریخت. در [فرانسه] با حمایت کشورهای غربی و عربی، به‌ویژه [عراق]، علیه [پیروزی انقلاب اسلامی] فعالیت کرد و در ۱۵ مرداد ۱۳۷۰ در [پاریس] کشته شد.

در استخراج اطلاعات به شکل خودکار، که از مدل‌های زبانی بزرگ استفاده شده است خطاها با نظارت انسانی اصلاح شدند. به‌عنوان مثال در شکل ۲، «بختیاری» موجودیتی از نوع «گروه قومی» یا «ایل» است که مدل به‌عنوان مکان شناسایی کرده است.

علاوه بر تحلیل متون، به‌منظور شناسایی اصطلاح‌های کلی استفاده‌شده در هستی‌شناسی‌ها و رده‌بندی‌های موجود، از رویکردی ترکیبی شامل تحلیل محتوای دستی و به‌کارگیری ابزارهای نرم‌افزاری استفاده شد. در این مرحله، منابع زیر به‌صورت نظام‌مند موردبررسی قرار گرفته و اصطلاح‌های مرتبط استخراج شدند. برای تحلیل و تطبیق مفاهیم از نرم‌افزار

پروتزه^۱ و ابزارهای اسپارکیوال کوئری^۲ جهت مرور، و استخراج مفاهیم از فایل های آر.دی.اف / او.دبلیو.ال.^۳ بهره گرفته شد. منابع مورد استفاده عبارت اند از:

- هستی شناسی دی بی پدیا
- هستی شناسی گراف دانش فارسی^۴
- هستی شناسی بیفریم^۵ (مدل کتاب شناسی)
- رده بندی کتابخانه کنگره: DSR تاریخ ایران

در گام سوم اصطلاحات گردآوری شده با استفاده از رویکرد تحلیل محتوا و طبقه بندی مفهومی و بر پایه اصول طراحی هستی شناسه دامنه در یک ساختار سلسله مراتبی سازمان دهی شده و سپس هر اصطلاح به یک رده (کلی ها^۶) مرتبط شد. این مرحله بخشی از فرایند مدل سازی مفهومی در روش پژوهش ترکیبی حاضر محسوب می شود. برای قالب بندی اصطلاحات اصول زیر مورد استفاده قرار گرفتند (Arp et al., 2015):

۱. استفاده از اسامی مفرد
 ۲. استفاده از حروف کوچک ایتالیک برای اسامی عام.
 ۳. پرهیز از مخفف ها
 ۴. اختصاص شناسه الفبایی - عددی منحصر به فرد به هر اصطلاح.
 ۵. تضمین یک معنایی اصطلاحات.
 ۶. تضمین یک معنایی عبارت های رابطه ای
 ۷. پرهیز از اسامی غیر قابل شمارش^۷
- پس از قالب بندی، برای ایجاد ساختار سلسله مراتبی از تحلیل تطبیقی هستی شناسی ها و رده بندی های موجود بر اساس اصول مربوط به طبقه بندی ها در هستی شناسی استفاده شد (Arp et al., 2015):

۱. ساختار بندی هر هستی شناسی حول یک سلسله مراتب «هست-یک»^۸ به عنوان ستون فقرات
 ۲. تضمین کامل بودن سلسله مراتب «هست-یک»
 ۳. تضمین توارث منفرد (تک والدی) بیان شده
 ۴. رعایت فرض دنیای باز^۹
 ۵. پایبندی به قاعده عینیت.
- بر اساس اصول مربوط به اصطلاح شناسی، تعاریف و طبقه بندی ها ساختار اولیه سلسله مراتبی هستی شناسی دامنه بر اساس مراحل زیر شکل گرفت:

1. Protégé
2. SPARQL Query
3. RDF/OWL
4. <http://farsbase.net/>
5. Bibframe
6. Universals
7. Mass nouns
8. is_a
9. open-world assumption

۱- برای نظم‌دهی به رده‌ها از عام به خاص با توجه به محدوده تعیین شده هستی‌شناسی، ساختار سلسله‌مراتبی هستی‌شناسی دی‌بی‌پدیا به‌عنوان مبنا قرار گرفت و از سایر رده‌بندی‌ها و هستی‌شناسی‌ها رده‌های مرتبط در جایگاه پیشنهادی اولیه قرار گرفتند. این پیشنهاد بر اساس مقایسه بین ساختار سلسله‌مراتب سایر هستی‌شناسی‌ها و رده‌بندی‌ها و همچنین پیشنهادی مدل‌های زبانی بزرگ مانند گروک، دیپ‌سیک^۱ و کیوون^۲ انجام گرفته است. سپس نمونه‌های استخراج‌شده از تحلیل متون بر اساس شباهت‌ها و ویژگی‌های مشترک خوشه‌بندی شده و به رده مرتبط نگاشت شدند.

۲- در مورد موجودیت‌هایی که دارای رده‌ای در دی‌بی‌پدیا یا رده‌بندی تاریخ ایران نبودند رده‌ای با مشورت با خبرگان و به کمک مدل‌های زبانی بزرگ پیشنهاد شد. برای مثال برای موجودیت «**دادگاه ویژه روحانیت**» و مانند آن، رده «**نهاد انقلاب اسلامی**» ذیل رده «**سازمان**» ایجاد شد.

در نهایت به‌منظور اعتبارسنجی رده‌های استخراج‌شده از روش گروه اسمی^۳ استفاده شد. در این گروه ۱۱ نفر از مدیران و خبرگان (شامل اعضای هیئت علمی) مرکز اسناد انقلاب اسلامی شرکت کردند که بر اساس معیارهایی چون سابقه فعالیت پژوهشی و مدیریتی در حوزه اسناد تاریخی، تسلط بر نظام‌های رده‌بندی و سازمان‌دهی، تسلط به مفاهیم تاریخ معاصر ایران و تجربه در تحلیل محتوای اسناد آرشیوی انتخاب شدند.

نشست این گروه در دو جلسه سه‌ساعته انجام شد. با توجه به تعداد بالای رده‌ها، این روش به‌عنوان جایگزینی سریع و ساختارمند برای روش دلفی به کار گرفته شد؛ روشی که به‌جای تکرارهای زمان‌بر دلفی، در یک یا دو جلسه ۲ تا ۳ ساعته با حضور ۸ تا ۱۲ نفر از خبرگان و تمرکز بر اجماع و اولویت‌بندی نظرات انجام می‌شود. معیارهای اصلی اعتبارسنجی شامل (۱) منطقی سلسله‌مراتبی، (۲) صحت تاریخی/فرهنگی، (۳) هم‌ترازی و معادل‌سازی مفاهیم، و (۴) کامل بودن و سادگی ساختار در نظر گرفته شدند.

یافته‌ها

سؤال اول پژوهش: موجودیت‌ها و رده‌های حوزه دامنه موضوعی اسناد آرشیو دیجیتال مرکز اسناد انقلاب اسلامی کدام‌اند و چگونه می‌توان آنها را در قالب هستی‌شناسی دامنه‌بازنمایی کرد؟

پاسخ: بر اساس تحلیل متون آرشیو دیجیتال مرکز اسناد انقلاب اسلامی و رده‌های هستی‌شناسی دی‌بی‌پدیا، رده‌های هستی‌شناسی ببفریم و رده‌بندی تاریخ ایران بر اساس کتابخانه‌کنگره آمریکا، یافته‌های این بخش شامل ۵۳۵ اصطلاح بر اساس محدوده موضوعی تعیین شده هستی‌شناسی بود که بر اساس رده‌های اصلی زیر تنظیم شدند (جدول ۱):

1. Deepseek
2. Qwen
3. Nominal Group Technique (NGT)

جدول ۱. تعداد رده‌های اصلی و زیررده‌ها هستی‌شناسی دامنه مرکز اسناد انقلاب اسلامی ایران

رده	زیر رده	زیر رده	تعداد زیر رده
			۵۳۴
			۲۵۳
	عامل	شخص	۱۷۱
		سازمان	۷۵
		خانواده	۶
	رویداد		۳۷
چیز	رسانه		۰
	اثر		۱۸
	دوره حکومت		۵۲
	مفهوم موضوعی		۱
	مکان		۳۷
	سازه معماری		۶۸
	گروه قومی		۶۷

سؤال دوم پژوهش: ساختار سلسله‌مراتبی و روابط میان رده‌های هستی‌شناسی دامنه چگونه باید باشد؟

پاسخ: پیش‌نویس ساختار سلسله‌مراتبی اصطلاح‌های استخراج‌شده در یافته‌های بخش اول، بر پایه رده‌های اصلی و ساختار هستی‌شناسی دی‌بی‌پدیا شکل گرفت. پس از اعتبارسنجی توسط گروه خبرگان و اجماع اکثریت، در ساختار رده‌بندی ۴۵۷ رده تأیید شد، ۷۰ رده اضافه شد، جایگاه ۵۳ رده در سلسله‌مراتب تغییر کرد، ۱۹ رده حذف شد، اصطلاح ۵ رده تغییر کرد و ۲ رده هم در جایگاه سلسله‌مراتب تغییر کرده و هم اصطلاحشان عوض شد (جدول ۲).

جدول ۲. نتیجه اعتبارسنجی رده‌های هستی‌شناسی دامنه مرکز اسناد انقلاب اسلامی

تعداد	نتیجه اعتبارسنجی
۴۵۷	تأیید رده
۷۰	اضافه کردن رده
۵۳	تغییر رده در سلسله‌مراتب
۱۹	حذف رده
۵	تغییر اصطلاح
۲	تغییر اصطلاح / تغییر در سلسله‌مراتب

عمده رده‌های افزوده‌شده مربوط به رده‌های اصلی عامل ← شخص و مکان هستند. برخی از رده‌ها نیز به‌عنوان مفهوم موضوعی در نظر گرفته شدند. خبرگان مرکز اسناد افزودن مفاهیمی همچون جریان‌ها مانند جریان‌های سیاسی (نوگرایی، روشنفکری و...)، اقتصادی، اجتماعی و همچنین مکتب‌های فکری و ایدئولوژی‌ها، فعالیت‌های سیاسی و فعالان سیاسی را ضروری دانستند که ذیل مفهوم موضوعی پوشش داده می‌شود. از آنجا که موضوعات و رده‌بندی علم بسیار وسیع است بنابراین در این هستی‌شناسی یک شاخه اصلی به‌عنوان مفهوم موضوعی قرار داده شده است تا مفاهیم موضوعی بر اساس نظام‌های موضوعی مختلف ذیل آن قرار گیرند که در پوشش این مطالعه نمی‌گنجد. سازه معماری نیز به این بخش منتقل

شد و رده‌های ذیل آن به عنوان تقسیم‌بندی‌های مکانی در نظر گرفته شدند. جدول ۳ نتیجه اجماع خبرگان این حوزه برای افزودن رده‌های ضروری به هستی‌شناسی است.

جدول ۳. فهرست رده‌های افزوده‌شده با تعیین جایگاه در سلسله‌مراتب

رده‌های افزوده‌شده	مکان در سلسله‌مراتب
ولیعهد، رئیس شورای شهر، عضو شورای شهر	چیز ← عامل ← شخص ← سیاستمدار
عضو حوزه علمیه (استاد یا مدرس)، فرد شبه‌نظامی، عارف	چیز ← عامل ← شخص
مداح، تعزیه‌خوان، قاری قرآن، مؤذن	چیز ← عامل ← شخص ← هنرمند ← خواننده ← خواننده آئینی
خواننده سنتی، خواننده پاپ، خواننده فولکلور (محلی)، خواننده رپ	چیز ← عامل ← شخص ← هنرمند ← خواننده
راهب بودایی	چیز ← عامل ← شخص ← پیشوای مذهبی
فرد نظامی نیروی زمینی، فرد نظامی نیروی هوایی، فرد نظامی نیروی دریایی، مأمور نیروی انتظامی	چیز ← عامل ← شخص ← فرد نظامی (نیروی مسلح)
مقام امنیتی، کارشناس امنیتی	چیز ← عامل ← شخص ← فرد امنیتی
شرکت	چیز ← عامل ← سازمان
شرکت دولتی، شرکت عمومی غیردولتی	چیز ← عامل ← سازمان ← شرکت
نافرمانی مدنی	چیز ← رویداد ← رویداد اجتماعی
منطقه	چیز ← مکان ← کشور ← سکونتگاه (مکان مسکونی)
جلگه، خور، آبشار، رود فصلی (مسیل)، تالاب، باتلاق، بیابان، کلوت، دشت، چشمه، قله کوه	چیز ← مکان ← کشور ← مکان طبیعی
مکان انسان ساخته	چیز ← مکان ← کشور
سالن موسیقی (کنسرت)، کاباره	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان فرهنگی و هنری
مکان مذهبی، مکان تاریخی، مکان نظامی، مکان تجاری / اقتصادی، مکان تفریحی / گردشگری، مکان صنعتی، سکوی نفتی، سکوی گازی، مکان کشاورزی (زراعی)، مکان فرهنگی و هنری، مکان خدماتی / عمومی	چیز ← مکان ← کشور ← مکان انسان ساخته
عبادتگاه (معبد)، زیارتگاه، حسینیه (تکیه)	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان مذهبی
امامزاده	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان مذهبی ← زیارتگاه
بیمارستان، زایشگاه، تیمارستان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان خدماتی / عمومی
بازار، فروشگاه / مغازه	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تجاری / اقتصادی
مزرعه، شالیزار، نخلستان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان کشاورزی (زراعی)
آرامگاه (بقعه، مرقد، مدفن، مزار یا مقبره)	چیز ← مکان ← کشور ← مکان انسان ساخته ← آرامستان (قبرستان)
ایستگاه درون‌شهری، ایستگاه برون‌شهری	چیز ← مکان ← کشور ← مکان انسان ساخته ← ایستگاه
پایانه مسافربری، قنات (کاهریز)، چاه آب، کانال مصنوعی	چیز ← مکان ← کشور ← مکان انسان ساخته

علاوه بر رده‌های افزوده‌شده، برخی رده‌ها در ساختار رده‌بندی جابه‌جا شدند و در جایگاه جدید قرار گرفتند. جدول ۴ فهرست این رده‌ها و جایگاه جدید آنها در سلسله‌مراتب هستی‌شناسی را نشان می‌دهد.

جدول ۴. فهرست رده‌های جابجا شده در سلسله‌مراتب رده‌بندی

رده تغییر یافته / مسیر سلسله‌مراتب قبلی	مسیر سلسله‌مراتب (محل قرارگیری جدید)
شهردار / چیز ← عامل ← شخص	چیز ← عامل ← شخص ← سیاستمدار ← زیررده (سیاستمدار)
رهبر ارکستر / چیز ← عامل ← شخص	چیز ← عامل ← شخص ← هنرمند ← زیررده (هنرمند)
امام‌جماعت / چیز ← عامل ← شخص ← پیشوای مذهبی	چیز ← عامل ← شخص ← پیشوای مذهبی ← روحانی ← روحانی مسلمان ← زیررده (روحانی مسلمان)
مقام ساواک، مقام واجا / چیز ← عامل ← شخص ← مأمور امنیتی	چیز ← عامل ← شخص ← فرد امنیتی ← زیررده (مقام امنیتی)
بازجو / چیز ← عامل ← شخص ← مأمور امنیتی	چیز ← عامل ← شخص ← فرد امنیتی ← زیررده (کارشناس امنیتی)
روزنامه‌نگار / چیز ← عامل ← شخص	چیز ← عامل ← شخص ← خبرنگار ← زیررده (خبرنگار)
شرکت خصوصی / چیز ← عامل ← سازمان	چیز ← عامل ← سازمان ← شرکت ← زیررده (شرکت)
اتلاف / چیز ← عامل ← سازمان	چیز ← عامل ← سازمان ← گروه ← زیررده (گروه)
میدان، خیابان، شهرک، برج / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← سکونتگاه (مکان مسکونی) ← منطقه ← استان ← شهرستان ← بخش ← شهر ← زیررده (شهر)
دهستان / چیز ← مکان ← کشور ← سکونتگاه (مکان مسکونی) ← منطقه ← استان ← شهرستان ← بخش ← شهر	چیز ← مکان ← کشور ← سکونتگاه (مکان مسکونی) ← منطقه ← استان ← شهرستان ← بخش ← شهر ← زیررده (بخش)
چشمه آب گرم / چیز ← مکان ← کشور ← مکان طبیعی	چیز ← مکان ← کشور ← مکان طبیعی ← چشمه ← زیررده (چشمه)
سالن سینما، سالن تئاتر / چیز ← سازه معماری ← ساختمان ← محل برگزاری	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان فرهنگی و هنری ← زیررده (مکان فرهنگی و هنری)
موزه، کتابخانه / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان فرهنگی و هنری ← زیررده (مکان فرهنگی و هنری)
کلیسا، صومعه، مسجد، آتشکده، کنیسه / چیز ← سازه معماری ← ساختمان مذهبی	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان مذهبی ← عبادتگاه (معبد) ← زیررده (عبادتگاه)
زیارتگاه / چیز ← سازه معماری ← ساختمان مذهبی	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان مذهبی ← زیررده (مکان مذهبی)
قلعه / چیز ← سازه معماری ← زیرساخت ← سازه نظامی	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تاریخی ← زیررده (مکان تاریخی)
ورزشگاه / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← زیررده (مکان انسان ساخته)
زندان / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← زیررده (مکان انسان ساخته)
معدن / چیز ← مکان ← کشور ← مکان طبیعی	چیز ← مکان ← کشور ← مکان انسان ساخته ← زیررده (مکان انسان ساخته)
آرامستان (قبرستان) / چیز ← مکان ← کشور	چیز ← مکان ← کشور ← مکان انسان ساخته ← زیررده (مکان انسان ساخته)

رده تغییر یافته / مسیر سلسله‌مراتب قبلی	مسیر سلسله‌مراتب (محل قرارگیری جدید)
فرودگاه، سد، دیوار دریایی، سکوی پرتاب، بندر، پست برق، آسیاب، بنای یادبود، ایستگاه / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← زیررده (مکان انسان ساخته)
رستوران، مرکز خرید (مال) / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تجاری / اقتصادی ← زیررده (مکان تجاری / اقتصادی)
کازینو / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تفریحی / گردشگری ← زیررده (مکان تفریحی / گردشگری)
پارک (بوستان) / چیز ← مکان ← کشور	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تفریحی / گردشگری ← زیررده (مکان تفریحی / گردشگری)
هتل / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تفریحی / گردشگری ← زیررده (مکان تفریحی / گردشگری)
باغ‌وحش / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان تفریحی / گردشگری ← زیررده (مکان تفریحی / گردشگری)
کارخانه / چیز ← سازه معماری ← ساختمان	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان صنعتی ← زیررده (مکان صنعتی)
نیروگاه / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان صنعتی ← زیررده (مکان صنعتی)
باغ / چیز ← مکان ← کشور	چیز ← مکان ← کشور ← مکان انسان ساخته ← مکان کشاورزی (زراعی) ← زیررده (مکان کشاورزی (زراعی))
ایستگاه مترو، ایستگاه تراموا، ایستگاه مسیر / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← ایستگاه → ایستگاه درون‌شهری ← زیررده (ایستگاه درون‌شهری)
ایستگاه راه‌آهن / چیز ← سازه معماری ← زیرساخت	چیز ← مکان ← کشور ← مکان انسان ساخته ← ایستگاه → ایستگاه برون‌شهری ← زیررده (ایستگاه برون‌شهری)
اقلیت مذهبی / چیز ← گروه قومی	چیز ← زیررده (چیز)

طی فرایند اعتبارسنجی، برخی رده‌ها حذف شده و اصطلاح برخی رده‌ها نیز تغییر کرد. در بخش رده‌های اشخاص، «دانشجو، مأمور ساواک، افسر پلیس و مأمور واجا» حذف شدند. همچنین اصطلاح «شطرنج‌باز» به جای «بازیکن شطرنج» اصلاح شد تا اصطلاح دقیق‌تر و متعارف‌تری داشته باشد.

در حوزه سازمان‌ها و نهادها، اصلاحات اصطلاحی قابل توجهی انجام شد: «مؤسسه آموزشی» به «نهاد آموزشی» تغییر یافت و به زیررده دانشگاه منتقل شد، «سازمان قانون‌گذار» به «نهاد قانون‌گذار»، «سازمان مشورتی» به «نهاد مشورتی» و «مؤسسه رسانه‌ای» به «سازمان رسانه‌ای» اصلاح شد. برخی رده‌ها نیز حذف شدند، از جمله «گروهک سیاسی و فرقه ضاله» حذف شده و «رسانه» به دلیل همپوشانی با «سازمان رسانه‌ای» نیز از هستی‌شناسی خارج شد.

در بخش مکان‌ها و سازه‌ها، سازه معماری که شامل دو زیررده «ساختمان و زیرساخت» بود حذف شدند و رده‌های مرتبط با آنها شامل «ساختمان مذهبی، معبد، ساختمان تاریخی، آسمان‌خراش، محل برگزاری، سازه نظامی و قلعه نظامی» حذف شدند و معادل آن‌ها به عنوان مکان ذیل رده مکان اضافه شد. همچنین برخی ویرایش‌های اصطلاحی انجام شد، مانند تغییر «مسیر حمل‌ونقل» به «مسیر انتقال» و «موتور بادی» به «توربین بادی».

به‌طور کلی، این بازنگری و اعتبارسنجی باعث افزودن رده‌های لازم، پاک‌سازی رده‌های غیرضروری، اصلاح اصطلاحات غیرمعارف و بهبود جایگاه سلسله‌مراتب رده‌ها شد و هم‌زمان انسجام، منطق و دقت هستی‌شناسی را در سطوح مختلف افزایش داد.

سؤال سوم پژوهش: چه هستی‌شناسی‌ها و رده‌بندی‌هایی برای بازاستفاده در تهیه هستی‌شناسی دامنه مرکز اسناد انقلاب اسلامی مطلوب هستند؟

پاسخ: بر اساس جست‌وجوی موضوعی در داخل کشور هستی‌شناسی‌ای برای تاریخ معاصر ایران یا مرتبط با آن یافت نشد. ولی گسترش رده‌بندی تاریخ ایران بر اساس نظام کتابخانه کنگره آمریکا و رده‌بندی تاریخ ایران بر اساس نظام ده‌دهی دیویی موردبررسی تطبیقی قرار گرفتند. با توجه به اینکه گسترش رده‌بندی کنگره کلیه رده‌های دیویی را پوشش داده است، لذا کنار گذاشته شد. علاوه بر این هستی‌شناسی‌های عمومی داخلی و خارجی شامل هستی‌شناسی‌های گراف دانش فارسی، دی‌بی‌پدیا، ویکی‌دیتا^۱ و یاگو^۲ که حوزه‌هایی از دامنه مورد مطالعه را پوشش داده بودند بررسی شدند. بر اساس مطالعه عسگری و همکاران (۲۰۱۹)، دی‌بی‌پدیا به‌عنوان هستی‌شناسی هسته در وب معنایی شناخته می‌شود که سایر هستی‌شناسی‌های عمومی را پوشش می‌دهد. هستی‌شناسی گراف دانش فارسی نیز تنها در ۱۰ رده با هستی‌شناسی دی‌بی‌پدیا تفاوت دارد که قابل چشم‌پوشی بود. در حوزه رده‌های مربوط به «اثر» و زیررده‌های آن نیز از هستی‌شناسی بیفریم به‌عنوان مرجع قابل‌استناد در این حوزه استفاده شد.

بحث و نتیجه‌گیری

آنچه در این مطالعه به دست آمد، یک رده‌بندی هستی‌شناسی مبنا برای حوزه تاریخ معاصر ایران بر اساس اسناد و منابع موجود در آرشیو دیجیتال مرکز اسناد انقلاب اسلامی است. این رده‌بندی بر اساس چارچوب هستی‌شناسی دامنه و رعایت اصول آن استخراج شده و قابل توسعه است. بر اساس این رده‌بندی، موجودیت‌های قابل استخراج از منابع آرشیو دیجیتال مرکز اسناد انقلاب اسلامی در قالب رده‌های اصلی شخص، سازمان، خانواده، دوره حکومت، رویداد، مکان، اثر، اقلیت مذهبی، گروه قومی و زیررده‌های آنها (۵۳۵ اصطلاح) مشخص شدند. این نتیجه با پژوهش‌های امامی (۱۴۰۱) و مرادی و همکاران (۱۳۹۴) هم‌راستا است که هر دو استخراج خودکار یا نیمه‌خودکار روابط و مفاهیم از متون فارسی را آزموده‌اند. باین حال، تمایز مهم پژوهش حاضر در این است که داده‌های آن از متون واقعی آرشیوی و تاریخی استخراج شده است، درحالی که پژوهش‌های یادشده عمدتاً بر داده‌های ساخت یافته یا متون عمومی تکیه داشته‌اند. به‌علاوه، گوئیرز و همکاران^۳ (۲۰۱۶) و ویمالاسوریا و دوو^۴ (۲۰۱۰) نیز بر استخراج ترکیبی موجودیت‌ها با استفاده از مدل‌های هستی‌شناختی تأکید کرده‌اند که با رویکرد ترکیبی این تحقیق هم‌خوانی دارد. باین حال، پژوهش حاضر به‌جای تمرکز بر الگوریتم‌های استخراج، بر ساختار مفهومی بومی و تناسب آن با دامنه تاریخی ایران تمرکز کرده است.

1. Wikidata.com
2. <https://yago-knowledge.org>
3. Gutierrez et al.
4. Wimalasuriya & Dou

نتیجه اعتبارسنجی خبرگان منجر به اصلاح و بازنگری در ساختار نهایی هستی‌شناسی شامل تأیید ۴۵۷ رده، افزودن ۷۰ رده، حذف ۱۹ رده، تغییر ۵ اصطلاح، جابه‌جایی ۵۳ رده در سلسله‌مراتب و تغییر اصطلاح و جابجایی هم‌زمان ۲ رده شد. این فرایند نشان داد که ساختارهای عمومی مانند دی‌بی‌پدیا یا گراف دانش فارسی برای پوشش دقیق مفاهیم بومی تاریخ ایران کفایت ندارند و نیاز به بومی‌سازی گسترده دارند. این یافته با نتایج مینایی و سجادی (۱۳۹۸) و عسگری و همکاران (۲۰۱۹) هم‌راستا است؛ زیرا آن‌ها نیز در توسعه گراف دانش فارسی بر ضرورت اصلاح و تکمیل ساختارهای عمومی تأکید کرده‌اند. درعین‌حال، با یافته‌های تان و همکاران (۲۰۲۱) که گراف دانش حوزه ترافیک را بر مبنای داده‌های ساختاریافته بنا کردند، تفاوت دارد؛ چراکه در پژوهش حاضر داده‌ها از اسناد تاریخی نیمه‌ساخت‌یافته استخراج شدند و ماهیت پویاتر و نامتجانس‌تری داشتند. افزوده شدن رده‌های مرتبط با جریان‌های فکری، سیاسی و مذهبی نیز بیانگر لزوم درک مفهومی و تاریخی داده‌ها است، موضوعی که در پژوهش‌های پیشین کمتر به آن توجه شده است.

مقایسه این پژوهش با پژوهش‌های مرتبط در حوزه طراحی هستی‌شناسی انسانی و کاربردی نشان می‌دهد که نتایج آن با یافته‌های ترابی و همکاران (۱۴۰۱)، نیارکی و همکاران (۱۳۹۷) و تقی‌زاده نائینی و همکاران (۱۳۹۸) مشابهت دارد، از آن‌جهت که همه آن‌ها طراحی هستی‌شناسی را مبتنی بر تحلیل دامنه و مشارکت خبرگان دانسته‌اند؛ اما تمایز پژوهش حاضر در آن است که این طراحی صرفاً مفهومی نبوده، بلکه با بازخورد عملی خبرگان آرشیوی و در بستر واقعی مرکز اسناد انجام گرفته است؛ درنتیجه، خروجی آن قابلیت به کارگیری مستقیم در نظام آرشیو دیجیتال را دارد.

یافته‌های این پژوهش در زمینه غنی‌سازی ساختار مفهومی و معنایی داده‌های تاریخی، با نتایج جعفری پاورسی و همکاران (۱۳۹۹) و وو (۲۰۱۸) هم‌راستا است که بر نقش هستی‌شناسی در بهبود بازیابی معنایی اطلاعات تأکید دارند. باین‌حال، تفاوت اصلی در نوع داده و کاربرد نهایی است؛ پژوهش حاضر بر اسناد تاریخی و روابط مفهومی میان شخصیت‌ها، رویدادها و مکان‌ها متمرکز بوده، درحالی‌که تحقیقات پیشین بیشتر در حوزه‌های پزشکی، آموزشی یا چندرسانه‌ای انجام شده‌اند.

نتایج پژوهش حاضر ضمن تأیید کارآمدی رویکرد ترکیبی در طراحی هستی‌شناسی، بر اهمیت بومی‌سازی مدل‌های مفهومی در آرشیوهای تاریخی تأکید دارد. هم‌راستایی نتایج با مطالعات پیشین در حوزه‌های استخراج خودکار داده‌ها، توسعه گراف دانش، پاسخگویی معنایی به پرس‌وجوها و طراحی سیستم‌های مبتنی بر هستی‌شناسی نشان می‌دهد که استفاده از گراف دانش می‌تواند بستر مؤثری برای سازمان‌دهی معنایی، بازیابی هوشمند و تحلیل محتوای تاریخی فراهم کند. باین‌حال، یافته‌های بومی این پژوهش شکاف میان الگوهای جهانی و نیازهای محتوایی آرشیوهای ایرانی را برجسته می‌سازد و راه را برای توسعه مدل‌های دانش‌محور بومی هموار می‌کند.

برای ادامه زنجیره این پژوهش، پیشنهاد می‌شود حوزه‌های زیر برای مطالعات آتی موردتوجه قرار گیرند:

۱. توسعه هستی‌شناسی‌ها دامنه در حوزه‌های موضوعی دیگر با رعایت اصول تعامل‌پذیری میان هستی‌شناسی‌ها
۲. پیاده‌سازی گراف دانش عملیاتی و ارزیابی کارایی آن در بازیابی اطلاعات
۳. توسعه روش‌های خودکارتر برای استخراج مفاهیم و روابط با استفاده از مدل‌های زبانی بزرگ
۴. مطالعه الگوریتم‌های وزن‌دهی روابط در گراف دانش

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

نویسندگان اصول اخلاقی را در انجام و انتشار این پژوهش علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آن‌هاست.

مشارکت نویسندگان

مقاله حاضر مستخرج از رساله دکتری علیرضا انتهای سرای است. نادر نقشینه (استاد راهنما): طراحی پژوهش، نظارت بر مراحل انجام پژوهش، بررسی و کنترل نتایج، اصلاح، بازبینی و نهایی‌سازی مقاله؛ علیرضا انتهای سرای: طراحی ایده، گردآوری داده‌ها، تحلیل متن و استخراج موجودیت‌ها، مدل‌سازی هستی‌شناسی، تحلیل یافته‌ها و تهیه پیش‌نویس مقاله؛ بهروز مینایی بیدگلی (استاد راهنما) و علی شعبانی (استاد مشاور): مشارکت در طراحی پژوهش، نظارت بر پژوهش، مطالعه و بازبینی مقاله. کلیه نویسندگان نسخه منتشرشده مقاله را مطالعه و با آن موافقت کرده‌اند.

دسترسی به داده‌ها

داده‌ها در صورت درخواست از نویسندگان در دسترس خواهند بود.

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرایند نگارش

در طول آماده‌سازی این اثر، نویسنده(گان) از مدل‌های زبانی بزرگ گروک، دیپ سیک و کوون به‌منظور استخراج موجودیت‌ها از متن و نظم‌دهی به رده‌های سلسله‌مراتب هستی‌شناسی استفاده کردند. پس از استفاده از این ابزار/خدمت، نویسنده(گان) مطالب را در صورت نیاز بررسی و ویرایش کرده و مسئولیت کامل محتوای نشریه را بر عهده می‌گیرند.

تعارض منافع

بنا بر اظهار نویسندگان این مقاله تعارض منافع ندارد.

حامی مالی

این پژوهش هیچ کمک مالی مشخصی از سازمان‌های تأمین مالی در بخش‌های دولتی، تجاری یا غیرانتفاعی دریافت نکرده است.

سپاسگزاری

مراتب سپاس و قدردانی خود را از رئیس مرکز اسناد انقلاب اسلامی و معاون مرکز اسناد که با حمایت از این پژوهش و فراهم کردن داده‌ها و تأمین زیرساخت‌های مورد نیاز امکان تحقق آن را فراهم کردند، به عمل می‌آوریم. همچنین از گروه خبرگان که وقت و دانش ارزشمند خود را در اختیار این مطالعه قرار دادند سپاسگزاریم.

ORCID

Nader Naghshineh



<https://orcid.org/0000-0002-0509-5454>

Alireza Entehaee Saray



<https://orcid.org/0000-0002-7996-0287>

Behrouz Minaei-Bidgoli



<https://orcid.org/0000-0002-9327-7345>

Ali Shabani



<https://orcid.org/0000-0001-7323-4332>

منابع

- امامی، حجت. (۱۴۰۱). ارائه روشی برای استخراج اطلاعات ساختاریافته محدود به دامنه از صفحات وب فارسی. *پردازش علائم و داده‌ها*، ۵۲، ۱۳۳-۱۴۶. <http://jsdp.rcisp.ac.ir/article-1-1102-fa.html>
- ترابی، لاله، میرحسینی، زهره، اباذری، زهرا، و حسینی بهشتی، ملوک‌السادات. (۱۴۰۱). طراحی و پیاده‌سازی هستان‌شناسی پزشکی هسته‌ای. *مدیریت اطلاعات سلامت*، ۱۸ (۲)، ۸۸-۸۱. <https://doi.org/10.22122/him.v18i2.4357>
- تقی‌زاده نایینی، جواد، فهیم‌نیا، فاطمه، و نقشینه، نادر. (۱۳۹۸). استخراج آنتولوژی هویت دیجیتال مبتنی بر تحلیل حوزه. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۴ (۴)، ۱۶۶۹-۱۷۰۰. <https://doi.org/10.35050/JIPM010.2019.018>
- ثروتی، لیلا، ولوی، محمدرضا، و حورعلی، مریم. (۱۳۹۶). توسعه یک هستان‌شناسی جغرافیایی برای استفاده در کاربردهای نظامی. *اطلاعات جغرافیایی (سپهر)*، ۲۶ (۱۰۲)، ۵-۱۷. <https://www.magiran.com/paper/1738409>
- جعفری پاورسی، حمیده، حریری، نجلا، علیپورحافظی، مهدی، باب‌الحوائجی، فهیمه، و خادمی، مریم. (۱۳۹۹). ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی. *مطالعات کتابداری و سازمان‌دهی اطلاعات*، ۳۱ (۱)، ۱۸-۳۸. <https://www.magiran.com/paper/2127226>
- حسینی بهشتی، ملوک‌السادات، و اژه‌ای، فاطمه. (۱۳۹۴). طراحی و پیاده‌سازی هستی‌شناسی علوم پایه بر اساس مفاهیم و روابط موجود در اصطلاح‌نامه‌های مرتبط. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۰ (۳)، ۶۷۷-۶۹۶. https://jipm.irandoc.ac.ir/article_699309.html
- سجادی، سید محمدباقر، و مینایی بیدگلی، بهروز. (۱۳۹۸). معماری سامانه گراف دانش زبان فارسی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۵ (۲)، ۴۲۵-۴۶۲. <https://doi.org/10.35050/JIPM010.2020.057>
- صادقی نیارکی، ابوالقاسم، برزگر، مریم، و شاکری، مریم. (۱۳۹۷). طراحی و توسعه آنتولوژی تجربیات مکانی در مسیریابی شهری. *رایانش نرم و فناوری اطلاعات*، ۷ (۲)، ۱-۱۵. https://jscit.nit.ac.ir/article_80956.html
- عالیشان کرمی، نادر، حاجی‌زین‌العابدینی، محسن، رداد، ایرج، و قاضی میرسعید، سید جواد. (۱۳۹۶). کاربرد و نقش هستان‌شناسی در نظام‌های بازیابی اطلاعات زیست‌پزشکی. *انفورماتیک سلامت و زیست‌پزشکی*، ۴ (۴)، ۳۲۷-۳۴۰. <http://jhbmi.ir/article-1-222-fa.html>
- مرادی، مهدی، وزیرنژاد، بهرام، و بحرانی، محمد. (۱۳۹۴). ساخت هستان‌شناسی دانش عرفی زبان فارسی با رویکردی تلفیقی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۱ (۱)، ۱۰۹-۱۲۴. https://jipm.irandoc.ac.ir/article_699334.html

References

- Alishan Karami, N., Haji Zeinolabedini, M., Radad, I., & Ghazi-Mirsaeid, S. J. (2017). Application and role of ontologies in biomedical information retrieval systems. *Journal of Health and Biomedical Informatics*, 4(4), 327-340. <http://jhbmi.ir/article-1-222-fa.html> [In Persian]
- Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with basic formal ontology*. MIT Press. <http://www.jstor.org/stable/j.ctt17kk7vw>
- Asgari-Bidhendi, M., Hadian, A., & Minaei-Bidgoli, B. (2019). FarsBase: The Persian knowledge graph. *Semantic Web*, 10(6), 1169-1196. <https://doi.org/10.3233/SW-190369>
- Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The use of ontology in retrieval: A study on textual, multilingual, and multimedia retrieval. *IEEE Access*, 7, 21662-21686. <https://doi.org/10.1109/ACCESS.2019.2897849>
- Choi, J. (2022). Graph embedding-based domain-specific knowledge graph expansion using research literature summary. *Sustainability*, 14(19), Article 12299. <https://doi.org/10.3390/su141912299>

- Emami, H. (2022). A method for extracting domain-specific structured information from Persian web pages. *Signal and Data Processing*, 52, 133–146. <http://jsdp.rcisp.ac.ir/article-1-1102-fa.html> [In Persian]
- Gutierrez, F., Dou, D., Fickas, S., Wimalasuriya, D., & Zong, H. (2016). A hybrid ontology-based information extraction system. *Journal of Information Science*, 42(6), 798–820. <https://doi.org/10.1177/0165551515610989>
- Homayouni, R., Ding, D., Roy, S., & Ebeid, I. A. (2018). MedGraph: A semantic biomedical information retrieval framework using knowledge graph embedding for PubMed. *Frontiers in Big Data*, 5, Article 965619. <https://doi.org/10.3389/fdata.2022.965619>
- Hosseini Beheshti, M. S., & Ejei, F. (2015). Designing and implementing basic sciences ontology based on concepts and relationships of relevant thesauri. *Iranian Journal of Information Processing and Management*, 30(3), 677–696. https://jipm.irandoc.ac.ir/article_699309.html [In Persian]
- Jafari Pavarsi, H., Hariri, N., Alipour Hafezi, M., Babolhavaeji, F., & Khademi, M. (2020). Enhancing semantic information retrieval using tagging and ontology. *Librarianship and Information Organization Studies*, 31(1), 18–38. <https://www.magiran.com/paper/2127226> [In Persian]
- Kejriwal, M., Knoblock, C. A., & Szekely, P. (2021). *Knowledge graphs: Fundamentals, techniques, and applications*. MIT Press.
- Li, J. (2019). Research on search engine technology of knowledge graph analysis. *AIP Conference Proceedings*, 2073(1), Article 020094. <https://doi.org/10.1063/1.5090759>
- Liao, J., Li, L., & Liu, X. (2015). An integrated, ontology-based agricultural information system. *Information Development*, 31(2), 150–163. <https://doi.org/10.1177/0266666913510716>
- Moradi, M., Vazirnezhad, B., & Bahrani, M. (2015). Commonsense knowledge extraction for Persian language: A combinatory approach. *Iranian Journal of Information Processing and Management*, 31(1), 109–124. https://jipm.irandoc.ac.ir/article_699334.html [In Persian]
- Park, J.-R., & Howarth, L. C. (2013). *New directions in information organization*. Emerald Group Publishing.
- Rahman, M. M., & Takasu, A. (2017). Entity oriented action recommendations for actionable knowledge graph generation. In *Proceedings of the International Conference on Web Intelligence* (pp. 686–693). ACM. <https://doi.org/10.1145/3106426.3106546>
- Rajabi, E., & Etminani, K. (2022). Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*. Advance online publication. <https://doi.org/10.1177/01655515221112844>
- Sadeghi Niaraki, A., Barzegar, M., & Shakery, M. (2018). Design and development of ontology for spatial experiences in urban routing. *Journal of Soft Computing and Information Technology*, 7(2), 1–15. https://jscit.nit.ac.ir/article_80956.html [In Persian]
- Sajjadi, S. M. B., & Minaei Bidgoli, B. (2019). Architecture of Persian language knowledge graph system. *Iranian Journal of Information Processing and Management*, 35(2), 425–462. <https://doi.org/10.35050/JIPM010.2020.057> [In Persian]
- Servati, L., Valavi, M. R., & Hourali, M. (2017). Development of a geographic ontology for use in military applications. *Geographical Information (Sepehr)*, 26(102), 5–17. <https://www.magiran.com/paper/1738409> [In Persian]
- Shen, X., Chen, J., Chen, J., Zeng, C., & Xiao, Y. (2022). Diversified query generation guided by knowledge graph. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining* (pp. 897–907). ACM. <https://doi.org/10.1145/3488560.3498431>
- Taghizadeh Nayini, J., Fahimnia, F., & Naghshineh, N. (2019). Ontology extraction of digital identity based on domain analysis. *Iranian Journal of Information Processing and Management*, 34(4), 1669–1700. <https://doi.org/10.35050/JIPM010.2019.018> [In Persian]
- Tan, J., Qiu, Q., Guo, W., & Li, T. (2021). Research on the construction of a knowledge graph and knowledge reasoning model in the field of urban traffic. *Sustainability*, 13(6), Article 3191. <https://doi.org/10.3390/su13063191>
- Tong, P., Zhang, Q., & Yao, J. (2019). Leveraging domain context for question answering over knowledge graph. *Data Science and Engineering*, 4(4), 323–335. <https://doi.org/10.1007/s41019-019-00109-w>

- Torabi, L., Mirhosseini, Z., Abazari, Z., & Hosseini-Beheshti, M. S. (2022). Design and construction of nuclear medicine ontology. *Health Information Management*, 18(2), 81–88. <https://doi.org/10.22122/him.v18i2.4357> [In Persian]
- Wang, P., Jiang, H., Xu, J., & Zhang, Q. (2019). Knowledge graph construction and applications for web search and beyond. *Data Intelligence*, 1(4), 333–349. https://doi.org/10.1162/dint_a_00019
- Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3), 306–323. <https://doi.org/10.1177/0165551509360123>
- Wu, Y. (2018). Enriching a thesaurus as a better question-answering tool and information retrieval aid. *Journal of Information Science*, 44(4), 512–525. <https://doi.org/10.1177/0165551517706219>
- Yang, L., Guo, H., Dai, Y., & Chen, W. (2023). A method for complex question-answering over knowledge graph. *Applied Sciences*, 13(8), Article 5055. <https://doi.org/10.3390/app13085055>
- Zaki, N., Tennakoon, C., Al Ashwal, H., Al Jaber, A., & Al Ameri, A. (2019). Methods of creating knowledge graph by linking biological databases. In *Advances in Intelligent Systems and Computing* (Vol. 803, pp. 73–83). Springer. https://doi.org/10.1007/978-3-319-98702-6_7