ScienceON Knowledge Graph System: Exploring New Frontiers in Science and Technology Information Integration System

Chanuk Lim [a,b], Nam-Gyu Kang [b], Suhyeon Yoo [b], Hyun Ji Jeong [c,*] and Sungsu Lim [b,**]

[a] *Department of Computer Science and Engineering, Chungnam National University, Daejeon 34134, South Korea*
E-mails: chanuklim@gmail.com, sungsu@cnu.ac.kr
[b] *Korea Institute of Science and Technology Information, Daejeon 34141, South Korea*
E-mails: chanuklim@kisti.re.kr, ngkang@kisti.re.kr, yoosu@kisti.re.kr
[c] *Department of Artificial Intelligence, Kongju National University, Cheonan 31080, South Korea*
E-mail: hjjeong@kongju.ac.kr

Abstract. The increasing complexity and volume of scientific and technological data necessitate advanced tools for effective data-driven analysis. Knowledge Graphs, with their capacity to encapsulate complex relationships among interconnected entities, have emerged as pivotal structures for organizing this vast amount of information. They enable a deeper understanding and exploration of data across various domains, notably in science and technology where the rapid proliferation of research outputs presents both opportunities and challenges. This paper presents the ScienceON Knowledge Graph System, a comprehensive framework designed to address integral challenges of integrating and analyzing scientific and technological data. We have developed the comprehensive infrastructure of ScienceON, a data ecosystem that harmonizes a wide spectrum of scientific and technological information. This encompasses everything from national R&D projects, scholarly papers, and patents to reports, author profiles, organizational details, keywords, and thematic categories. Our approach significantly advances the field not only by streamlining the aggregation of data via an Extract, Transform, Load process but also by facilitating the creation of a sophisticated knowledge graph. This knowledge graph meticulously interlinks research data, incorporating extensive metadata to accurately reflect the complex web of relationships within the science and technology domains. Our contributions are threefold: Firstly, we detail the creation of the ScienceON data ecosystem, highlighting an automated pipeline that ensures ongoing updates and expansion of data. Secondly, we describe the design of the ScienceON Knowledge Graph, which provides a detailed and interconnected representation of scientific and technological data. Lastly, we explore the application of the ScienceON Knowledge Graph in conducting graph-related experiments and in developing user-centric applications, demonstrating its versatility and utility. By employing rigorous data curation practices and utilizing the Resource Description Framework for data representation, we ensure the high quality and accessibility of our dataset, positioning the ScienceON Knowledge Graph as a gold standard in the realm of science and technology knowledge management. This initiative not only augments data management practices but also fosters the development of innovative applications and services, enhancing access to and understanding of the vast landscape of science and technology.

Keywords: Knowledge graph, Ontology, Science and technology information, Data ecosystem

*Corresponding author. E-mail: hjjeong@kongju.ac.kr.
**Corresponding author. E-mail: sungsu@cnu.ac.kr.

1. Introduction

Knowledge graphs (KGs) play a pivotal role in structuring vast amounts of interconnected data, enabling the organization and accessibility of information across various scenarios. These advanced data structures enhance the intelligent capabilities of numerous services by providing a foundation of structured, factual information [1, 2]. For instance, Google's and Microsoft's Knowledge Graphs enhance search capabilities through advanced question-answering features, utilizing a broad spectrum of global knowledge. Similarly, Facebook's vast social graph incorporates personal and global interests, while eBay's Product Knowledge Graph aims to semantically link products with their broader context [3]. These instances highlight how knowledge graphs are key in enriching user interactions and understanding complex relationships.

Particularly in the science and technology domain, the publication of scientific papers and patents has rapidly increased. Accordingly, various scholarly KGs such as Microsoft Academic Graph (MAG) [4], Aminer [5], Open Academic Graph (OAG) [6], OpenCitations [7], PubMed Knowledge Graph [8] have been proposed. Scholarly KGs have been used in various fields such as article retrieval, recommendation, and mining. For instance, methods for recommending scientific papers have been proposed [9–11], methods for document retrieval have been proposed [12–14], and community detection based approaches have been proposed to discover hidden relationships among researchers in an academic network [15–17]. Similarly, KGs built from patent data are used for keyword-based patent searches and for recommending patents to meet industry needs [18–20].

It is evident that information on the complex relationships between science, papers, patents, reports, and R&D projects is needed to analyze an entire process of how individual research achievements permeate industry and ultimately promote social development. These documents act as perfect channels for showcasing knowledge generated within the science and technology fields. Hence, a joint analysis of such documents across both domains offers significantly deeper insights than individual analysis. There have been numerous prior studies that have explored the relationship between the fields of science and technology from various perspectives, integrating these fields to investigate complex interrelations [21–27].

Previous research has highlighted two main challenges that KG systems face in accurately representing the science and technology domain: **(i) The Unexplored Potential of Knowledge Graphs in Existing Systems:** The National Technical Information Service (NTIS)[1] allows for searching U.S. government-funded technical reports but doesn't support KG creation. Similarly, the NIH ExPORTER[2] reveals healthcare-related research data without forming a KG, and the European Research Information Ontology (EURIO)[3] offers limited metadata for EU research documents despite constructing a KG. These systems often fail to provide comprehensive coverage and integration of heterogeneous data types, such as R&D projects and literature, limiting the systematic organization, access, and in-depth analysis of data. This deficiency impairs the investigation of complex data relationships, limiting the insights and discoveries that could be facilitated by a more structured and relational approach to data management and analysis. **(ii) The Limited Entity & Relation Types in Scientific and Technological Knowledge Graphs:** Table 1 provides a comparison of existing KGs within this domain. Despite considerable efforts to model literature, projects, authors, and other components as entities and to establish connections among them, the range of relationships used to link these entities remains limited. The simplicity of KGs is often a limitation in capturing the complexity and nuance of real-world relationships and data. This inherent simplicity can result in oversimplification of complex ideas and relationships, hindering deeper understanding and analysis of the knowledge domain represented.

We propose a system, called ScienceON Knowledge Graph System, designed to construct a comprehensive and systematic KG to address the challenges of data-driven analysis within the science and technology domains. This paper has two primary goals: **(i)** Firstly, to introduce ScienceON, a data ecosystem that includes R&D projects, papers, patents, and research reports from multiple sources. Notably, the data fueling this system is supported by the government and curated by the Korea Institute of Science and Technology Information (KISTI), offering a wealth of knowledge spanning all science and technology fields. ScienceON also features an automated pipeline for the seamless update and expansion of science and technology data. **(ii)** Secondly, to detail how the ScienceON Knowledge Graph leverages this infrastructure by connecting research project information with their outcomes and incorporating comprehensive metadata for each (such as titles, abstracts, and references). This methodology enables the creation of the ScienceON Knowledge Graph adept at mapping the intricate web of real-world relationships that

Table 1
Comparison of knowledge graphs in the science and technology domain.

Graph	Entity Type	Relationship Type
Cora[4][28]	Paper(P)	P-P
CiteSeer[5][28]	Paper(P)	P-P
cit-Patents Graph (SNAP)[6]	Patent(P)	P-P
Unified Database Management Systems(UDBMS)[7]	Patent(P), Inventor(I), Assignee(A), Class(C), Category(T)	P-P, P-A, P-C, P-I, C-T
PubMed KG[8]	Article(A), Author(Au), Affiliation(Af), Funding(F), Project(P)	A-A, A-P, A-Au, Au-Af, Au-F, F-P
Open Academic Graph(OAG)[8]	Paper(P), Author(A), Affiliation(Af), Venue(V)	P-P, P-A, P-Af, A-Af, P-V
AIDA[29]	Paper(P), Author(A), Affiliation(Af), industrialSector(S), DBpediaCategory(C), Topic(T), Patent(Pt)	P-A, A-Af, P-Af, P-S, P-C, Af-S, Af-C, P-T, Pt-T, Pt-S, Pt-C
ScienceON KG	Top-Project(T), Sub-Project(S), Author(A), Organization(O), Paper(P), Journal(J), PaperCategory(PC), Patent(Pt), IPC(I), Report(R), ReportCategory(RC), Keyword(K)	T-S, S-P, S-R, S-Pt, P-J, P-P, P-A, P-PC R-P, R-R, R-RC, R-A, Pt-Pt, Pt-I, Pt-A A-O, P-K, R-K

span the science and technology domains. The rigorously vetted data is highly accurate and publicly accessible, establishing the ScienceON Knowledge Graph as the gold standard in the realm of science and technology.

To summarize, our key contributions are as follows:

- **Development of a Data Ecosystem:** We have developed ScienceON, a comprehensive data ecosystem that aggregates science and technology-related data using an Extract, Transform, Load (ETL) process from various legacy systems. This system features an automated pipeline designed for efficiently collating information on papers, patents, reports, national R&D projects, and authors, and introduces a structured data management system.
- **Construction of the ScienceON Knowledge Graph:** Leveraging collected science and technology data, we have created the ScienceON Knowledge Graph. This KG integrates a wide spectrum of entities, attributes, and relationships relevant to a variety of science and technology disciplines, and is encoded using the Resource Description Framework (RDF), a standard endorsed by the World Wide Web Consortium (W3C).
- **Utilization of the ScienceON Knowledge Graph:** We harness the potential of the ScienceON Knowledge Graph by engaging in graph-level experiments and employing knowledge representation learning models for both comprehensive quantitative and qualitative analysis. Furthermore, by leveraging the collected data and the ScienceON Knowledge Graph, we develop user-centric applications specifically tailored to science and technology information. This initiative is distinguished by the introduction of an intuitive interface, which significantly enhances the users' ability to navigate and interact with the ScienceON Knowledge Graph, thereby democratizing access to complex data for a wider audience.

The remainder of this paper is structured as follows: Section 2 presents the motivation and background of our study. Section 3 reviews previous work related to KGs. In Section 4, we detail the ScienceON Knowledge Graph System, encompassing the collection and management of science and technology data, as well as the development of the ScienceON Knowledge Graph, which structures this data into a KG format. Section 5 discusses the results of graph-related tasks utilizing the ScienceON Knowledge Graph. Section 6 presents various applications leveraging the ScienceON Knowledge Graph. Finally, Section 7 concludes the paper.

[1]https://ntrl.ntis.gov/NTRL/
[2]https://report.nih.gov/exporter-data-dictionary
[3]https://cordis.europa.eu/
[4]https://relational.fit.cvut.cz/dataset/CORA
[5]https://relational.fit.cvut.cz/dataset/citeseer
[6]https://snap.stanford.edu/data/cit-Patents.html
[7]https://www.helsinki.fi/en/researchgroups/unified-database-management-systems-udbms/datasets/patent-dataset
[8]https://www.microsoft.com/en-us/research/project/open-academic-graph/

2. Motivation and Background

As the volume of science and technology information expands, the need to search and analyze this data intensifies, accompanied by a growing diversity of infrastructures to access such knowledge. Data-driven analysis requires multiple stages, including data discovery, acquisition, preparation, knowledge discovery and sharing [30]. Nonetheless, current datasets and infrastructures dedicated to science and technology information frequently fail to fully support this comprehensive data-driven analysis. We outline the challenges faced by existing systems in the science and technology domain from the standpoint of data-driven analysis, detailing the motivation and background for tackling these issues.

From the perspective of data discovery, the scattered distribution of data across disparate systems and repositories poses significant challenges. KISTI, under Article 40, CHAPTER V, of the Enforcement Decree of the Framework Act on Science and Technology, plays a crucial role in the comprehensive collection of domestic and international science and technology knowledge, supported by government funding. This enables the systematic construction of a wide array of science and technology information and continuous efforts in data discovery. The scientific and technological data utilized in this study are collected under national policies, forming the basis of our research.

Regarding data acquisition, KISTI collaborates extensively with both domestic and international institutions. It collects academic information through partnerships with various academic institutions and patents information in cooperation with the Korea Institute of Patent Information (KIPI). ScienceON serves as a data ecosystem for integrating data collected by KISTI, employing an ETL process to continuously aggregate large volumes of data from internal and external sources.

In the realm of data preparation, the science and technology data amassed necessitates a meticulous preparation process for subsequent data analysis. At KISTI, the metadata collected from academic and R&D projects undergoes a manual verification process. In addition, when service users submit feedback regarding any data issues, system operators proceed to review the data. Most importantly, as these datasets are public data, they must adhere to government database standardization guidelines to ensure standardization across databases and are validated by external independent audits. This verification process enables the data constructed by KISTI to serve as a gold standard in the field of science and technology.

The data, collected from various fields and refined and verified, enables knowledge discovery in ways that existing systems in the science and technology domain could not. By employing ontology, we clearly identify complex semantic relationships between data items, deriving knowledge not explicitly expressed and fostering a deeper understanding and discovery of new knowledge. In addition, leveraging graph representation techniques in the knowledge graph enhances our ability to conduct graph-related tasks, including node classification and link prediction, thereby discovering new significance.

Aiming for open science, ScienceON emphasizes data sharing. KISTI holds the rights to replicate, distribute, and transmit the collected scientific and technological information, offering free copyright usage and providing the information via REST APIs. Serving a diverse group of stakeholders, including students, researchers, and policymakers, ScienceON endeavors to supply scientific and technological information to various stakeholders and engage in technological cooperation.

In summary, traditional systems and research institutions have faced policy and technical challenges in collecting and analyzing scientific and technological data from a data-driven analysis perspective. This paper sets out the background and motivation for addressing these long-standing issues, spanning from improved data discovery mechanisms to enhanced data sharing capabilities. The proposed system represents a new initiative in the field of science and technology information to address and bridge these critical gaps.

3. Related Work

3.1. Knowledge graphs in science and technology domain

In 2012, Google announced a strategy to enhance the search engine, which emphasized utilization of KGs[9]. Moreover, research on construction and refinement of KGs in a scientific domain has been active, and is currently being applied to many real world applications. In a medical domain, KGs have been constructed from multiple sources, notably during the COVID-19 pandemic, enabling development of treatments to combat the virus [31]. These KGs have played crucial roles in supporting clinical decision-making and detecting inaccurate health information [32, 33]. In a education domain, KGs were built using pedagogical data to learn assessment data or to find relationships between learning topics and basic concepts, helping personalized learning [34]. These KGs were used to help students understand mathematical concepts [35]. In the field of Information and Communication Technology (ICT), the KGs were built for preventing or detecting various types of cyber-attacks [36] or were used to efficiently manage telecommunication networks [37, 38]. In an academic domain, scholarly communication accumulates data generated from papers and other scholarly writings, and it is building scholarly infrastructures to provide this data. Academic search engines provide scholarly literature search services using these infrastructures, and they define entity and relation from metadata such as authors, venues, and affiliations and use them to build scholarly KG.

Microsoft has built the Microsoft Academic Graph (MAG) [4], a Web-scale academic graph for semantic search engines. MAG consists of heterogeneous entities such as papers, patents, authors, affiliations, fields of study and venues. Researchers have used MAG to augment an academic graph or combine it with other infrastructures to create new scholarly KGs. M. Färber [39] proposed a large RDF dataset of over 8 billion triples based on MAG and linked to data sources such as DBpedia, Wikidata, OpenCitations, and the Global Research Identifier Database. Aminer [5] integrated researcher profiles and publications by collecting publications from online digital libraries such as DBLP bibliography, ACM Digital library, and CiteSeer and automatically collecting researcher profiles from the web. This was stored in MySQL and KGs were configured with a probability-based model to provide search services such as expertise search and association search. Open Academic Graph (OAG) [6] integrated MAG and Aminer with a LinKG framework to build a large-scale linked entity graph that integrates MAG and Aminer.

In addition, researchers have built scholarly KGs from various data sources. J. Xu *et al.* [8] built KGs by integrating multi-source information obtained from infrastructure such as PubMed literature, ORCID, MapAffil, and NIH ExPORTER project data. PubMed knowledge graph was defined to connect entities from multiple sources by integrating bio-entities extracted from 29 million PubMed abstracts by applying BioBERT based a Named Entity Recognition (NER) model, unique authors obtained from ORCID and MapAffil, and funding data obtained from NIH ExPORTER. B. Yaman *et al.* [40] built a SciGraph of around 1.5 to 2 billion triples containing research entities such as funders, research projects, affiliations and publications from Springer Nature. They also enhanced KGs by linking entities from DBpedia infrastructure including structured information from Wikipedia by a crowd-sourced community with entities generated by applying Named Entity Recognition(NER) to document abstracts, which are unstructured data in SciGraph.

Some not-for-profit organizations utilize scholarly KGs to provide infrastructure. OpenCitations [7] provides open scholarly information built using semantic web technologies. OpenCitations offers bibliographic metadata to researchers and authors, bibliometricians, librarians, funders, academic administrators of research institutes and universities, research managers, data repositories, publishers, and computer scientists and software providers to build large-scale scholarly KGs. OpenAIRE [41] is an EU e-Infrastructure dedicated to Open Science, an infrastructure that hosts not only research publications, but also scientific products generated by experiments, such as research data, research software, and experiments. OpenAIRE Graph builds KGs by utilizing richer objects than existing scholarly graphs such as publications, research data, research software, organizations, funding, etc.

Analyzing documents in the science and technology domains together can yield much deeper insights than evaluating them individually. Previous studies have explored these areas collectively, adopting various perspectives. Anderson examined the insights and viewpoints on the knowledge regarding academy-industry relationships and

[9] https://blog.google/products/search/introducing-knowledge-graph-things-not/

their influence on higher education [21]. Grimpe *et al.* analyzed that the complementary relationships between formal and informal university knowledge and technology transfer [22]. Huang *et al.* investigated the contemporary trend of collaboration between academia and industry in the field of fuel cells by analyzing papers and patents [23]. Larivière *et al.* revealed a global shift in basic research from industrial and governmental sectors to universities [24]. Bikard *et al.* found that academic scientists collaborating with industry partners generate more subsequent publications and fewer patents [25]. Salatino *et al.* presented ResearchFlow, a method that combines semantic technologies with machine learning to quantify the trends of research topics within both academic and industrial spheres [26]. McManus *et al.* assessed the output and influence of scientific publications originating from non-academic and industrial entities [27].

3.2. Knowledge Representation

Knowledge representation is the systematic structuring and organization of information. The goal of knowledge representation is to convey real-world details in a way that enables computers to comprehend and utilize this knowledge for solving complex problems, while encapsulating the richness of human knowledge within the limitations of computer memory, processing speed, and comprehension. Knowledge representation learning adeptly navigates the semantics of entities and relationships within a lower-dimensional space, effectively addressing the curse of dimensionality challenge. This enhancement significantly improves the performance of tasks such as knowledge acquisition and reasoning.

Many developments in learning knowledge graph representations depend on translating relations in the embedding space, transforming KG representation learning into vector computations, and evaluating the plausibility of facts based on the distance between two entities after translation by the relation. TransE [42] is an approach that interprets relationships by modeling them as translations applied to the low-dimensional embeddings of entities. TransR [43] addresses the challenge of a single space inadequacy for both entities and relations by introducing distinct spaces for entities and relations. TransH [44] characterizes a relation by utilizing a hyperplane in conjunction with a translation operation. KG2E [45] incorporates Gaussian distribution to address the uncertainties associated with entities and relations, drawing inspiration from Gaussian word embeddings. ManifoldE [46] expands embedding into a manifold-based approach to overcome the limitations of overly strict geometric forms imposed by certain methods using subspace projection. The traditional knowledge representation learning method simplifies design and training by focusing on semantic associations within triplets, but the accuracy of learned entities and relationships has limitations.

Graph Neural Networks(GNNs) have emerged for learning graph representations with the success of deep learning in various domains. They have broken the performance barrier of the existing knowledge representation learning models. GNNs utilize a propagation module composed of aggregation and update to apply the deep learning framework to graphs. GNNs learn representations of nodes, edges, and graphs by iteratively integrating and updating feature vectors from neighbors. The goal of GNNs is to learn to convert a graph into a low-dimensional vector in order to solve downstream tasks such as node-level tasks, edge-level tasks, and graph-level tasks [47]. The node-level task includes classification, regression and clustering of nodes. The edge-level task includes link prediction and edge classification. Finally, graph-level tasks aim to learn representations at the level of entire graphs or subgraphs, for like classification or regression.

With the architecture of Convolutional Neural Networks(CNNs), there have been efforts to generalize CNNs to graphs based on graph signal processing. GNNs with convolution operators are categorized into two main streams; spectral-based methods and spatial-based methods [48]. First, spectral-based GNNs were developed, where the convolution operators are defined in the Fourier domain. Spectral GCNs [49] were the first spectral-based approaches to receive attention, defining the filter as a learnable diagonal matrix. However, Spectral GCNs suffered from problems caused by using a diagonal matrix: as the graph changes, the eigenvectors also change, so the eigendecomposition is computationally inefficient. ChebNet [50] improved computational efficiency by defining filters as Chebyshev polynomials instead of simple diagonal matrices. GCN [51] simplified ChebNet's filters to a first-order approximation to mitigate overfitting and applied a renormalization trick to solve a problem of exploding and vanishing gradients. However, spectral-based GNNs required an entire graph and could not apply learned filters to graphs with different structures, so they were applied only under transductive conditions.

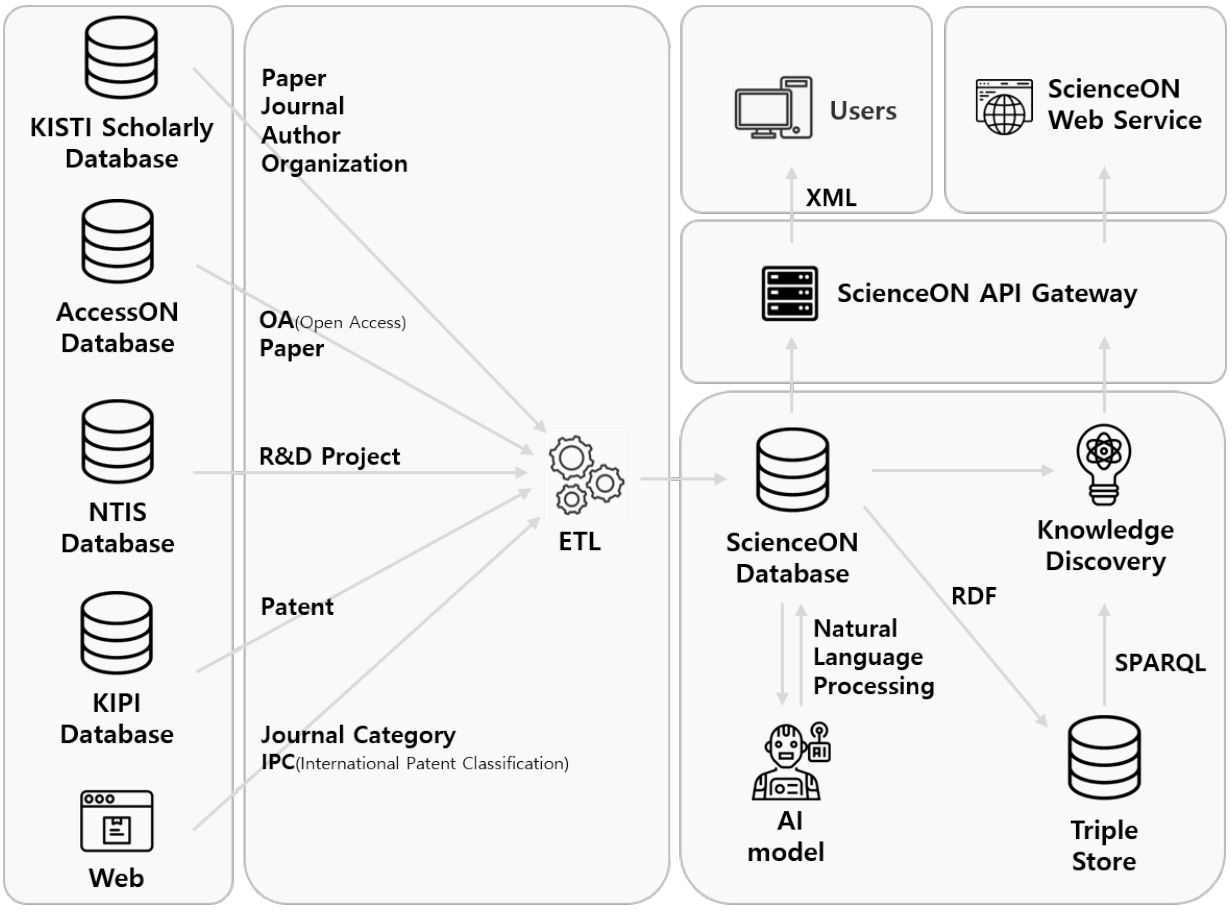

Fig. 1. The ScienceON Knowledge Graph System's architecture, depicting the integration and processing of science and technology information. From multiple data sources like KISTI Scholarly Database, AccessON, NTIS, KIPI and Web, to the utilization of an ETL pipeline for data aggregation, and culminating in knowledge discovery via RDF, SPARQL, and AI models, this system facilitates the structured dissemination and in-depth analysis of scientific data.

On the other hand, spatial-based GNNs solved the problem of spectral-based methods by defining convolution operators to aggregate features of neighboring nodes in the graph domain. GraphSAGE [52] was a framework for inductive node embedding, which used multiple aggregator functions to aggregate feature information from local neighborhood of a sampled node instead of using a full neighbor set. Also, with the success of the attention mechanism and transformer [53], several models employed the attention mechanism. GAT [54] applied self-attention into an aggregator function, allowing neighborhoods to be assigned different weights. GaAN [55] applied GRU to a multi-head attention mechanism to reflect importance of the attention head.

However, these GNNs are methods for homogeneous graphs, which cannot reflect different types of nodes and edges in the real world. Therefore, a GNN framework for heterogeneous graphs has emerged to model complex relationships between entities in the real world by defining node and edge types. HAN [56] was inspired by GAT to reflect the importance between nodes and meta-path based neighbors. To overcome the drawback of requiring domain knowledge to design meta-paths, HGT [57] designed heterogeneous mutual attention, heterogeneous message passing and target-specific aggregation to learn node- and edge-type dependent representations.

4. ScienceON Knowledge Graph System

In this section, we present the ScienceON Knowledge Graph System, a comprehensive data ecosystem designed to integrate science and technology data from a variety of legacy systems. The ScienceON Knowledge Graph System encompasses essential components for data-driven analysis such as data discovery, acquisition, preparation, knowledge discovery and sharing. As depicted in Figure 1, the ScienceON Knowledge Graph System architecture facilitates the progression from diverse data sources to the extraction of actionable knowledge.

4.1. Data Discovery

ScienceON has the capability to collect a wide array of scientific and technological data, including papers, journals, authors, organizations, patents, research reports, and R&D projects. It aggregates extensive science and technology data from four legacy systems and one web source. The status of the collected data is as presented in Table 2. This comprehensive approach allows ScienceON to serve as a pivotal resource, enabling researchers and practitioners to access a rich repository of scientific and technological information, thereby facilitating advanced research and development activities across various domains.

Table 2
Summary of Science and Technology Data

Data Type	Number	Source
Paper	112,515,346	KISTI Data Center
Journal	412,482	KISTI Data Center
National R&D Report	343,756	KISTI Data Center
Author	947,860	KISTI Data Center
Organization	43,329	KISTI Data Center
OA(Open Access) Paper	47,841,476	AccessON
Patent	45,260,490	KIPRIS
National R&D Project	1,081,266	NTIS
Journal, Patent Category	-	Web

Scholarly-related data from KISTI Data Center. KISTI provides data files including paper, journal, R&D report, author, affiliation and more. **Paper data** is assigned a unique ID and includes attributes like title, author(s), affiliation(s), associated journal, publication date, DOI, and abstract. Moreover, there are a citation relationship between a referencing paper and a referenced paper. Since all collected papers are identified by unique IDs, a referencing paper can be linked exactly to a referenced paper. **Journal data** has a unique ID and includes attributes such as the journal title, subject term(s). **R&D report** is a document that describes outcomes of a national R&D project. It includes attributes such as a unique ID, report title, author(s), author's affiliated organization(s), publication data and abstract. Also, our collected R&D reports are assigned to hierarchical category, called Korea National Science and Technology Standards Classification Codes by Korea Institute of Science & Technology Evaluation and Planning (KISTEP)[10]. While this categorization system is divided into three levels, we utilize only the first and second levels to classify our data. We find that the third level categories are overly granular, making it difficult to appropriately group and represent the nodes. **Author & Organization** indicates the author and affiliation of papers, patents, and reports. Identifying authors and affiliations suffers from the problem of homonyms. Seol *et al.* [58] proposed a method to identify authors and affiliations. Each author and affiliation has a unique ID and the attribute becomes the name. Additionally, author data are mapped to recognized identifiers such as ISNI (International Standard Name Identifier) and ORCID (Open Researcher and Contributor ID), ensuring excellent extensibility. As a result, we can capture relationships between paper(s), patent(s) and report(s) published by the authors. The data inventory includes 112,515,346 papers, 412,482 journals, 343,756 national R&D reports, 947,860 authors, and 43,329 organizations.

[10]https://www.kistep.re.kr/

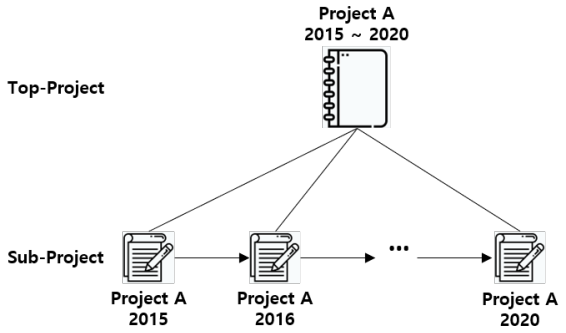

Fig. 2. Parent-child relationships of national R&D Project. The Top-Project is an object that represents a research project, and the Sub-Project is an object that belongs to the Top-Project and represents the research projects that are conducted each year.

Open Access scholarly-related data from AccessON[11]**.** AccessON facilitates free access to Open Access (OA) papers globally, encompassing a comprehensive array of all versions of OA papers, including published, post-print, and pre-print editions. This service ensures the provision of timely academic information through a systematic process. Each OA paper is uniquely identified and cataloged with details such as the paper title, author(s), affiliations, the journal of publication, publication date, DOI, abstract, and the type of OA. AccessON boasts a repository of approximately 47 million OA papers, significantly contributing to the democratization of knowledge by making scholarly work freely available to researchers worldwide.

Patent-related data from KIPRIS[12]**.** KIPRIS provided by the Korea Institute of Patent Information (KIPI) includes extensive patent records. Each patent record is uniquely identified and detailed with information such as the patent title, inventor(s), affiliations, and abstract. Additionally, patents are categorized according to the International Patent Classification (IPC) system, allowing for the organization of patents into specific fields of invention. With an approximate count of 47 million patents, KIPRIS serves as a vital resource for accessing a wide range of patents filed or registered worldwide, thereby supporting innovation and research by offering insights into existing patents and their classifications.

National R&D project-related data from NTIS[13]**.** National R&D projects cataloged by NTIS feature attributes such as the project title, abstract, and duration. Typically, these projects span several years, and for representational clarity, a node that encapsulates the project's entire timeline can be designated as a parent node. Subsequently, individual yearly projects are considered as child nodes linked in a chronological sequence. For instance, as illustrated in Figure 2, a multi-year R&D project running from 2015 to 2020 could be structured with the entire project as the parent node and each annual segment as a child node. Crucially, the metadata concerning R&D projects from NTIS forms the backbone of the ScienceON knowledge graph, enabling the interconnection of disparate science and technology data. Researchers produce various R&D deliverables, including papers, patents, and reports. These deliverables, submitted by the researchers, are manually correlated by operators with the unique IDs of scientific literature previously assigned by KISTI, ensuring accurate mapping with the corresponding research projects. This hands-on approach guarantees that each output is meticulously linked to its relevant project, enhancing the integrity and utility of the ScienceON knowledge graph. This methodology allows for the creation of links between a project entity and other entities such as papers, patents, and reports. The projects are methodically organized and depicted using parent-child relationships, enhancing the knowledge graph's ability to represent complex project timelines and associated research outputs.

Journal's category from web. The paper data compiled in ScienceON are invariably linked to a journal. However, these journals lack a taxonomy of scientific concepts. To address this, we perform web crawling on Google Scholar's journal categories[14]. The categories adopt a structure as depicted in Figure 3, where a journal may cor-

[11] https://accesson.kisti.re.kr/
[12] http://www.kipris.or.kr/
[13] https://www.ntis.go.kr/
[14] https://scholar.google.com/citations?view_op=top_venues&hl=en

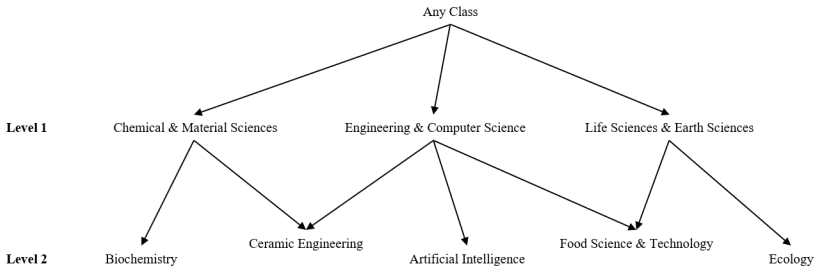

Fig. 3. An example of hierarchical classification in the paper domain. The taxonomy of papers is represented by a tree that expresses a hierarchical structure.

respond to multiple categories, allowing for a one-to-many (1:n) mapping. Additionally, the categories exhibit a two-level hierarchical structure.

International Patent Classification (IPC) from web. While the patent information acquired from KIPRIS is mapped to the IPC, it lacks the descriptive information about IPC itself. To remedy this, we obtain the necessary explanatory details of IPC through web crawling from the World Intellectual Property Organization (WIPO)[15]. The IPC is elaborately divided into four hierarchical levels for detailed patent categorization: section, class, subclass, and group. We use the class and subclass levels to categorize patents and the IPC's descriptive information to provide context and clarity to the classifications.

4.2. Data Acquisition

To construct a comprehensive database from various systems, our ETL process plays a crucial role. It involves systematically extracting data from authoritative sources such as the KISTI Scholarly Database, AccessON, NTIS, and KIPRIS. This information, once extracted, undergoes transformation to align with uniform standards for codes, terms, domains, and metadata, in adherence to the government's guidelines for database standardization. This ensures that the data integrated into the ScienceON system is consistent and maintains integrity.

A pivotal aspect of the ETL process is the assignment of unique identifiers to each document within ScienceON, which is essential for eliminating duplicates and establishing explicit relationships between different entities like papers, authors, and institutions. This is particularly important when reconciling datasets from various origins, where DOIs are instrumental in ensuring the uniqueness and authenticity of each record. Papers from KISTI and AccessON, for instance, are cross-referenced by their DOIs to filter out duplicates, with unique IDs bestowed upon the remaining distinct AccessON papers.

The process is also expanded to include web-crawled data, such as IPC and journal categories, which enrich the structural data from existing databases. By mapping this supplemental information to the established datasets, we enhance the database's depth, providing a richer context for analysis and knowledge discovery. Algorithm 1 is essential for categorizing journals within ScienceON, especially since these journals do not inherently have unique identifiers within our system. By employing a text pattern comparison methodology [59], as outlined in the algorithm, we can effectively categorize journals based on the correlation of titles collected from Google with those in the ScienceON database. This categorization process significantly augments the ScienceON knowledge graph, enabling it to support complex queries and facilitate advanced research initiatives.

4.3. Data Preparation

In the process thus far, there have been limitations in conducting knowledge discovery using the data collected. Therefore, to address these deficiencies and support data analysis based on machine learning, we augment the ScienceON Knowledge Graph with language models. Large language models (LLMs) such as BERT [60], GPT [61], and Llama [62], pre-trained on extensive corpus, have demonstrated exceptional performance across various natural

[15]https://www.wipo.int/classifications/ipc/en/

Algorithm 1: Assign journal's category from Google Scholar to ScienceON.

Input: $P = T_1, T_2, \cdots T_K$; Paper's title set of one journal from google by web crawling, C ; Category of one journal from google by web crawling, Q ; Query for searching paper's title on ScienceON, S ; paper's title list from ScienceON, F ; Sequence match function, J ; List of candidate journal's unique ID
Output: *Journal set*
for $k = 1 \cdots K$ **do**
  $S = Q(T_k)$; **for** $l = 1 \cdots L$ **do**
    if $F(T_k, S_l) > 0.9$ **then**
      $J \leftarrow S_l$'s unique ID of the journal;
$H \leftarrow$ Select the item with the highest count from J;
return $\{H, C\}$

language processing (NLP) tasks. LLMs enhance KGs by leveraging their advantages [63]. Traditional KGs are frequently incomplete, and existing methods for KG construction often neglect to consider textual information. To tackle these challenges, LLMs contribute to enhancing KGs by incorporating textual information.

We leverage LLMs to enhance the ScienceON Knowledge Graph in two ways. Firstly, we extract keywords from scientific and technical articles to generate additional nodes. The paper node, patent node, and report node, as collected, include the title and abstract as attributes. We leverage the abstract data prepending with their article title, utilizing LLMs to extract keywords and incorporate them as additional nodes to augment the knowledge graph. Specifically, OpenAI's GPT-3.5[16], one of the LLMs, is employed for keyword extraction, outperforming existing models for document keyphrases in scientific domain [64]. We extract top 5 keywords from the title and abstract of paper nodes, patent nodes, and report nodes, defining them as nodes within the ScienceON Knowledge Graph.

Secondly, we employ node attributes to define the node feature vector for knowledge representation learning. The attributes of specific nodes consist of text extracted from titles or abstracts. To represent this textual information as vectors, we utilize Sentence-BERT [65, 66]. Sentence-BERT is a variant of the BERT model tailored to produce semantic embeddings for sentences. These embeddings capture the semantic similarities between pairs of sentences, enabling tasks such as semantic text similarity. Each node for papers, patents, reports, and projects possesses a feature vector created by embedding text that combines the title and abstract. Similarly, journals and categories are represented as vectors derived from their titles. For paper, journal, and keyword, which have long input lengths and require rich textual representation, we utilized the all-mpnet-base-v2 model, which has a large dimension size and is the most widely used. For report and patent, which contain a mix of English and Korean, we employed the distiluse-base-multilingual-cased-v1 model, which supports multilingualism. As for categories, compared to other entities, the text is less diverse and more limited. Therefore, to reduce computational complexity, we used a model with a smaller dimension size. For nodes lacking textual information, such as authors and organizations, we arbitrarily assign 10-dimensional random vectors. In this case, the dimension size does not hold much significance. Table 3 provides an overview of the dimensions of the node feature vectors and the models used for their vectorization. Through this rigorous data preparation, we are setting the stage for a more insightful and comprehensive knowledge discovery, ensuring that our Knowledge Graph remains an accurate and current reflection of scientific advancements.

4.4. Knowledge Discovery

4.4.1. Methodical Development and Structuring of ScienceON Knowledge Graph

We have aggregated data from a variety array of sources into a consolidated relational database format, establishing a robust foundation for the development of a comprehensive knowledge graph. This process begins with the careful identification of entities, relationships, and attributes within the ScienceON database, chosen to represent the intricacies of the science and technology domain accurately. This selected data then serves as the cornerstone for our ontology development, outlining the conceptual and relational structure of our domain of interest.

[16] https://platform.openai.com/

Table 3
Feature vectors of the ScienceON Knowledge Graph entities.

Entity Type	Dimension	Feature Extraction Model
Paper	768	Sentence-BERT (all-mpnet-base-v2)
Journal	768	Sentence-BERT (all-mpnet-base-v2)
Report	512	Sentence-BERT (distiluse-base-multilingual-cased-v1)
Author	10	Random vector
Organization	10	Random vector
Patent	512	Sentence-BERT (distiluse-base-multilingual-cased-v1)
Project	10	Sentence-BERT (distiluse-base-multilingual-cased-v1)
Categories	384	Sentence-BERT (all-MiniLM-L12-v2)
Keyword	768	Sentence-BERT (all-mpnet-base-v2)

To aid in the construction of this ontology, we utilize Protégé[17], a premier tool for ontology modeling, known for its effectiveness in facilitating complex ontology designs. Based upon the developed ontology, we adopt the Resource Description Framework (RDF), a World Wide Web Consortium (W3C) standard, to model the knowledge graph's structure explicitly. To further bolster the data's integrity and reliability, we apply the W3C Shapes Constraint Language (SHACL)[18] to validate the RDF data, significantly enhancing the credibility of the included information. A pivotal addition to our methodology involves the transformation of relational database content into RDF format, leveraging R2RML mappings to ensure precise alignment with the ontology. This step is crucial for converting structured relational data into the RDF triples that populate the knowledge graph. Upon successful RDF validation using SHACL, the ScienceON Knowledge Graph is stored in a triple store, enabling sophisticated query capabilities via SPARQL[19]. We selected GraphDB for its exemplary performance as both a SPARQL engine and triple store.

This cohesive and standardized approach not only streamlines the representation and exploration of complex scientific datasets but also establishes a benchmark for future endeavors in knowledge graph construction and application. By adhering to W3C standards and employing rigorous data transformation and validation processes, we ensure that the ScienceON Knowledge Graph stands as a testament to the potential of semantic web technologies in facilitating advanced knowledge discovery.

4.4.2. Structuring ScienceON Knowledge Graph Ontology

The ScienceON Knowledge Graph is meticulously engineered to elucidate the complex interconnections amongst a multitude of scientific and technological datasets. It achieves this by defining a comprehensive network of entities and their interrelations, which are depicted in Figure 4. The figure showcases the full data model of the ScienceON Knowledge Graph, where the entities are not merely standalone points of data but are intricately connected to represent a rich tapestry of knowledge within the realms of science and technology. The ontology is robust, encompassing a variety of entities such as R&D projects, papers, journals, patents, reports, authors, institutions, keywords, and categories. Each entity is represented by a colored box, with the color-coding indicating the provenance of data—differentiating between data sourced internally within ScienceON and those aggregated from external databases and repositories. The incorporation of external schemas such as RDFS, XSD, and OWL enriches the ontology with a globally recognized standard, ensuring compatibility and interoperability across different knowledge systems. This alignment with international standards not only enhances the SKG's credibility but also facilitates its expansion and integration with other knowledge graphs and databases.

In the ScienceON Knowledge Graph ontology, entities such as "skg:TopProject", "skg:SubProject", "skg:Paper", "skg:Patent", and "skg:Report" serve as the basis to articulate the complex web of scientific and technological relations. These entities are intrinsically linked to descriptive metadata—titles, abstracts, and publication years—which are instrumental for semantic analysis and interpretation. "Skg:TopProject" signifies a high-level R&D initiative,

[17] https://protege.stanford.edu/
[18] https://www.w3.org/TR/shacl/
[19] https://www.w3.org/TR/sparql11-query/

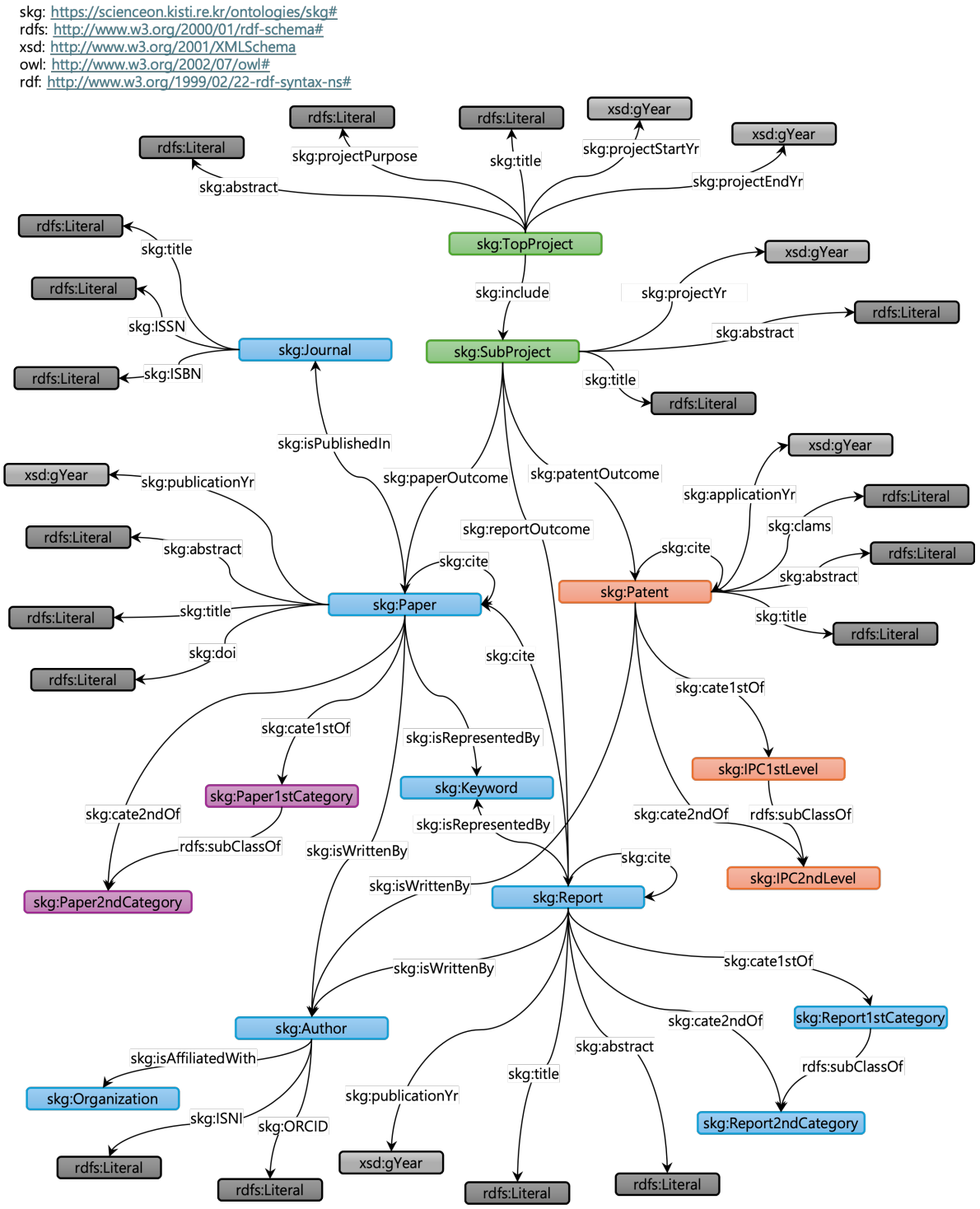

Fig. 4. The ScienceON Knowledge Graph ontology illustrating the intricate network of entities and their interrelationships representing scientific and technological data. This schema provides a comprehensive structure that mirrors the complex interconnections within the real-world science and technology domain.

which can encompass several "skg:SubProject" entities, illustrating the nested nature of extensive research endeavors. These subprojects are temporally ordered, with properties like "skg:projectYr" marking the commencement of each phase, thus mapping the project's timeline.

The entities "skg:Paper", "skg:Patent" and "skg:Report", which emerge as tangible outputs of "skg:SubProject", are detailed with comprehensive metadata, enriching the knowledge graph with actionable insights. Interlinked with these entities are authors, represented by "skg:Author", who are affiliated with institutions ("skg:Organization"), thus forming a network of intellectual contributions. And the "skg:Paper" entity, published within a "skg:Journal", carries unique identifiers such as ISSN and ISBN, facilitating its distinction and retrieval. This relational structure is further enriched by keywords ("skg:Keyword"), which are extracted from documents to semantically bridge the content within the knowledge graph. Categorization plays a crucial role, with "skg:cate1stOf" and "skg:cate2ndOf" denoting primary and secondary hierarchical levels, respectively. These categories provide a multi-layered classification system that enhances the discoverability and analytical potential of the documents within the graph.

We have incorporated a variety of relationships to support complex queries between classes. Each relationship carries the following meanings:

- "skg:include" connects the phases of an R&D project over several years.
- "skg:paperOutcome", "skg:patentOutcome", "skg:reportOutcome" link the papers, patents, and reports that are outcomes of R&D projects.
- "skg:cite" denotes the citation relationships among papers, patents, and reports.
- "skg:isPublishedIn" indicates that a paper has been submitted to a journal.
- "skg:isRepresentedBy" signifies that keywords extracted from the text data of papers and patents represent the respective documents.
- "skg:isWrittenBy" denotes the authors of papers, patents and reports.
- "skg:isAffiliatedWith" associates authors with specific institutions.
- "skg:cate1stOf" and 'skg:cate2ndOf" classify papers, patents, and reports into hierarchical categories.

By implementing such an extensive ontology, the ScienceON Knowledge Graph provides a foundational framework for representing and querying a wide range of scientific and technological data. Through the ScienceON Knowledge Graph, data is not only stored but also semantically linked, allowing for advanced queries and analyses that can uncover patterns, trends, and associations that are otherwise not apparent. This ontology-driven approach empowers researchers, data scientists, and technologists to delve deeper into the intricacies of scientific data, unlocking new insights and fostering innovation.

4.5. Data Sharing from ScienceON API Gateway

The ScienceON API Gateway[20], designed on a REST API, significantly enhances user interaction by providing streamlined access to a wealth of science and technology information as well as to processed knowledge through open API services. The RESTful architecture ensures scalability, simplicity, and flexibility, enabling seamless integration with a wide range of applications. By facilitating data retrieval in XML format, the ScienceON Web Service ensures that both raw data and value-added knowledge are structured and readily accessible to end-users. This approach, leveraging the REST API framework, underscores our commitment to fostering a user-centric ecosystem for sophisticated knowledge discovery and application, offering both raw data and processed knowledge products via API services.

This robust infrastructure is instrumental in empowering users with the tools to transform extensive datasets into actionable insights, thereby enriching the research landscape. ScienceON's integrated system simplifies the intricacies of managing and utilizing complex data, markedly increasing its value for a wide array of stakeholders, including researchers, developers, and policymakers. The open API service not only guarantees data accessibility and reliability but also ensures that the data is optimized for in-depth analysis, driving forward the pursuit of knowledge and innovation.

[20]https://scienceon.kisti.re.kr/apigateway/

5. Measuring Quantitative and Qualitative Aspects Based on Graph-Related Tasks

5.1. Quantitative Analysis of Graph-Related Tasks

5.1.1. Knowledge Representation

We experiment on several knowledge representation learning tasks such as node classification and link prediction in the science and technology domain. These tasks enable the performance of applications such as citation analysis, conference analysis, and trend analysis. [67] We employ the GNNs framework, which has recently demonstrated superior performance among various knowledge representation learning models, to learn the ScienceON Knowledge Graph. A graph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consists of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$ with node feature vector $\mathcal{X}_v$ for $v \in \mathcal{V}$. A general GNN framework composes two operations of the aggregation function and the update function to learn a node representation vector, denoted as h_v.

$$Aggregation : a_v^{(\ell)} = \text{Aggregate}^{(\ell)}(h_u^{(\ell-1)}, \forall u \in \mathcal{N}_v), \tag{1}$$

$$Update : h_v^{(\ell)} = \text{Update}^{\ell}(h_v^{(\ell-1)}, a_v^{\ell}), \tag{2}$$

where $\mathcal{N}_v$ is the neighborhood of the node v, $h_u^{(\ell-1)}$ is the feature vector of node v at the $\ell-th$ layer of a GNN.

We emphasize various GNN frameworks to address tasks such as node classification, link prediction, and clustering within the scope of the the ScienceON Knowledge Graph. However, since most existing GNN frameworks use loss functions for single-label problems, we need to define a loss function to solve multi-class and multi-label problems. For graph representation learning with multi-class, multi-label, we can optimize the GNN model by binary cross entropy loss with combing a sigmoid layer.

$$L(x) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \sigma(h_i) + (1 - y_i) \log(1 - \sigma(h_i)), \tag{3}$$

where N is the number of training nodes, h is an embedding vector of training nodes, σ is a sigmoid function and y is a label vector of training nodes.

5.1.2. Experimental Setup

Dataset. In our investigation of the ScienceON Knowledge Graph, we have engaged in a selective extraction of data to train graph representation models and conduct experiments on various tasks. Given the extensive size of the entire knowledge graph, we faced computational constraints that necessitated the extraction of a representative subset of data. Our extraction criteria focused on R&D projects conducted from 2011 to 2020, from which we meticulously selected entities and relations that were central to our study.

The count of the primary entities extracted for this exercise is as follows:

- Top-Projects: 36,230
- Sub-Projects: 64,883
- Papers: 360,612
- Patents: 102,703
- Reports: 21,487

In addition to these entities, we have categorized our papers, patents, and reports into primary and secondary categories to enable a nuanced classification that aligns with the intricate structure of scientific research. The status of these categories is as outlined:

- Paper 1st Level Category: 8
- Paper 2nd Level Category: 100
- Patent 1st Level Category: 54
- Patent 2nd Level Category: 139
- Report 1st Level Category: 17

– Report 2nd Level Category: 88

Tasks and Evaluation. We evaluate several baseline models on the ScienceON Knowledge Graph. For node classification, we predict classes such as paper 1st level category, paper 2nd level category, patent 1st level category, patent 2nd level category, report 1st level category, and report 2nd level category. We measure F1-scores for both 1st level and 2nd level category node classification. For link prediction, we predict links between papers. After learning baseline models for node representations of papers, the models are used to compute a probability of each pair of papers to be linked. We evaluate link prediction with the Area Under the ROC Curve (AUC) and Average Precision (AP).

Baselines. We apply various state-of-the-art GNN approaches to the ScienceON Knowledge Graph. All baselines are implemented using the PyTorch Geometric (PyG) package [68]. The first group of baselines is designed for homogeneous graphs, which includes the following approaches:

– GCN [51] is a semi-supervised graph convolutional network for homogeneous graphs, which aggregates embeddings of neighboring nodes through average and subsequent linear projection.
– GraphSAGE [52] is an inductive approach that generates node embeddings for unseen data by employing node features and learning a function that samples and aggregates features from a node's local neighborhood.
– GAT [54] is a semi-supervised homogeneous GNN that incorporates an attention mechanism in graph spatial domain convolutions.

The second group of baselines is models for heterogeneous graphs, including:

– Metapath2vec [69] is a traditional heterogeneous graph embedding model that utilizes metapath-based random walks for skip-gram embeddings.
– HAN [56] utilizes hierarchical attentions to aggregate neighbor information from diverse metapaths, forming a heterogeneous graph neural network that simultaneously employs node-level and semantic-level attention.
– HGT [57] defines meta relations as the types of two adjacent nodes and their link, assigning diverse attention weight matrices to these meta relations, enabling the model to incorporate type information.

For metapath2vec, we set the window size to 7, the walk length to 100, the walks per node to 500, the number of negative samples to 7 and the size of each embedding vector to 128. For all GNNs including GCN, GraphSAGE, GAT, HAN and HGT, we initialize parameters randomly and leverage the Adam [70] with the learning rate set to 0.001, the weight decay set to 0.001. We set the dropout rate to 0.5 and use early stopping with a patience of 50. We set the hidden dimension to 256 for all baselines. We set the number of heads to 6 for all multi-head attention-based models.

5.1.3. Node Classification

To assess the ScienceON Knowledge Graph's contribution to machine learning tasks, we evaluated its impact on document classification efficacy. The comparison of classification performance, with and without the integration of the ScienceON Knowledge Graph, is illustrated in Figure 5. We employed two distinct models for this experiment: the baseline **sentenceBERT** [65, 66] and the enhanced **sentenceBERT + Graph Embedding**. Both models were trained on paper nodes, leveraging the 'abstract' attribute with a train-to-test data ratio of 80:20. The baseline **sentenceBERT** model utilizes only the abstract text for embedding and employs logistic regression for classification. Conversely, **sentenceBERT + Graph Embedding** enriches this approach by incorporating graph-structured data as an additional feature set alongside the textual information.

The evaluation, as depicted in Figure 5, benchmarks the weighted F1 score against the number of training iterations for both models. The graph indicates that while **sentenceBERT** achieves a quicker initial convergence, the **sentenceBERT + Graph Embedding** model ultimately attains a superior F1 score. When utilizing both textual information and graph information together, there is an improvement in the f1-score of over 10% compared to not using them in combination. This demonstrates that the integration of graph information provides a significant performance advantage, underscoring the value of including the ScienceON Knowledge Graph in the model. Hence, the evidence suggests that **sentenceBERT + Graph Embedding**, by leveraging the interconnected structure of the knowledge graph, enhances the model's predictive accuracy compared to the text-only baseline.

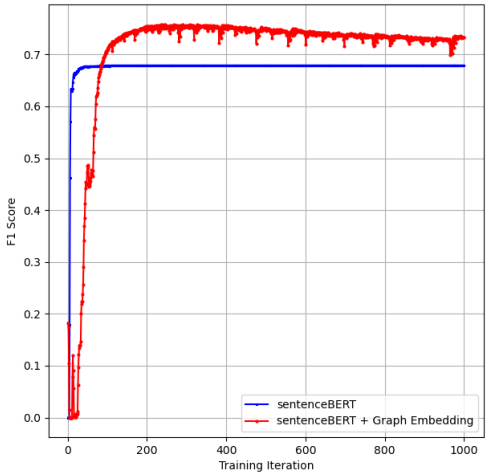

Fig. 5. Comparision of document classification performance both with and without the incorporation of the ScienceON Knowledge Graph.

We continue to report averaged Micro-F1 and Macro-F1 scores of repeating the process 10 runs for all classification tasks. The results, detailed in Table 4, present the node classification performance when categorizing paper nodes into their 1st level category, noting that a paper may be classified into multiple categories. We only present experimental results for the rest of the tasks with HAN and HGT in Table 5, since the F1-scores of metapath2vec, GCN, GAT and GraphSAGE converge to 0 in the rest of node classification tasks.

As shown in Table 4, HGT outperforms other baselines in terms of both metrics on the node classification for the 1st level category of the paper. HGT also outperforms other baselines beyond comparison on all of node classification tasks in Table 5. However, it does not necessarily mean that heterogeneous methods always outperform homogeneous methods. In certain scenarios, such as when comparing against the HAN model, homogeneous methods like GCN, GAT, and GraphSAGE show superior performance. Additionally, metapath2vec, the traditional graph embedding model, has been shown to be empirically better than GNNs in Table 4. This highlights the nuanced nature of model performance, where the choice of model should be contextually driven, accounting for the specific characteristics of the dataset and task at hand.

In Table 5, to validate our model to properly learn representation by aggregating information from the neighborhood, we further evaluate HAN and HGT on subgraphs in which some node types are eliminated. In other words, if the model learns representation vectors by effectively aggregating information from neighboring nodes in the ScienceON Knowledge Graph, it is expected to outperform the model trained on its subgraph. To illustrate, for paper node classification, we extract and examine a subgraph containing paper nodes, project nodes, author nodes, organization nodes, and their interlinking relationships. Similarly, for patent and report node classifications, we consider corresponding subgraphs populated with relevant nodes and their connections.

From Table 5, let HGT represents the model that learns the entire ScienceON Knowledge Graph, and HGT-Subgraph represents the model that learns a subgraph of the ScienceON Knowledge Graph. We expected that HGT to be better than HGT-Subgraph, but HGT has shown similar performance to HGT-Subgraph and in some cases or HGT-Subgraph has shown even better performance. That said, HGT model, has the restriction of aggregating information from the neighborhood. This suggests that HGT model is limited in scaling up model, as there are obstacles related to oversmoothing and expensive computation.

5.1.4. Link Prediction

In the realm of network analysis, the link prediction task serves as a crucial benchmark for assessing the predictive power of GNN models. In our study, we define a positive node pair as two papers connected by a link, signifying a relationship such as citation or thematic similarity. On the other hand, a negative node pair consists of two papers

Table 4

Quantitative results on the node classification task of 1st level category of the paper.

	Baseline	**Metapath2Vec**		**GCN**		**GAT**		**GraphSAGE**		**HAN**		**HGT**	
Classfication Case	Test Size	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Paper 1st Category	20%	0.6219	0.4414	0.6550	0.4280	0.6733	0.4161	0.7063	0.4471	0.5276	0.3649	**0.7217**	**0.6151**
	40%	0.6098	0.4304	0.6558	0.4247	0.6736	0.4170	0.7070	0.4494	0.4155	0.2644	**0.7149**	**0.5997**
	60%	0.5900	0.4035	0.6524	0.4230	0.6715	0.4160	0.6951	0.4469	0.4146	0.2628	**0.7048**	**0.5806**
	80%	0.5546	0.3558	0.6579	0.4245	0.6724	0.4153	0.6965	0.4436	0.5156	0.3551	**0.6983**	**0.5410**

Table 5

Quantitative results on the node classification task of all category with HAN and HGT. HGT is a measurement of node classification on the ScienceON Knowledge Graph, and HGT-Subgraph is the results on its subgraph. Bold numbers represent the highest performance in each row.

	Baseline	**HAN**		**HGT**		**HGT-Subgraph**	
Classification Case	Test Size	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Paper 1st Level Category	20%	0.5276	0.3649	0.7217	0.6151	**0.7339**	**0.6212**
	40%	0.4155	0.2644	0.7149	0.5997	**0.7217**	**0.6036**
	60%	0.4146	0.2628	0.7048	0.5806	**0.7073**	**0.5903**
	80%	0.5156	0.3551	0.6983	0.5410	**0.7024**	**0.5609**
Paper 2nd Level Category	20%	0.2329	0.2037	0.4245	0.4201	**0.4254**	**0.4232**
	40%	0.2305	0.2034	0.4065	0.3952	**0.4192**	**0.4106**
	60%	0.2618	0.2359	0.3962	0.3850	**0.3993**	**0.3873**
	80%	0.2145	0.1884	0.3580	0.3336	**0.3681**	**0.3460**
Patent 1st Level Category	20%	0.0491	0.0106	0.6005	**0.4188**	**0.6015**	0.4130
	40%	0.1566	0.0866	**0.5870**	**0.3978**	0.5828	0.3797
	60%	0.1407	0.0778	**0.5708**	**0.3625**	0.5680	0.3543
	80%	0.1423	0.0782	**0.5287**	**0.2991**	0.5204	0.2949
Patent 2nd Level Category	20%	0.0415	0.0208	0.4814	**0.3435**	**0.4819**	0.3284
	40%	0.0480	0.0257	**0.4642**	**0.3256**	0.4617	0.3114
	60%	0.0528	0.0316	**0.4409**	**0.2931**	0.4335	0.2783
	80%	0.0522	0.0298	**0.3879**	**0.2369**	0.3806	0.2328
Report 1st Level Category	20%	0.0033	0.0033	**0.5158**	0.4345	0.5143	**0.4591**
	40%	0.0030	0.0030	**0.4806**	**0.3993**	0.4727	0.3991
	60%	0.0296	0.0296	**0.4497**	**0.3673**	0.4222	0.3496
	80%	0.0086	0.0086	**0.4110**	**0.3242**	0.3649	0.2799
Report 2nd Level Category	20%	0.0032	0.0032	**0.2835**	**0.2678**	0.2432	0.1740
	40%	0.0029	0.0029	**0.2562**	**0.2415**	0.2322	0.1765
	60%	0.0026	0.0026	**0.2172**	**0.1928**	0.2014	0.1554
	80%	0.0024	0.0024	**0.1777**	**0.1557**	0.1571	0.1111

with no connection between them, indicating that no direct relationship is known. Through this binary classification framework, we aim to gauge the models' ability to infer hidden or potential connections within the knowledge graph.

As evidenced by the results in Table 6, the HGT model outstrips other baseline models in terms of the AUC and AP metrics. Notably, while the AUC of 0.9824 and AP of 0.9810 for HGT are impressive, the superiority of HGT over other models is not as pronounced in link prediction as it is in node classification tasks. This suggests that while HGT is adept at uncovering complex patterns in node features for classification purposes, its edge in discerning latent links is somewhat more modest, highlighting areas for potential enhancement in future iterations of model development.

Table 6
Quantitative results on the link prediction task.

Metric	GCN	GraphSAGE	GAT	HAN	HGT
AUC	0.9717	0.9357	0.9380	0.9278	**0.9824**
AP	0.9739	0.9449	0.9180	0.8991	**0.9810**

5.2. Qualitative Insights from Graph-Related Tasks

5.2.1. Graph Embedding Visualization

Beyond the numerical analysis, we delve into a qualitative evaluation of our Knowledge graph through the visualization of node embeddings. This section focuses on the embeddings generated by the HGT, which previously showcased superior performance. For an intuitive understanding of these embeddings, we employed t-distributed Stochastic Neighbor Embedding (t-SNE) [71], a technique renowned for its efficacy in reducing high-dimensional data into a two-dimensional space conducive to visual interpretation. This process facilitates the visualization of node embeddings, with nodes color-coded according to their categorical affiliations, offering insights into the spatial distribution of various scientific disciplines within our knowledge graph.

A closer look at the interplay between the categories of “Chemical & Material Sciences” and “Engineering & Computer Science” is provided in Figure 6b, with the former depicted in blue and the latter in red. Papers that embody both fields are shown in purple, serving as a bridge between the two domains and situated at their juncture, thereby illustrating the overlapping knowledge areas. This visualization not only evidences the model’s capacity to distinguish between singular and intersecting domains but also its sensitivity to the complex landscape of scientific research. Subsequent figures, Figure 6c and Figure 6d, extend this analysis to other category pairs, consistently revealing the model’s adeptness at mapping the multifaceted relationships between various scientific disciplines.

Figure 7 provides a granular view of paper embeddings within the specific lower fields of "Chemical & Material Sciences" and "Engineering & Computer Science", showcasing the model’s adeptness at classification. In Figure 7a, the visualization illustrates a clear demarcation between papers classified under these higher fields, affirming the model’s precision in maintaining distinct domain boundaries. Figure 7b reveals an interesting phenomenon where nodes associated with "Engineering & Computer Science" (teal-colored) are positioned close to the intersection with "Chemical & Material Sciences". This proximity suggests a conceptual overlap or interdisciplinary nature of certain papers that bridge the two fields, underscoring the model’s capacity to reflect the complex, multidimensional structure of scientific knowledge. Further analysis through Figures 7c and 7d plots the distribution of nodes by their lower field classifications, confirming that each class is distinctly separated.

These visual analyses provide deep insights into the ScienceON Knowledge Graph’s ability to intricately map the complex and interwoven landscape of the science and technology domain. The clear delineation of classes, coupled with the identification of interdisciplinary nodes at the intersection of major fields, demonstrates the model’s refined comprehension of the scientific domain. This qualitative assessment sheds light on the knowledge representation’s depth and sophistication, revealing the graph’s rich structural intricacies and the dynamic interrelationships it encapsulates.

5.2.2. Network Analysis of Patent and Paper

Traditional knowledge graphs often fail to capture the heterogeneous nature of the relationships between patents and papers, a limitation that overlooks the interplay between scientific research and technological innovation. The ScienceON Knowledge Graph, however, stands out by bridging this gap, offering a more comprehensive framework that enables the interpretation of the intricate relationships between patents and papers. To enhance our understanding of the intricate relationships among various node types within the ScienceON Knowledge Graph, we delve into the analysis of co-occurrence frequencies, specifically between patents and journals. This approach involves calculating the frequency with which a target node (e.g., a IPC) appears alongside a source node (e.g., a paper), thereby gauging the proximity and relevance between them. As co-occurrence frequency escalates, the relational closeness between the target and source nodes intensifies, providing a quantitative measure of their association.

Illustrated in Figure 8, our analysis zeroes in on the dynamic interplay between journal nodes and IPC nodes. Here, IPC nodes serve as source nodes, while journals associated with related patents act as target nodes. The

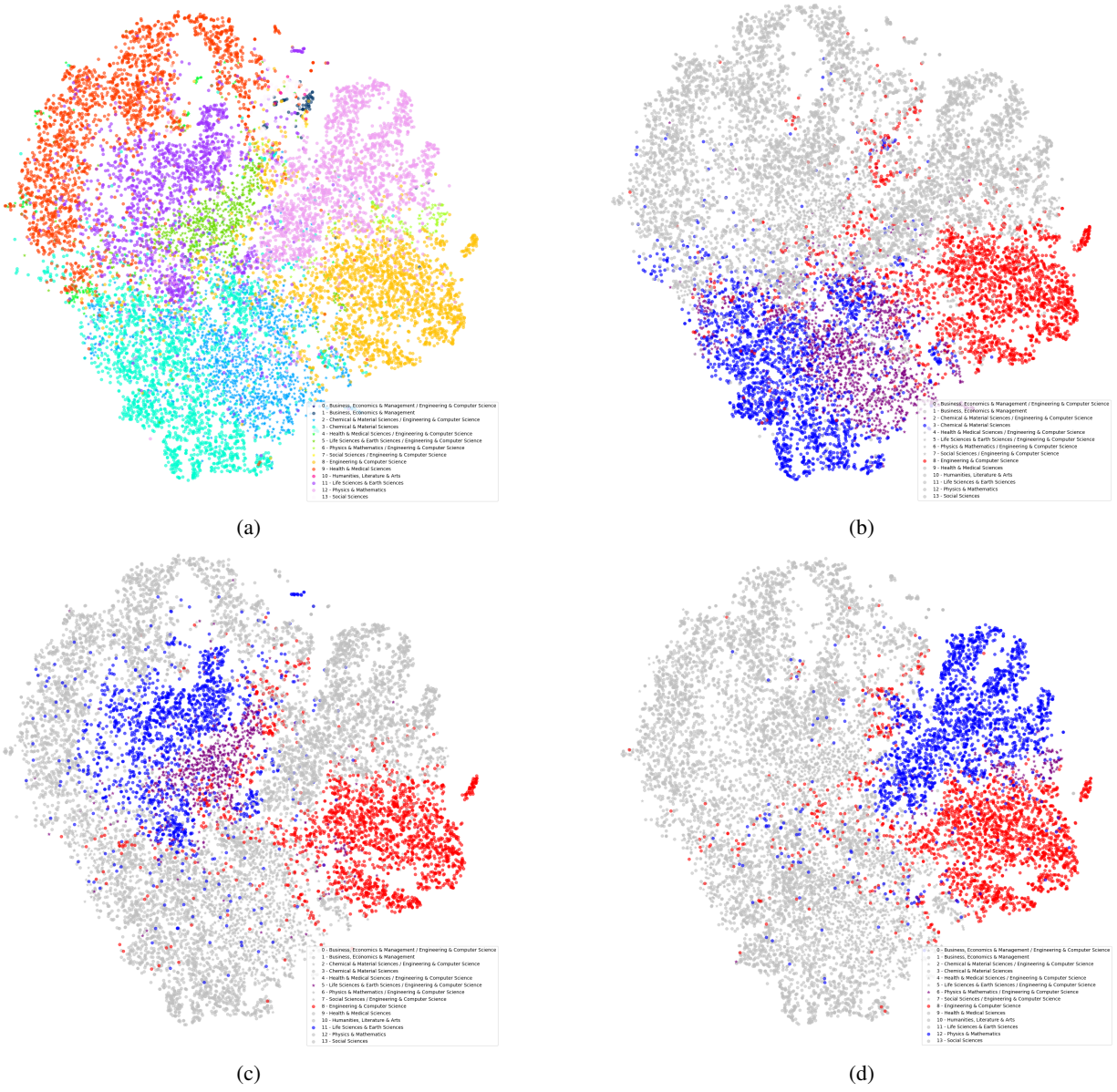

Fig. 6. Embedding of 8 paper's class in 1st level category. (a) Node colors denote classes. Nodes marked with a circle are single-label classes, and nodes marked with a asterisk are multi-label classes. (b) The class "Chemical & Material Sciences" is colored blue, the class "Engineering & Computer Science" is colored red, nodes that belong to both are colored purple, and the rest are gray. (c) The class "Life Sciences & Earth Sciences" is colored blue, the class "Engineering & Computer Science" is colored red, nodes that belong to both are colored purple, and the rest are gray. (d) The class "Physics & Mathematics" is colored blue, the class "Engineering & Computer Science" is colored red, nodes that belong to both are colored purple, and the rest are gray.

visualization employs varying thicknesses of links to denote the degree of proximity between the IPC and its corresponding journals; a thicker link suggests a stronger relationship. In Figure 8a, we spotlight the "Electric communication technique" IPC category (denoted as H04), a higher field of patent classification. The analysis indicates a pronounced connection between this category and journals within domains such as “Engineering & Computer Science”, “Computer Networks & Wireless Communication”, “Signal Processing”, “Microelectronics & Electronic Packaging”, and “Computing Systems”, highlighting significant interdisciplinary overlaps. Figure 8b presents the

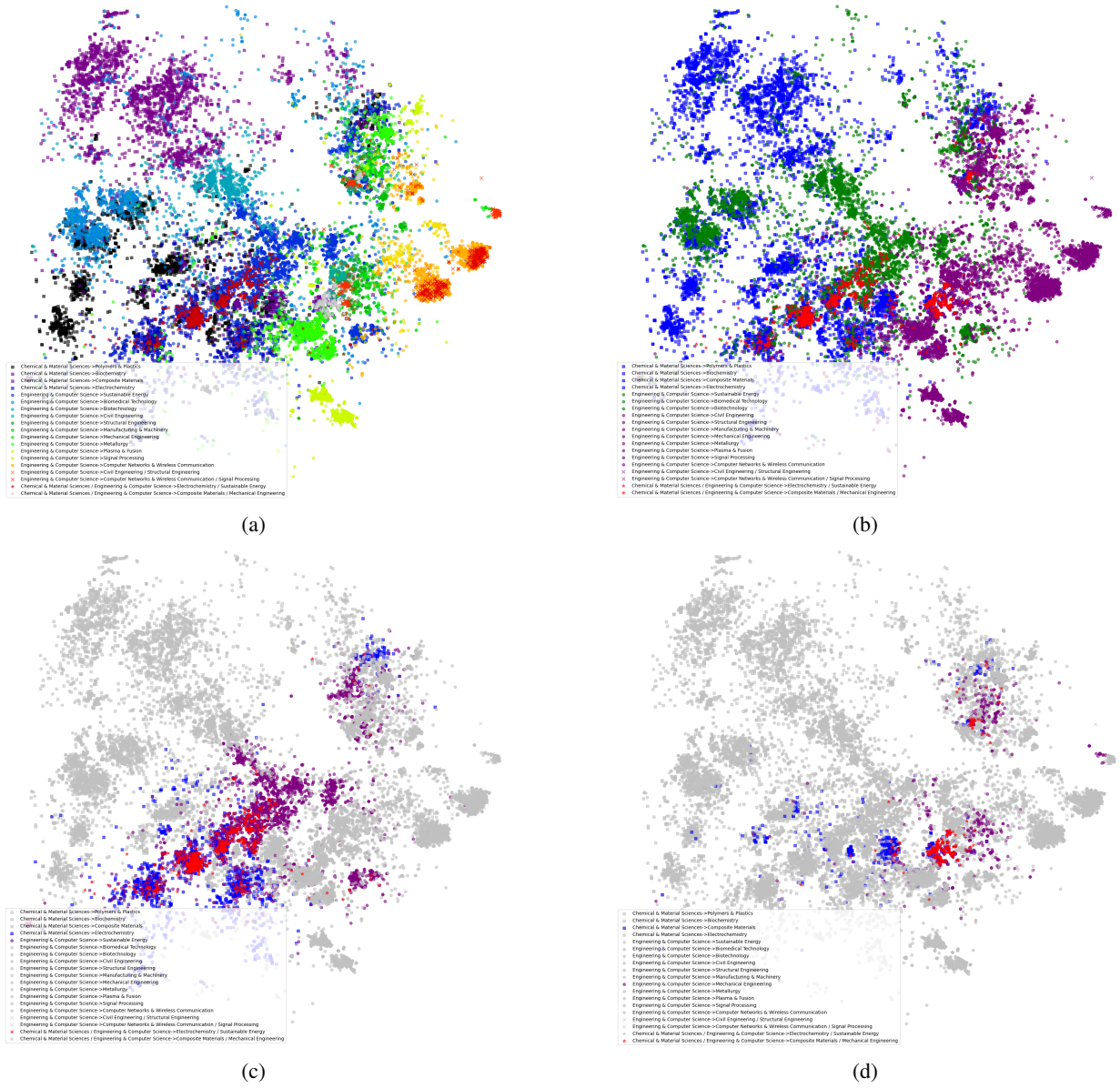

Fig. 7. Embedding of selected papers belonging to "Chemical & Material Sciences" and "Engineering & Computer Science". (a) Node colors denote classes. Nodes marked with a rectangle are "Chemical & Material Sciences" classes, nodes marked with a circle are "Engineering & Computer Science" classes, nodes with an X are multi-label classes that belong to "Engineering & Computer Science" class, and nodes with an asterisk are multi-label classes that belong to both classes. (b) "Chemical & Material Sciences" nodes are colored blue and "Engineering & Computer Science" nodes are colored purple, papers similar to "Chemical & Material Sciences" in "Engineering & Computer Science" are colored teal, and nodes belonging to both are colored red. (d) Chemical & Material Sciences" nodes are colored blue, "Engineering & Computer Science" nodes are colored purple, nodes that belong to both are colored red.

same results as Figure 8a. Further exploration in Figure 8c examines the "Image data processing or generation, in general" IPC category (G06T), a lower field of patent classification. The proximity of this category to journals like “Sensors”, “Electronics Letters”, “IEEE Access”, “Scientific Reports”, and “Multimedia Tools and Applications” underscores meaningful relationships, suggesting thematic commonalities that bridge these areas of research. Similarly, Figure 8d shows the same results as Figure 8c.

Fig. 8. The IPC related the journal based on the frequency of co-occurrence. The IPC nodes are colored teal in the top center, and the journal nodes are colored blue. As the thickness of the link increases, it indicates that the IPC and the journal are getting closer to each other. (a) and (b) are IPC nodes belonging to the 1st level patent category, and (c) and (d) are IPC nodes belonging to the 2nd level patent category.

Through this analysis, the ScienceON Knowledge Graph emerges not only as a repository of data, but also as a dynamic map of the confluence of scientific disciplines and technological innovation. The co-occurrence frequency approach reveals how patents and papers, as proxies for broader research trends, are interconnected within the vast ecosystem of science and technology, providing a perspective on the collaborative nature of modern research efforts.

5.2.3. Network Analysis of Scientific Keyword

The ScienceON Knowledge Graph offers valuable insights into the interconnections among scientific keywords across different fields of study. This visualization underscores the ways in which the same keyword may align with different sets of related terms, depending on the academic or professional field. It highlights the critical need for

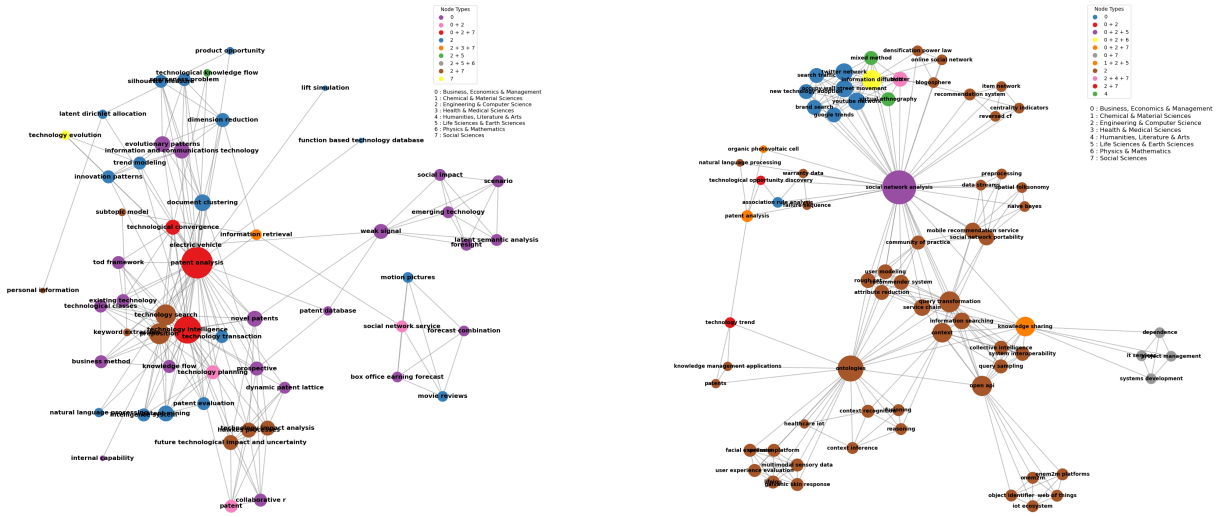

Fig. 9. Scientific keyword projection network. (a) represents a network of keywords centered on "patent analysis", and (b) depicts a network of keywords centered on "social network analysis". Keywords with varying characteristics are linked depending on the research area.

precise field specification when recommending or analyzing keyword lists within the science and technology domain to ensure the relevance and accuracy of the results.

As depicted in Figure 9, the graph visualizes how certain keywords cluster together, revealing their varying associations across disciplines. For instance, Figure 9a highlights the keyword "patent analysis" and its connections to distinct sets of keywords based on their disciplinary context. Keywords such as "document clustering," "dimension reduction," and "latent Dirichlet allocation" are closely linked within the "Engineering & Computer Science" field, indicating a technical focus. In contrast, within the "Business, Economics & Management" domain, terms like "novel patents," "existing technology," and "business method" cluster together, reflecting a more business-oriented perspective. Similarly, Figure 9b explores the keyword "social network analysis" and its diverse linkages across fields. Here, keywords related to "Business, Economics & Management" and "Engineering & Computer Science" converge around the theme of social network analysis yet form distinct groupings that mirror their respective disciplinary lenses.

6. ScienceON Knowledge Graph Applications

In this section, we demonstrate the capabilities of the ScienceON Knowledge Graph System through the development of various functionalities aimed at providing science and technology information to the general public via the ScienceON web service. By showcasing examples of the user interface (UI) along with SPARQL queries designed to facilitate these services, we illustrate how the ScienceON Knowledge Graph System can effectively support a wide range of services for disseminating scientific and technological knowledge to non-expert audiences. The UI examples and SPARQL queries serve to highlight the system's ability to curate and deliver relevant information in an accessible and user-friendly manner.

The first application provides access to scientific papers collected by ScienceON. As shown in Figure 10, this service delivers various metadata related to papers, including titles, journals, authors, affiliations, publication years, DOIs, abstracts, keywords, and references. This allows users to easily understand detailed information about a paper. On the right side of the figure, an example SPARQL query is presented, showcasing how data can be queried in a manner simpler than the equivalent SQL queries for relational databases.

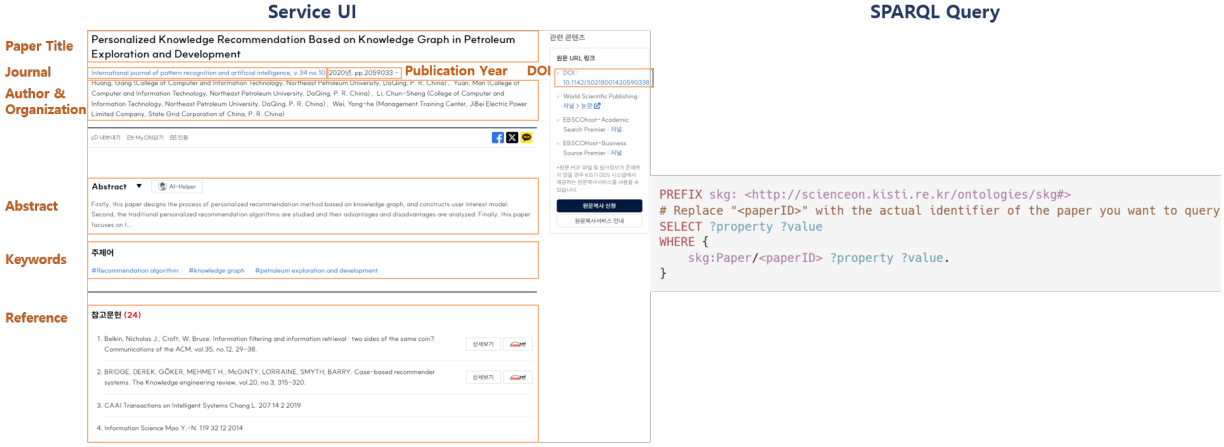

Fig. 10. Detailed view of a scientific paper's metadata provided by the ScienceON service, alongside an example SPARQL query for accessing this information.

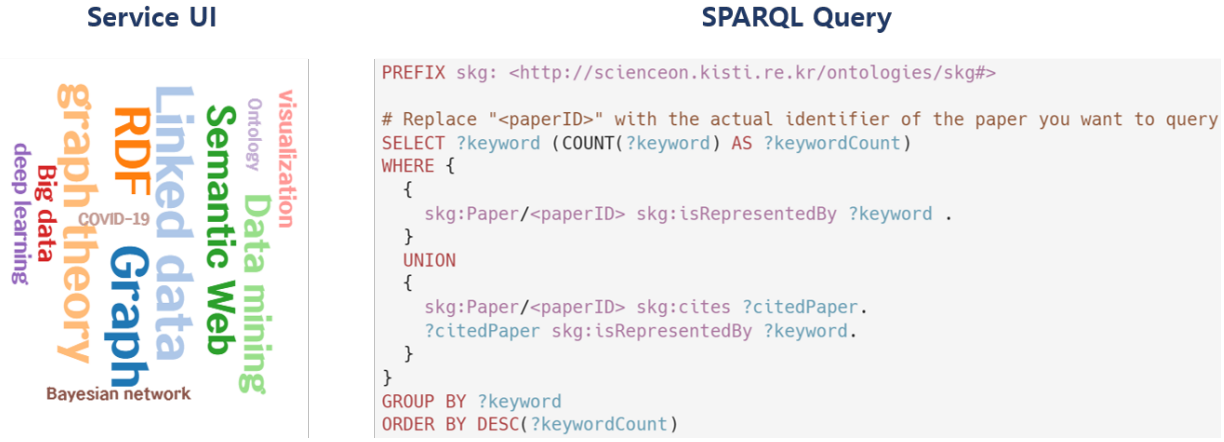

Fig. 11. Visualization of a keyword cloud generated from a selected paper and its references, demonstrating the network of scientific topics.

The second application generates keyword clouds for papers. Illustrated in Figure 11, this service searches for keywords associated with a selected paper and its references, visualizing them all together. The corresponding query, as shown on the right, retrieves and groups keywords from all related papers, sorting them by frequency. The more frequently a keyword appears, the more significant it is, which is represented by its larger size in the keyword cloud.

Lastly, we offer a feature that retrieves national R&D projects by year and their resulting outputs. In Figure 12, projects from 2014 to 2016 are displayed along with their yearly published papers, patents, and reports in a timeline format. The SPARQL query, as demonstrated, locates all Sub-Projects belonging to a Top-Project and fetches all research outputs produced from these Sub-Projects.

7. Conclusion

The pursuit of integrating and analyzing scientific and technological information is fraught with challenges. The exponential growth in data volume, coupled with the complexity of relationships within this domain, necessitates sophisticated solutions. Such challenges underscore the need for advanced data structures that can not only integrate vast amounts of interconnected data but also enable meaningful knowledge discovery.

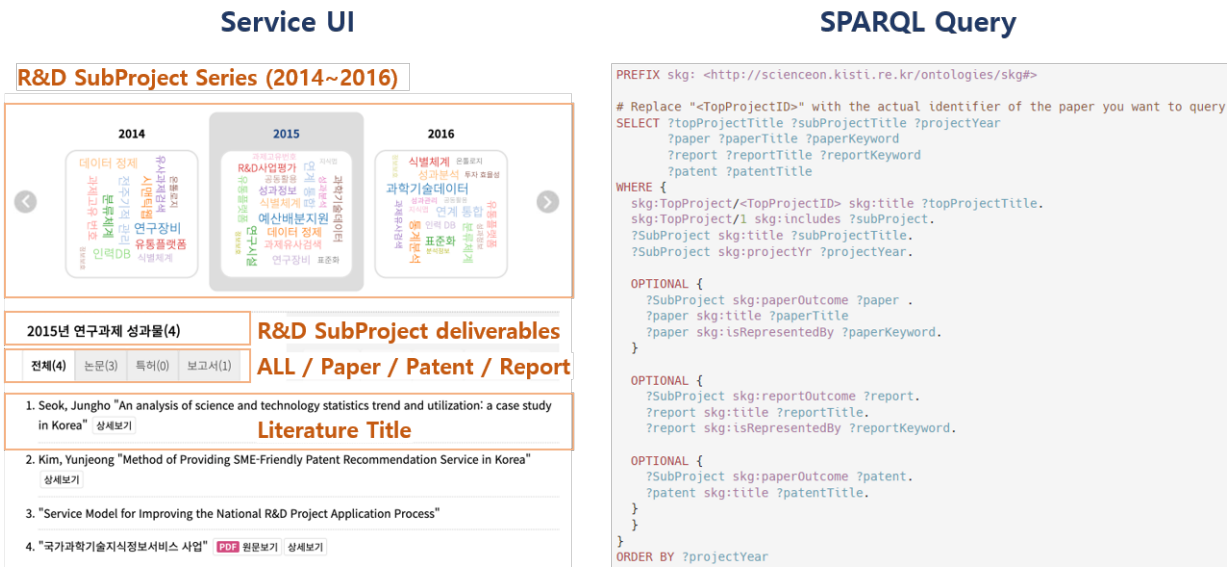

Fig. 12. A timeline visualization of national R&D projects from 2014 to 2016, including their yearly outputs of papers, patents, and reports, with an example SPARQL query for data retrieval.

In addressing the complexities inherent in data-driven analysis of scientific and technological domains, we have developed the ScienceON Knowledge Graph System. This advanced system is built upon a solid data ecosystem that aggregates a wide range of information, encompassing R&D projects, papers, patents, reports, authors, organizations, keywords, and categories. A key strength of the ScienceON Knowledge Graph System is its commitment to strict curation and standardization protocols. This approach ensures the high quality and reliability of the data, setting a foundation that supports intricate queries and analyses. These capabilities surpass those offered by conventional systems, offering a more comprehensive exploration of the interconnected web of scientific and technological knowledge.

The value of the ScienceON Knowledge Graph System has been further demonstrated through a series of experiments and the development of user-centric applications. These endeavors have showcased the system's capability to enhance knowledge discovery, offering deeper insights into the scientific and technological landscape. By conducting experiments for graph analysis, we have validated the efficacy of the knowledge graph in capturing the complicated relationships between various data points. Moreover, the applications developed on this system have proven to be instrumental in disseminating scientific knowledge, thereby reinforcing the practical utility and significance of the ScienceON Knowledge Graph System.

Looking ahead, the evolution of the ScienceON Knowledge Graph System will be guided by feedback from a broad range of stakeholders, including researchers, policymakers, and industry professionals. This iterative process of feedback and refinement will not only aid in the continuous collection and curation of scientific and technological data but also in the enhancement and expansion of the knowledge graph. Furthermore, by fostering the development of diverse applications, we aim to extend the reach and impact of the ScienceON Knowledge Graph System, ensuring it remains a vital resource in the exploration and understanding of science and technology. Through this collaborative and dynamic approach, we aspire to unlock new horizons in data-driven science, paving the way for breakthroughs that could transform our understanding of the world.

Acknowledgements

This research was supported by Korea Institute of Science and Technology Information (KISTI)(No. K24L4M2C5) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00214065).

References

[1] N. Heist, S. Hertling, D. Ringler and H. Paulheim, Knowledge Graphs on the Web-An Overview, *Knowledge Graphs for eXplainable Artificial Intelligence* (2020), 3–22.

[2] H. Yuan, H. Yu, S. Gui and S. Ji, Explainability in Graph Neural Networks: A Taxonomic Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(5) (2023), 5782–5799.

[3] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson and J. Taylor, Industry-Scale Knowledge Graphs: Lessons and Challenges: Five Diverse Technology Companies Show How It's Done, *Communications of the ACM* **62**(8) (2019), 36–43.

[4] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J.P. Hsu and K. Wang, An Overview of Microsoft Academic Service (MAS) and Applications, in: *WWW*, 2015, pp. 243—246.

[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang and Z. Su, ArnetMiner: Extraction and Mining of Academic Social Networks, in: *KDD*, 2008, pp. 990—998.

[6] F. Zhang, X. Liu, J. Tang, Y. Dong, P. Yao, J. Zhang, X. Gu, Y. Wang, B. Shao, R. Li and K. Wang, OAG: Toward Linking Large-Scale Heterogeneous Entity Graphs, in: *KDD*, 2019, pp. 2585—2595.

[7] S. Peroni and D. Shotton, OpenCitations, An Infrastructure Organization for Open Scholarship, *Quantitative Science Studies* **1**(1) (2020), 428–444.

[8] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J.F. Rousseau, X. Li, W. Xu, V.I. Torvik, Y. Bu, C. Chen, I.A. Ebeid, D. Li and Y. Ding, Building a PubMed Knowledge Graph, *Scientific Data* **7**(1) (2020), 205.

[9] T. Ebesu and Y. Fang, Neural Citation Network for Context-Aware Citation Recommendation, in: *SIGIR*, 2017, pp. 1093—1096.

[10] S. Gupta and V. Varma, Scientific Article Recommendation by Using Distributed Representations of Text and Graph, in: *WWW*, 2017, pp. 1267—1268.

[11] S. Sharma, V. Rana and V. Kumar, Deep Learning based Semantic Personalized Recommendation System, *International Journal of Information Management Data Insights* **1**(2) (2021), 100028.

[12] L. Pang, Intelligent Big Information Retrieval of Smart Library Based on Graph Neural Network (GNN) Algorithm, *Computational Intelligence and Neuroscience* **2022** (2022), 1475069.

[13] H. Cui, J. Lu, Y. Ge and C. Yang, How Can Graph Neural Networks Help Document Retrieval: A Case Study on CORD19 with Concept Map Generation, in: *ECIR*, 2022, pp. 75–83.

[14] Y. Zhang, J. Zhang, Z. Cui, S. Wu and L. Wang, A Graph-based Relevance Matching Model for Ad-hoc Retrieval, *AAAI* (2021), 4688–4696.

[15] O. Shchur and S. Günnemann, Overlapping Community Detection with Graph Neural Networks, in: *DLG-KDD*, 2019.

[16] C. Wang, S. Pan, R. Hu, G. Long, J. Jiang and C. Zhang, Attributed Graph Clustering: A Deep Attentional Embedding Approach, in: *IJCAI*, 2019, pp. 3670–3676–.

[17] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu and P. Cui, Structural Deep Clustering Network, in: *WWW*, 2020, pp. 1400—1410.

[18] H. Zuo, Y. Yin and P. Childs, Patent-KG: Patent Knowledge Graph Extraction for Engineering Design, *Proceedings of the Design Society* **2** (2022), 821–830.

[19] S. Sarica, B. Song, E. Low and J. Luo, Engineering Knowledge Graph for Keyword Discovery in Patent Search, *ICED* **1**(1) (2019), 2249—2258.

[20] Y. Xiao, C. Li and M. Thürer, A Patent Recommendation Method based on KG Representation Learning, *Engineering Applications of Artificial Intelligence* **126**(A) (2023), 106722.

[21] M.S. Anderson, The Complex Relations between the Academy and Industry, *The Journal of Higher Education* **72**(2) (2001), 226–246. doi:10.1080/00221546.2001.11778879.

[22] C. Grimpe and K. Hussinger, Formal and Informal Knowledge and Technology Transfer from Academia to Industry: Complementarity Effects and Innovation Performance, *Industry and Innovation* **20**(8) (2013), 683–700. doi:10.1080/13662716.2013.856620.

[23] M.-H. Huang, H.-W. Yang and D.-Z. Chen, Industry–academia collaboration in fuel cells: A perspective from paper and patent analysis, *Scientometrics* **105** (2015), 1301–1318.

[24] V. Larivière, B. Macaluso, P. Mongeon, K. Siler and C.R. Sugimoto, Vanishing industries and the rising monopoly of universities in published research, *PloS one* **13**(8) (2018), e0202120.

[25] M. Bikard, K. Vakili and F. Teodoridis, When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity, *Organization Science* **30**(2) (2019), 426–445.

[26] A. Salatino, F. Osborne and E. Motta, Researchflow: Understanding the knowledge flow between academia and industry, in: *EKAW*, Springer, 2020, pp. 219–236.

[27] C. McManus, A.A. Baeta Neves and A.T. Prata, Scientific publications from non-academic sectors and their impact, *Scientometrics* **126**(11) (2021), 8887–8911.

[28] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher and T. Eliassi-Rad, Collective classification in network data, *AI magazine* **29**(3) (2008), 93–93.

[29] S. Angioni, A. Salatino, F. Osborne, D.R. Recupero and E. Motta, AIDA: A knowledge graph about research dynamics in academia and industry, *Quantitative Science Studies* **2**(4) (2021), 1356–1398. doi:10.1162/qss$_{a0}$0162.

[30] M.F. Sy, B. Roman, S. Kerrien, D.M. Mendez, H. Genet, W. Wajerowicz, M. Dupont, I. Lavriushev, J. Machon, K. Pirman et al., Blue Brain Nexus: An open, secure, scalable system for knowledge graph management and data-driven science, *Semantic Web* **14**(4) (2023), 697–727.

[31] C. Chen, I.A. Ebeid, Y. Bu and Y. Ding, Coronavirus Knowledge Graph: A Case Study, *arXiv preprint arXiv:2007.10287* (2020).

[32] K.M. Malik, M. Krishnamurthy, M. Alobaidi, M. Hussain, F. Alam and G. Malik, Automated Domain-Specific Healthcare Knowledge Graph Curation Framework: Subarachnoid Hemorrhage as Phenotype, *Expert Systems with Applications* **145** (2020), 113120.
[33] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang and D. Lee, DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation, in: *KDD*, 2020, pp. 492—502.
[34] P. Chen, Y. Lu, V.W. Zheng, X. Chen and B. Yang, KnowEdu: A System to Construct Knowledge Graph for Education, *IEEE Access* **6** (2018), 31553–31563.
[35] T. Zhao, C. Chai, Y. Luo, J. Feng, Y. Huang, S. Yang, H. Yuan, H. Li, K. Li, F. Zhu et al., Towards automatic mathematical exercise solving, *Data Science and Engineering* **4** (2019), 179–192.
[36] Y. Jia, Y. Qi, H. Shang, R. Jiang and A. Li, A Practical Approach to Constructing a Knowledge Graph for Cybersecurity, *Engineering* **4**(1) (2018), 53–60, Cybersecurity.
[37] E. Aumayr, M. Wang and A.-M. Bosneag, Probabilistic Knowledge-Graph based Workflow Recommender for Network Management Automation, in: *WoWMoM*, 2019, pp. 1–7.
[38] K. Krinkin, A. Vodyaho, I. Kulikov and N. Zhukova, Models of Telecommunications Network Monitoring Based on Knowledge Graphs, in: *MECO*, 2020, pp. 1–7.
[39] M. Färber, The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data, in: *ISWC*, 2019, pp. 113–129.
[40] B. Yaman, M. Pasin and M. Freudenberg, Interlinking SciGraph and DBpedia Datasets Using Link Discovery and Named Entity Recognition Techniques, in: *LDK*, Vol. 70, 2019, pp. 15:1–15:8.
[41] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen and P. Principe, The OpenAIRE Research Graph Data Model, Zenodo, 2019.
[42] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: *NIPS*, Vol. 26, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds, Curran Associates, Inc., 2013.
[43] Y. Lin, Z. Liu, M. Sun, Y. Liu and X. Zhu, Learning Entity and Relation Embeddings for Knowledge Graph Completion, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, AAAI Press, 2015, pp. 2181–2187–. ISBN 0262511290.
[44] Z. Wang, J. Zhang, J. Feng and Z. Chen, Knowledge Graph Embedding by Translating on Hyperplanes, *Proceedings of the AAAI Conference on Artificial Intelligence* **28**(1) (2014).
[45] S. He, K. Liu, G. Ji and J. Zhao, Learning to Represent Knowledge Graphs with Gaussian Embedding, in: *CIKM*, CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 623–632–. ISBN 9781450337946.
[46] H. Xiao, M. Huang and X. Zhu, From One Point to a Manifold: Knowledge Graph Embedding for Precise Link Prediction, in: *IJCAI*, S. Kambhampati, ed., IJCAI/AAAI Press, 2016, pp. 1315–1321.
[47] S. Wu, F. Sun, W. Zhang, X. Xie and B. Cui, Graph Neural Networks in Recommender Systems: A Survey, *ACM Computing Surveys* **55**(5) (2022), 97.
[48] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and P. Vandergheynst, Geometric Deep Learning: Going beyond Euclidean data, *IEEE Signal Processing Magazine* **34**(4) (2017), 18–42.
[49] J. Bruna, W. Zaremba, A. Szlam and Y. Lecun, Spectral Networks and Locally Connected Networks on Graphs, in: *ICLR*, 2014.
[50] M. Defferrard, X. Bresson and P. Vandergheynst, Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, in: *NIPS*, 2016, pp. 3844—3852.
[51] T.N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in: *ICLR*, 2017.
[52] W. Hamilton, Z. Ying and J. Leskovec, Inductive Representation Learning on Large Graphs, in: *NIPS*, Vol. 30, 2017, pp. 1024–1034.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser and I. Polosukhin, Attention is All you Need, in: *NIPS*, Vol. 30, 2017, pp. 5998–6008.
[54] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, Graph Attention Networks, in: *ICLR*, 2018.
[55] J. Zhang, X. Shi, J. Xie, H. Ma, I. King and D.-Y. Yeung, GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs, in: *UAI*, 2018, pp. 339–349.
[56] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui and P.S. Yu, Heterogeneous Graph Attention Network, in: *WWW*, 2019, pp. 2022–2032–.
[57] Z. Hu, Y. Dong, K. Wang and Y. Sun, Heterogeneous Graph Transformer, in: *WWW*, 2020, pp. 2704—2710.
[58] J.-W. Seol, S.-H. Lee and K.-Y. Kim, Author Disambiguation Using Co-Author Network and Supervised Learning Approach in Scholarly Data, *International Journal of Software Engineering and Its Applications* **10** (2016), 73–82.
[59] J.W. Ratcliff, D. Metzener et al., Pattern Matching: The Gestalt Approach, *Dr. Dobb's Journal* **13**(7) (1988), 46.
[60] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *NAACL-HLT*, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, 2019, pp. 4171–4186.
[61] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language Models are Few-Shot Learners, in: *NeurIPS*, Vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, eds, Curran Associates, Inc., 2020, pp. 1877–1901.
[62] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al., Llama: Open and efficient foundation language models, *arXiv preprint arXiv:2302.13971* (2023).
[63] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, Unifying Large Language Models and Knowledge Graphs: A Roadmap, *ArXiv* **abs/2306.08302** (2023).
[64] R. Martínez-Cruz, A.J. López-López and J. Portela, ChatGPT vs State-of-the-Art Models: A Benchmarking Study in Keyphrase Generation Task, *arXiv preprint arXiv:2304.14177* (2023).

[65] N. Reimers and I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: *EMNLP/IJCNLP*, 2019, pp. 3980–3990.
[66] N. Reimers and I. Gurevych, Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, in: *EMNLP*, 2020, pp. 4512–4525.
[67] A. Dridi, M.M. Gaber, R.M.A. Azad and J. Bhogal, Scholarly Data Mining: A Systematic Review of Its Applications, *WIREs Data Mining and Knowledge Discovery* **11**(2) (2021), e1395.
[68] M. Fey and J.E. Lenssen, Fast Graph Representation Learning with PyTorch Geometric, in: *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
[69] Y. Dong, N.V. Chawla and A. Swami, Metapath2vec: Scalable Representation Learning for Heterogeneous Networks, in: *KDD*, 2017, pp. 135—144.
[70] D. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in: *ICLR*, 2015.
[71] L. van der Maaten and G. Hinton, Visualizing Data using t-SNE, *Journal of Machine Learning Research* **9**(86) (2008), 2579–2605.
[72] S. Dong, P. Wang and K. Abbas, A Survey on Deep Learning and Its Applications, *Computer Science Review* **40**(C) (2021), 100379.
[73] A. Krizhevsky, I. Sutskever and G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: *NIPS*, 2012, pp. 1106–1114.
[74] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M.P. Reyes, M.-L. Shyu, S.-C. Chen and S.S. Iyengar, A Survey on Deep Learning: Algorithms, Techniques, and Applications, *ACM Computing Surveys* **51**(5) (2018), 92.
[75] S. Verma, R. Bhatia, S. Harit and S. Batish, Scholarly Knowledge Graphs through Structuring Scholarly Communication: A Review, *Complex & Intelligent Systems* **9**(1) (2023), 1059–1095.
[76] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P.S. Yu, A Comprehensive Survey on Graph Neural Networks, *IEEE Transactions on Neural Networks and Learning Systems* **32**(1) (2021), 4–24.
[77] Y. Zhou, H. Zheng, X. Huang, S. Hao, D. Li and J. Zhao, Graph Neural Networks: Taxonomy, Advances, and Trends, *ACM Transactions on Intelligent Systems and Technology* **13**(1) (2022), 15.
[78] B. Abu-Salih, Domain-specific Knowledge Graphs: A Survey, *Journal of Network and Computer Applications* **185** (2021), 103076.
[79] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, in: *ICLR*, 2015.