Lawrence Berkeley National Laboratory
LBL Publications

Title
Learning Global Proliferation Expertise Evolution Using AI-Driven Analytics and Public Information

Permalink
https://escholarship.org/uc/item/5sn2r9cr

Journal
IEEE Transactions on Nuclear Science, 69(6)

ISSN
0018-9499

Authors
Glenski, Maria
Ayton, Ellyn
Soni, Sannisth
et al.

Publication Date
2022-06-01

DOI
10.1109/tns.2022.3162216

Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, availalbe at https://creativecommons.org/licenses/by-nc/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Learning Global Proliferation Expertise Evolution Using AI-Driven Analytics and Public Information

Maria Glenski, Ellyn Ayton, Sannisth Soni, Emily Saldanha, Dustin Arendt, Brian Quiter, Ren Cooper, *Member, IEEE*, and Svitlana Volkova

***Abstract*— Detecting and anticipating global proliferation expertise and capability evolution from unstructured, noisy, and incomplete public data streams is a highly desired, but extremely challenging task. In this article, we present our pioneering data-driven approach to support the non-proliferation mission to detect and explain the evolution of proliferation expertise and capability development globally from terabytes of publicly available information (PAI), focusing on our knowledge extraction pipeline and descriptive analytics. We first discuss how we fuse nine open-source data streams, including multilingual data, to convert 4 TB of unstructured data to structured knowledge and encode dynamically evolving proliferation expertise representations—content and context graphs. For this, we rely on natural language processing (NLP) and deep learning (DL) models to perform information extraction, topic modeling, and distributed text representation (aka embedding) learning. We then present interactive, usable, and explainable descriptive analytics to refine domain knowledge and present it in a human-understandable form. Finally, we introduce future work avenues that will leverage our dynamic knowledge representations and descriptive analytics to enable predictive and prescriptive inferences to achieve real-time domain understanding and contextual reasoning about global proliferation expertise and capability evolution.**

***Index Terms*— Artificial neural networks, big data applications, data mining, data visualization, decision support systems, knowledge discovery, knowledge representations, machine learning, natural language processing (NLP), prediction models.**

I. Motivation

Open source data analytics have been shown to have a tremendous success and impact across a variety of applications that support national security missions ranging from cognitive security (e.g., detecting influence operations in the information environment [1]–[3]) to biosecurity (e.g., infectious disease modeling [4], [5]), cybersecurity (e.g., threat hunting), and nuclear security and safeguards [6]–[9]. This work demonstrates how recent advances in data science and artificial intelligence (AI), such as machine learning, deep learning (DL), and natural language processing (NLP) in combination with large amounts of publicly available information (PAI), advance U.S. nuclear security and impact the nonproliferation mission to reduce the risk of catastrophic consequences from the use of a nuclear weapon.

Terabytes of PAI combined with AI-driven decision intelligence and analytics offer an excellent potential capability to detect, anticipate, and reason about global nuclear proliferation expertise and capability development [8], [10]. The ability to identify emerging proliferation expertise and technologies from open-source data streams to monitor its evolution and forecast potential proliferation risks, as well as unexpected shifts in research interests, are critical to support the nonproliferation mission. The U.S. withdrawal from the Joint Comprehensive Plan of Action (JCPOA)[1] and Iran's announcement of noncompliance with the agreement[2] are likely leading to diminished U.S. insights into Iranian nuclear expertise and capability development. Early research via GoogleScholar[3] indicates that Iranian scientists and subject matter experts (SMEs) are still actively publishing and developing domestic capabilities, sometimes in collaboration with chemical and material scientists. This is only one example of potential proliferation activity that may require additional open-source insights; it is critical to monitor the evolution of the nuclear expertise and technology development network globally to anticipate changes that may require a rapid response.

The existing efforts primarily focused on proliferation expertise detection in bibliometric data in English using co-citation network analysis (i.e., knowledge generation process via collaborative efforts), ignoring emerging scientific content (i.e., domain knowledge itself) [11]–[14]. In comparison, our approach (outlined in Fig. 1) considers both the knowledge generation process and the knowledge that is generated. Our approach fuses nine multilingual, heterogeneous open-source data streams, for example, academic publications,

This work was supported in part by the U.S. Department of Energy (DOE) National Nuclear Security Administration (NNSA) Office of Defense Nuclear Nonproliferation Research and Development (DNN R&D) Next-Generation Artificial Intelligence (AI) Research Portfolio and in part by the Pacific Northwest National Laboratory operated by the Battelle Memorial Institute for the U.S. DOE under Contract DE-AC05-76RLO1830.

Maria Glenski, Ellyn Ayton, Sannisth Soni, Emily Saldanha, Dustin Arendt, and Svitlana Volkova are with the Pacific Northwest National Lab-oratory, Richland, WA 99352 USA (e-mail: maria.glenski@pnnl.gov; ellyn.ayton@pnnl.gov; sannisthamitkumar.soni@pnnl.gov; emily.saldanha@pnnl.gov; dustin.arendt@pnnl.gov; svitlana.volkova@pnnl.gov).

Brian Quiter and Ren Cooper are with the Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA (e-mail: bjquiter@lbl.gov; rjcooper@lbl.gov).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNS.2022.3162216.

[1] https://www.whitehouse.gov/briefings-statements/president-donald-j-trump-ending-united-states-participation-unacceptable-iran-deal/

[2] https://www.bbc.com/news/world-middle-east-51001167

[3] https://scholar.google.com/

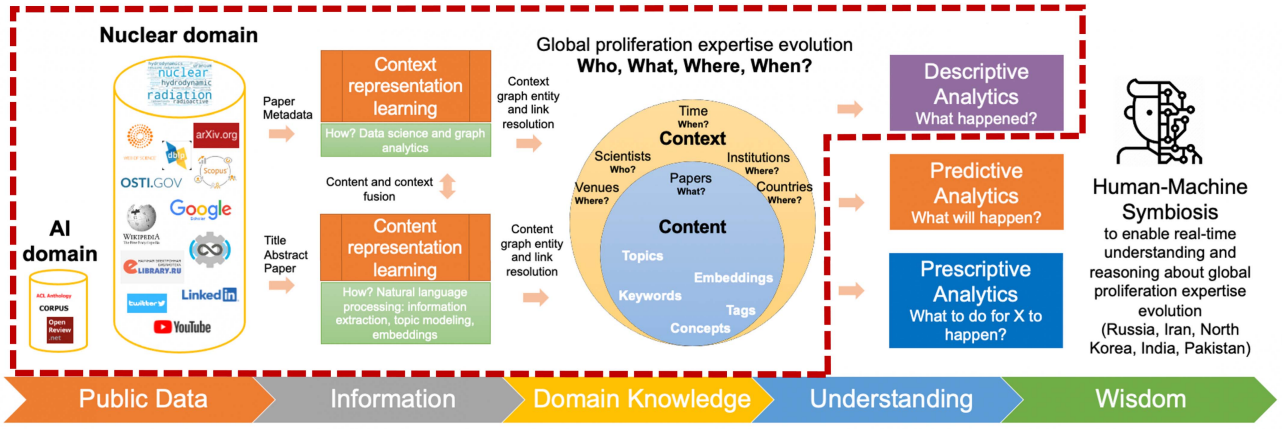

Fig. 1. Overview of the AI-enabled technologies to detect, anticipate, and reason about global proliferation expertise and capability evolution from unstructured dynamic multilingual real-world data, highlighting the data fusion and structure knowledge representation construction pipelines that are the focus of the current work.

and converts 4 TB of unstructured data (both scientific knowledge—content—and knowledge generation processes—context) into domain knowledge representations.

Our core contributions are an overview of our novel multilingual data collection, enrichment, fusion pipeline, and our process to extract structured knowledge representations over time using our dynamic context (who, where, when, with whom) and content (what) graphs. We also highlight interactive analytics that, using our dynamic context and content graphs, enable descriptive analytics to answer questions like *What are the existing capabilities and expertise for a country?*

In the discussion section, we highlight our avenues of future work which will leverage these dynamically evolving structured proliferation expertise and capability representations to enable predictive modeling [5], [15] and counterfactual reasoning [16], [17] to answer operational questions, including those shown below.

- *Predictive*: What are the next likely capabilities to be acquired by a country?
- *Prescriptive*: What should a country do to acquire a specific capability?
- *Counterfactual reasoning*: Could a country have acquired a capability if an alternate event had happened?

II. Open-Source Data Collection and Fusion

To construct dynamically evolving proliferation expertise knowledge graphs, we fuse nine multilingual, heterogeneous open-source data streams, for example, academic literature. Although there is a wealth of PAI, we leverage domain expertise to address the challenge of identifying nuclear-related information and relevant signals of proliferation expertise. We focus on two approaches to identify signals related to varying countries: the language in which scientific publications were written (or in which nuclear domain keywords are translated), which can be associated with international locations, and the locations linked to scientists, publications, and venues.

A. Cross-Lingual Data Collection

To ensure multilingual coverage, we constructed queries using 638 nuclear terms in English, Russian, Korean, and Arabic using a combination of domain knowledge from nuclear SMEs and nuclear resources–IAEA Safety Glossary.[4] First, we combined a set of English SME-identified nuclear terms and search queries with terminology extracted from the IAEA Safety Glossary, a document produced by the International Atomic Energy Agency that defines and explains technical terms related to or used in IAEA publications including safety standards. After identifying this set of English terms and queries, we translated the terms and queries into three other languages (Russian, Korean, and Arabic) to increase the coverage across these languages.

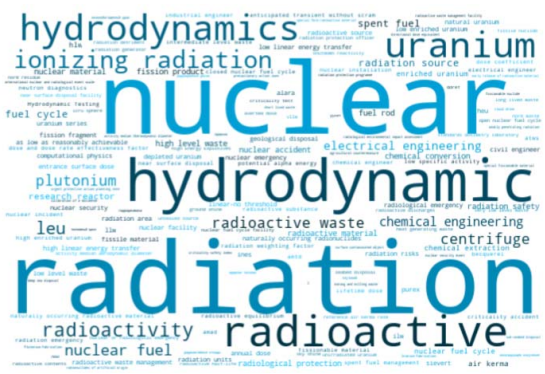

Fig. 2. Summary of the relative frequency of nuclear terms in our data collection. Size indicates the frequency of the query term's presence in the scientific publications collected to date.

Using these 638 multilingual terms, we collected relevant data from nine PAI sources. Fig. 2 illustrates the relative frequency of each term, with the query terms "nuclear," "radiation," "hydrodynamic," "radioactive," and "ionizing radiation" among the most frequently occurring. To date, we have processed more than 4 TB of publicly available scientific literature from SCOPUS, the Web of Science, OSTI.gov, arXiv, DBLP, and bioRXiv. In addition, we queried Google Scholar, ArabArXiv, and eLibrary.ru to expand our multilingual data representation.

[4]2018 edition of the IAEA Safety Glossary, collected from: https://www-pub.iaea.org/MTCD/Publications/PDF/PUB1830_web.pdf

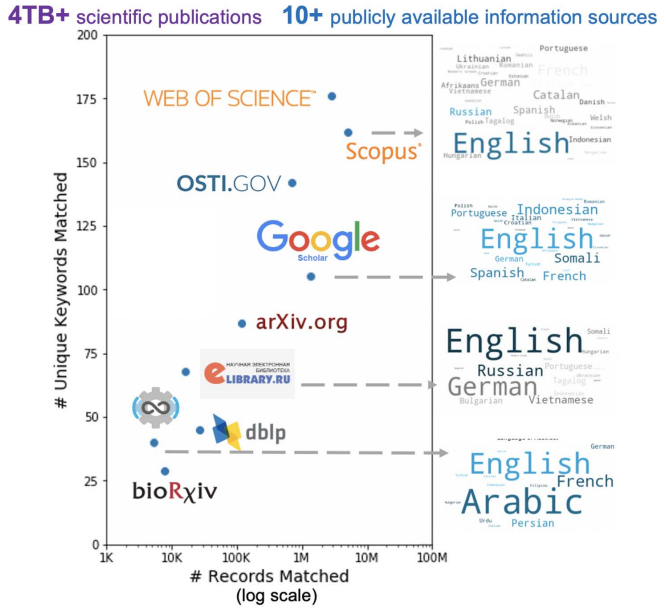

Fig. 3. Data summary plotting unique nuclear terms matched as a function of the number of records matched using multilingual queries for each data source, with wordclouds illustrating the prevalence of each language to the right.

We summarize the scale of nuclear-related data collected from each source and the representation of languages in the scientific publications collected in Fig. 3. Although we see that English is one of the top two most frequent languages for all data sources, the wordclouds illustrating the relative presence of all languages in data collected from each source illustrates the multilingual coverage—including Arabic, Russian, and Korean—across all data sources.

B. Global Coverage of Content and Context Information

Alongside examining the language of the content itself, we also evaluated the global worldwide representation of the locations referenced in the publication metadata or context to evaluate the global coverage. This includes locations linked to publications with the institutions' authors are associated with, or the locations of conferences or publication venues.

In Fig. 4, we illustrate the complementary nature of the global coverage identified during our location-based analyses across the data sources. While Scopus (top) has global coverage, the highest concentration is in the United States. In contrast, Google Scholar data (bottom) shows global coverage, with the highest concentration in India and China. As our data fusion approach leverages multiple PAI sources, it is able to take advantage of this diversity in coverage for a more complete global representation overall. This enables our descriptive analytics to support global proliferation expertise identification and summarization.

III. Topic Modeling and Representation Learning

After data collection and fusion, we employ topic modeling to summarize and filter signals about potential nuclear proliferation activities. Using rich probabilistic topics and embedding representations allows us to extract insights about key topics—more abstract summarizations of content and proliferation

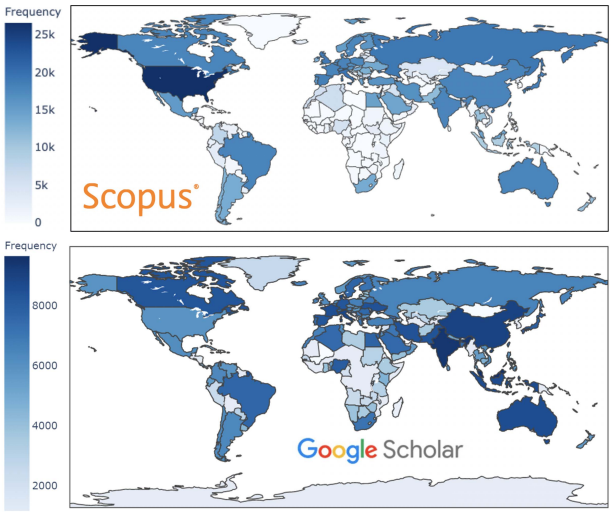

Fig. 4. Global coverage of data collection illustrated using Scopus (above) and Google Scholar (below) publications identified using the 638 multilingual nuclear queries.

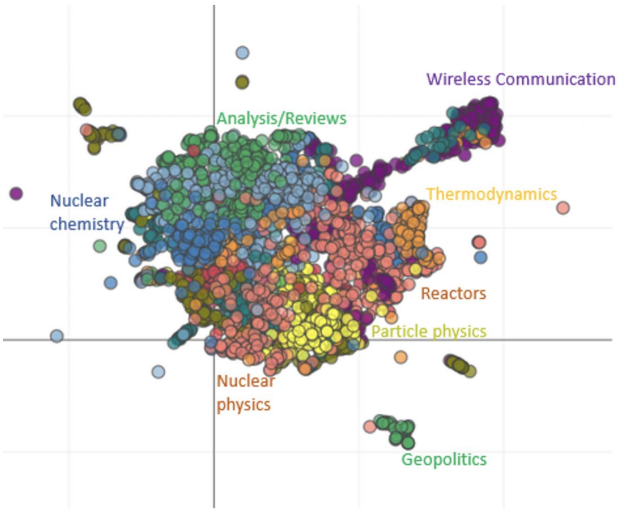

Fig. 5. Illustration of a 2-D UMAP representation of the top2vec embeddings generated from the OSTI, Scopus, and Web of Science datasets. Clusters are annotated by topics identified (e.g., nuclear chemistry, thermodynamics, and reactors).

signals than query terms—and increase the precision of the data collected.

We use Top2Vec [18], an algorithm developed for topic modeling and semantic search that is able to jointly embed topics, documents, and word vectors when given a corpus of documents. In contrast to other topic modeling approaches, Top2Vec is able to automatically detect the number of topics present, which reduces the number of parameters to tune (or identify *a priori*). In Fig. 5, we illustrate an example of how this approach allows us to explore the publications (i.e., abstracts) in the OSTI, Scopus, and Web of Science datasets relative to the topics covered in the publication content. Using these topics, we can both summarize the data and identify unrelated topics, for example, medical imaging and so on, which can be used to increase the precision (focus on nuclear-related documents).

TABLE I

IMPACT OF TOPIC MODEL-BASED FILTERING WITH DOMAIN EXPERT FEEDBACK TO REMOVE UNRELATED DATA (I.E., FALSE-POSITIVES)

	# Papers		
Data Source	Before	After	% Reduced
Scopus	499K	193K	-61%
WoS	531K	180K	-66%
OSTI	184K	108K	-41%
Google Scholar	700K	150K	-78%

Using domain knowledge and SME feedback (aka as the human-in-the-loop approach), we remove unrelated publications (included as a result of noisy matches to our queries). We identified several identifiable non-nuclear topics (e.g., biological applications) which allowed us to filter several unrelated publications. Intuitively, all of the bioRxiv publications were identified as noisy matches to our queries resulting from non-nuclear biological applications. Therefore, we removed bioRxiv publications from consideration. This filtering approach allows us to significantly reduce the noise in the collections from the four largest data sources, as illustrated in Table I. We see that the removal of identifiable irrelevant documents reduces the size of each collection by 41%–78%, which would have otherwise contributed noise to subsequent descriptive, predictive, or prescriptive analytics.

IV. CROSS-LINGUAL TEXT ENRICHMENTS USING NLP

We leverage several NLP enrichments to enhance text representations in our dataset, for example, abstracts, documents, and so on. We incorporate both structured information and distributed representations—word or document embedding vectors—that encode the semantic meaning of the information within each document and across the documents. We use AllenNLP [19] and spaCy[5] models to perform linguistic annotation of English text. Each token within a document is annotated with the following:

- a list of all other references to the token via *co-reference resolution* [20];
- grammatical relationships between the token and any modifiers via *syntactic parsing*;
- a dictionary denoting the token's placement in the latent predicate argument structure of the sentence through *semantic role labeling (SRL)* [21];
- any *noun phrase* involving the token (*noun phrase tagging*); and
- *part of speech (POS)* tagging for the token.

For each document, in addition to the NLP enrichments, we learn the word and document embedding vectors based on the state-of-the-art BERT model [23] from the HuggingFace *transformers* library.[6] This provides a semantic representation of the information within each document, which can be leveraged for downstream analytics using similarity measures to other documents, queries, or clustering.

Furthermore, we extract relevant scientific entities and relations from each document using the SpERT model [22], which

[5]https://spacy.io/

[6]https://huggingface.co/transformers/model_doc/bert.html

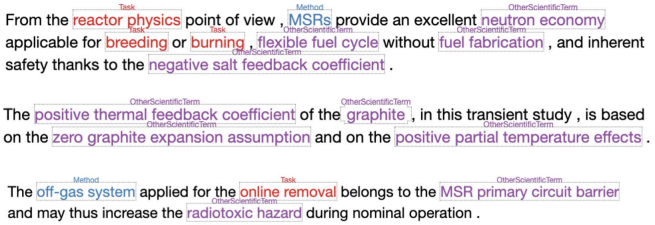

Fig. 6. Example entity and relation extraction applied to the academic papers using linguistic annotations and the pre-trained model from [22].

is trained on the SciERC [24] dataset. Extracting entities and relationships in combination with linguistic enrichments enables us to construct knowledge graph representations for each document. An example of the entities and relations extracted from a nuclear-related academic paper is illustrated in Fig. 6. For all non-English documents, we apply models from the Stanza NLP toolkit [25] to perform similar linguistic annotations-syntactic parsing, noun phrase, and part-of-speech tagging. We use these annotations to provide rich representations for the downstream analytics. Additional multilingual NLP resources support annotations (e.g., SRL, co-reference) for relevant non-English languages (e.g., Arabic) [26]–[29].

V. GRAPH CONSTRUCTION

We construct content and context graph representations that will enable the end-users with search capabilities to answer questions relevant to global proliferation expertise evolution. Content graphs consist of nuclear domain keywords, tags associated with the academic papers, and concepts, combined with rich probabilistic topics and embedding representations. Context graphs encode the relationships between scientists, venues, institutions, countries, and papers over time.

A. Content Graphs to Encode Domain Knowledge

Our NLP-driven approach allows us to encode the relationships between key concepts described in the abstracts and full text of the nuclear publications at two scales—global and local. Global content graphs encode the relationships of concepts across all publications, whereas local content graphs illustrate the relationships of concepts within a specific document (e.g., a scientific publication or technical report), enabling us to summarize general connections (*global*) and precise applications (*local*) for topics, methodology, and concepts.

First, we construct the *global content graph*, which is the structure of knowledge and expertise that is revealed by the co-occurrence of concepts across our full corpus of papers and abstracts. To do so, we leverage the observed relationships between context (scientific publications) and content (key concepts), such as the occurrence of entities, topics, keywords, and tags within the papers and abstracts or within the descriptions of conferences and journals. We construct a bipartite graph connecting concept nodes to documents, where documents are papers, conference descriptions, and journal descriptions and project this set of bipartite relationships to create a concept-to-concept projection (see Fig. 7). Using this projection, we can observe which concepts are closely related to each other by observing whether they tend to co-occur in the

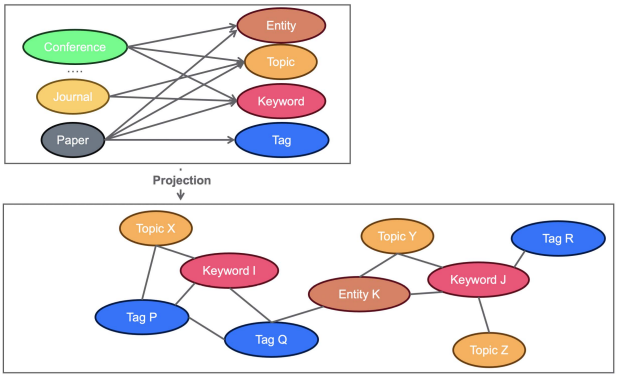

Fig. 7. Illustration of the process projecting a bipartite graph encoding the co-occurrence of concepts (entities, topics, keywords, and tags) in documents (papers, and conference or journal descriptions) to construct the global content graph.

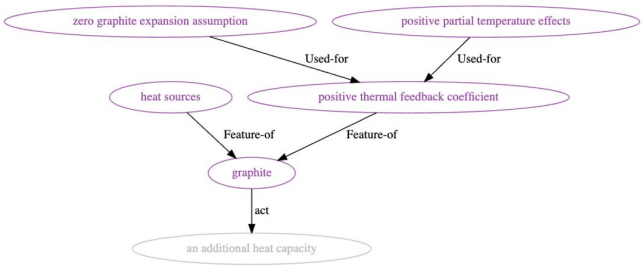

Fig. 8. Local content graph example, illustrating the knowledge extracted from the document in Fig. 6.

same documents. This reveals the underlying structure behind the concepts that we use to summarize expertise patterns.

While the global content graph reveals the general relationships that exist between key concepts, we are able to extract a much richer and deeper semantic understanding of the relationships between concepts by using the actual text of the abstracts and papers. We do so through our local content graph construction, where we aim to leverage the language used to relate concepts within a specific paper or abstract.

To construct a *local content graph* for each document, we combine two techniques for automatic concept and relationship extraction introduced in Section IV. The first is the SRL model which parses the sentences in the text and assigns words and phrases to semantic roles, such as the agent, action, and the object of the action. This allows us to relate pairs of concepts through the verbs that define their relationships, for example, "*graphite*" and "*heat capacity*" are related through the verb "*act*." The second uses the SpERT model trained to extract entities such as tasks, methods, and materials and how they are related such as "Feature-Of" and "Used-for" relationships. By combining the extracted entities and relationships from both methods, we can construct a local graph representation of the key concepts. Finally, to resolve nodes returned by these two approaches and preserve edges to and from these nodes, we use the Needleman–Wunsch algorithm [30]. In Fig. 8, we illustrate a sample of the local graph constructed from the same nuclear text used in Fig. 6.

B. Context Graphs to Encode Relationships Between Entities

Once we identified and pre-processed our corpus of nuclear-related scientific publications, we construct context

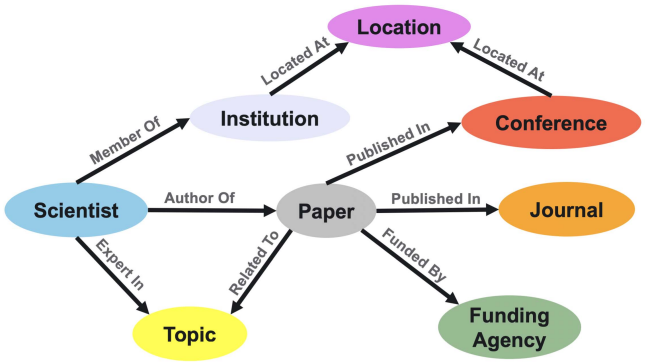

Fig. 9. Summary of the node types and the relationships in the context graph representations, extracted from the metadata and content of scientific publications.

TABLE II
SUMMARY OF THE CONTEXT GRAPHS EXTRACTED FROM PUBLICATIONS COLLECTED FROM EACH OF THE CORE DATASETS

	WOS	Scopus	OSTI	Google Scholar	Arxiv	DBLP
# Nodes	7.3M	4.8M	730K	793.7K	164.6K	9.8K
# Edges	22.5M	1M	4.1M	101.9K	354.5K	10.9K

graphs from both content and metadata. These context graphs represent how different papers, scientists, topics, locations, journals, conferences, and funding agencies are connected to each other in the scientific publication (aka scientific knowledge generation) space. We use the resulting heterogeneous graphs and relevant projections (e.g., Scientist-to-Scientist collaboration networks) to support identification, monitoring, and reasoning about global proliferation expertise evolution.

In Fig. 9, we illustrate the various relationships and node types that our approach extracts from scientific publication data. The resulting context graphs encode *what* is published ("Paper" and "Topic" nodes and "Related To" links), *who* funds the research ("Funding Agency" nodes), *who* performed the research ("Scientist" nodes, "Author Of" links), *where* scientists publish ("Conference" and "Journal" nodes, "Published In" links), which "*Institutions*" scientists are affiliated with, and *where* those institutions and conferences are located.

In Table II, we present the size of each data source's resulting context graphs in regard to the number of nodes (entities) and edges (relationships between entities), and the distribution across node and edge types in Fig. 10. Of the core datasets, Web of Science and Scopus have the largest context graphs, with millions of entities and relationships represented. To derive actionable insights, we can focus on subsets of interest using entity or relationship-based queries. For example, Fig. 11 illustrates a sample of the context graph extracted from the Scopus data source, focusing on papers linked to Russian query terms.

C. Entity Resolution

Real-world scientific publication data is noisy [31]. Even within the same dataset, ambiguous entities are often present, for example, different spellings of the same author's name or two different authors having the same name [32]–[34]. It is necessary to address the issue of entity resolution to

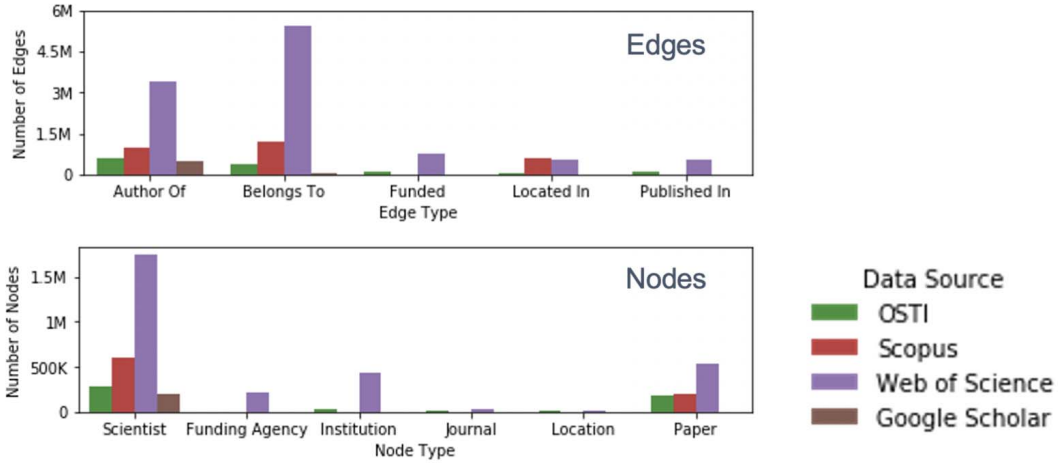

Fig. 10. Summary of the distribution across node (entity) types and edge (relationships between entities) types for the four largest data sources.

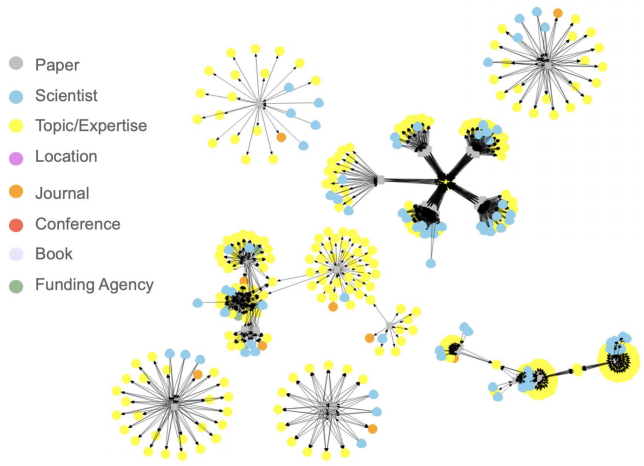

Fig. 11. Sample of the context graph extracted from the Scopus data source, focused on papers linked to Russian query terms.

increase the reliability of analytics run on and the quality of knowledge extracted from publication records. A human in the loop approach is necessary to address edge cases where simple rule-based approaches cannot be used with confidence, for example, common names and incomplete formatting, such as "J. Jones," that requires additional context (email address, institution, coauthors, etc.) to resolve.

The solution we developed for entity resolution in both content and context graphs relies on a combination of the graph- and language-based heuristics to address this challenge at scale but allows a user (i.e., analyst) to quickly validate the proposed results and fix any issues. First, a graph is formed connecting entities (e.g., scientists) whose edges are weighted by a combination of character-level distance metrics (text similarity of names using Levenshtein distance, similarity of substrings within names that minimize Levenshtein distance, word-level tfidf cosine similarity) and bibliographic distance metrics (similarity of projection graphs that represent shared coauthors, institutional affiliations). We also incorporate likelihoods of how frequently names occur in the dataset overall, to down-weight the similarity of names that are most likely to be common names (e.g., "Jones") rather than multiple representations of the same person. Then, clusters are found by running community detection (greedy modularity optimization [35], [36]) on each connected component of the graph.

Within each cluster, we compute the minimum spanning tree. If the user validates all edges in the spanning tree of a cluster, then using the transitive property, we can assume that all nodes in the cluster are equivalent. The spanning tree reduces the number of edges for the user to validate from $O(n^2)$ to $O(n)$.

We developed a Jupyter widget[7] to allow a user to quickly validate edges and clusters (see Fig. 12). The tool presents a potential edge to merge with the user, showing the names of the two nodes on the edge's endpoints. It also presents contextual information including what papers, institutions, and topics overlap (or not). The user can accept or reject individual nodes or entire clusters. Accepting a cluster accepts all edges in that cluster (every node is treated as the same entity), and rejecting a cluster rejects every edge (every node is treated as a different/unique entity). A limitation of the overall approach is that it will not allow nodes to be merged if the spanning tree path includes an edge that the user chooses to reject. To address this, we allow the clustering and merging process to be repeated as necessary after merging occurs.

VI. Descriptive Analytics: Nuclear Expert Knowledge Evolution

In order to understand how nuclear domain knowledge and expertise is evolving over time globally and contrasting these across countries in the context of real-world events, for example, before and after JCPOA, we leverage an in-house developed descriptive analytics tool called Evaluating Spatiotemporal Embeddings (ESTEEM) [37]. To do so, we fine-tuned BERT [23] models to learn country- and time-specific word embeddings for India, Iran, North Korea, Pakistan,

[7]https://jupyter.org/widgets

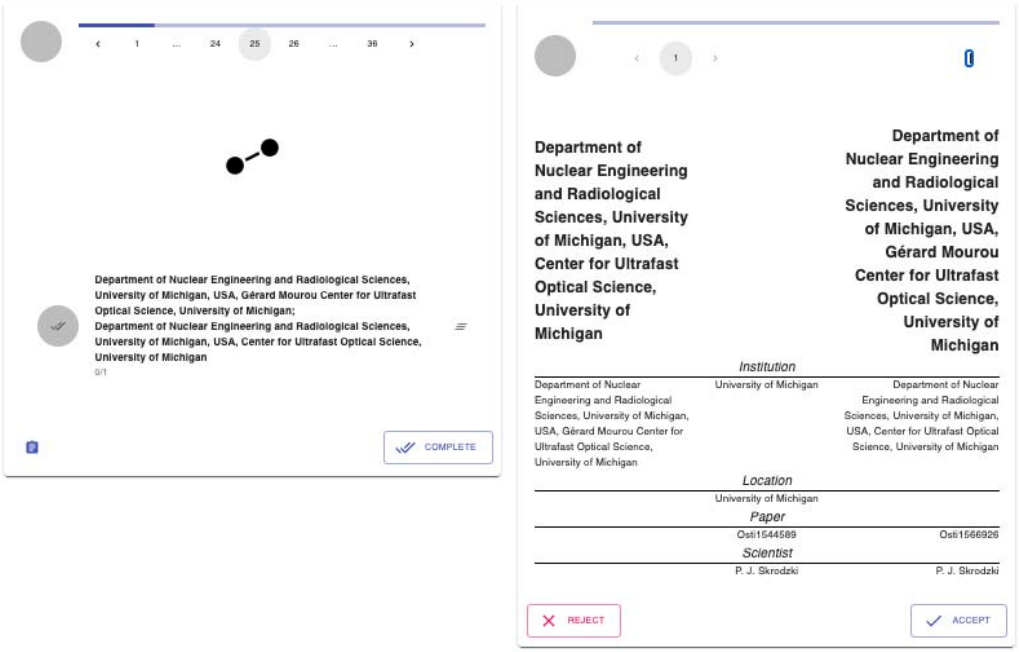

Fig. 12. Entity resolution tool used to disambiguate entities in context and content graphs (e.g., scientists, institutions, and venues) using an iterative, human-in-the-loop approach supported by automated text and graph similarity-based clustering techniques.

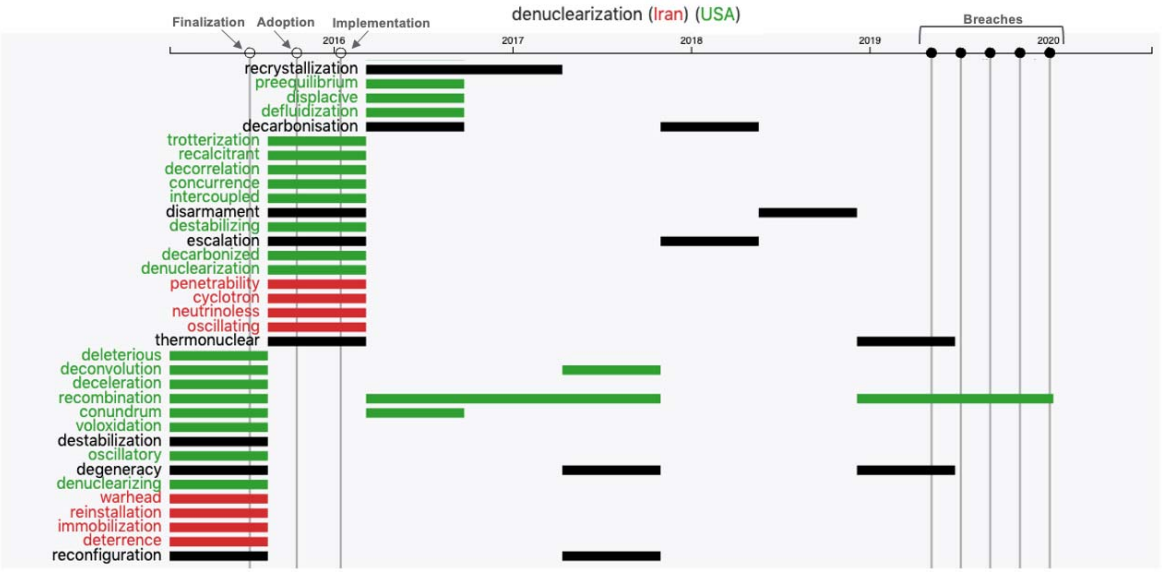

Fig. 13. ESTEEM tool illustrating the search query "denuclearization," comparing USA and Iran in a single, combined view.

Russia, and the U.S. per year. We summarize the volume of data used to fine-tune these models in Table III.

ESTEEM allows us to rapidly explore how the semantics (i.e., associated meaning) of different queries of interest (e.g., "*denuclearization*" in Fig. 13) varies depending on the country and year in which the term is used. We can analyze countries independently in separate visualizations or jointly in a combined plot (as shown in Fig. 13). To replicate the analyses supported by ESTEEM with a manual human approach would require not only significant time and cognitive effort to read thousands of scientific papers, but also domain expertise to identify and summarize semantic associations, which would need to be replicated for all additional queries.

When we examine the use of "*denuclearization*," we can identify shared context and country-specific context over time in a combined plot. Terms highlighted in black reflect nearest neighbors in the embedding space of the term "*denuclearization*" (terms used in the same context) for both U.S.-linked scientific publications and Iran-linked scientific publications, while green terms are U.S.-specific and red terms are Iran-specific. We have also highlighted key real-world events related to the establishment of the JCPOA (finalization in

TABLE III

SUMMARY OF THE NUMBER OF PUBLICATIONS LINKED TO EACH COUNTRY OF INTEREST USED TO FINE-TUNE A COUNTRY-SPECIFIC BERT EMBEDDING MODEL PER YEAR FOR THE ESTEEM ANALYSES

	# Papers per Year						
Country	2015	2016	2017	2018	2019	2020	2021
India	6,719	7,097	7,604	7,752	7,540	4,816	45
Iran	1,620	1,906	2,011	2,060	2,163	1,449	17
North Korea	29	21	31	29	44	17	1
Pakistan	753	1,029	1,280	1,314	1,367	998	11
Russia	7,625	8,469	8,355	8,855	9,223	5,088	32
USA	14,208	17,233	14,739	14,269	13,992	9,832	85

July 2015, adoption in October 2015, and implementation in January 2016) and the five breaches in 2019 and early 2020. There are several shared associations with "*denuclearization*" between the two countries (e.g., "*destabilization*," "*degeneracy*," "*reconfiguration*") and several distinctions between the two, including a focus on weaponization associations in Iran-linked publications around the time of adoption, that do not persist after the JCPOA is implemented: "*warhead*," "*reinstallation*," "*immobilization*," "*deterrence*."

VII. DISCUSSION

In this article, we demonstrate how PAI in combination with data science and AI can contribute to the reduction of nuclear risks worldwide: analysts will have access to a finer level of details on descriptive analytics for increased situational awareness, leading to a higher level of confidence in insights and analyses at the scale used for real-time informed decision-making. This aligns with previous work highlighting the value of data analytics and open-source information for nuclear security, safeguards, and risk reduction [6], [9], [38]–[41]. We presented our NLP and DL approaches to automatically learn dynamically evolving proliferation expertise representations from open data sources that have extreme volume, velocity, and complexity—terabytes of unstructured scientific publications over the last five years—focusing on six countries of interest—Russia, North Korea, Pakistan, India, Iran, and the United States. Specifically, we fused multiple multilingual open-source data streams and converted unstructured data to information then generated knowledge representations to encode dynamically evolving proliferation expertise with the goal to use these dynamically evolving knowledge representations to enable predictive and prescriptive inferences to achieve real-time global proliferation expertise evolution understanding and contextual reasoning.

A. Predictive and Prescriptive Interactive Analytics

We have illustrated an example of the *descriptive analytics* our nontraditional approach can support using the ESTEEM tool, summarizing the shift in the semantic and associated meaning of nuclear-related concepts across locations over time to explain how the use of concepts and how the context surrounding the use of concepts changes over time in each location of interest. Future work will focus on expanding our descriptive analytics for the rapid, spatiotemporal understanding of proliferation expertise and capabilities and the development of novel predictive and prescriptive analytics.

To develop novel predictive analytics, we will build on our prior work using DL models, for example, graph convolutional networks (GCNs) and long short-term memory networks (LSTMs), for anticipating the future [5], [15] and take advantage of recently emerged DL architectures for link prediction [42], [43]. We will leverage structured representations of nuclear domain knowledge using our context and content graphs combined with node embeddings and topic vectors to anticipate future proliferation expertise evolution. This will allow us to answer operational questions like "In what venue will a given country publish next?," "What topics will a given country publish on?," and "Which institutions will publish from a given country?"

A second avenue of future work focuses on identifying causal relationships between domain knowledge evolution and the knowledge generation processes (e.g., as has been done for the computational linguistics research community [44]). We will estimate the causal effects of varying treatments on outcomes of interest (e.g., acquisition of a nuclear capability, speed of capability acquisition) and support the recommendation of interventions to achieve the desired outcome.

As a result, our future work, expanding beyond the descriptive analytics introduced in this article, will support forecasting, prescription, counterfactual reasoning, and prescriptive intervention by allowing a user to first focus on an actor or capability of interest and then understand this visually in the context of related actors, capabilities, events, venues, and so on. These answers to the aforementioned key questions will then be presented in terms of this context. For example, when the user queries specific actor X, that actor's publications, affiliations, events, and capabilities will be visible. The tool will also show actor X's next likely capabilities in this context. Clicking on a predicted future capability will then reveal what relationships, events, and so on are likely to occur to obtain that capability. This then provides the end-user the means to have the model predict the effectiveness of certain interventions including early identification of potential failure to adhere to ITAR/export or treaty terms.

Example workflows could include examining how specific countries' scientists are reacting to the advances in hypersonic and ASATs with concern for its effect on deterrence and determine what scientists located in two countries entering hostilities are discussing related to nuclear issues (increased or decreased publications while hostilities occur versus during interims).

B. Ethical Considerations

Our approach leverages PAI [10] to identify and explain the evolution of global proliferation expertise and capability development. The use of PAI raises ethical considerations; often individuals who contribute or create content online are not aware or do not have the ability to opt in/out of research using their content or public interactions [45], [46]. In contrast to PAI mined from social media platforms, scientific publications and the knowledge they contain are meant to be publicly disseminated and leveraged to advance the field of study within and outside the specified scientific domains. Compared to traditional use wherein subsequent research studies leverage

published findings to motivate, support, or build upon the current state of the art, we leverage this public knowledge to identify and understand signals of proliferation and support reasoning and anticipation of future activity and expertise evolution.

C. Data Access and Reproducibility

Data and structured knowledge representations (context and content graphs) described in this article are maintained within documented data collections in the Berkeley Data Cloud (BDC),[8] under the "Global Expertise Forecasting" project. The data is stored in line-delimited json[9] files for ease of access and consistent readability across file and operating systems. The code developed on the project is publicly available on GitHub: https://github.com/pnnl/expert, to support the reproducibility of our analyses and extension to additional analyses.

VIII. Conclusion and Mission Impact

Our novel approach to learning global proliferation expertise evolution from PAI using AI-driven descriptive, predictive, and prescriptive analytics allows users to quickly uncover hidden patterns and relationships across multiple disparate public data sources and otherwise incomprehensible masses of open-source data. This work will:

- add strong multilingual, knowledge representation, and modeling components to traditional efforts;
- allow analysts to move away from traditional reactive analyses and take a proactive posture;
- provide a deeper understanding of how publicly available data could be used to detect, monitor, forecast, and prevent proliferation; and
- provide the quality, scale, and timeliness required for operational monitoring capability.

More generally, this work will enable more effective global allocation of resources, for example, better and faster validation of declarations and monitoring of safeguards and will allow non-proliferation efforts to move to a more predictive posture versus the current reactionary posture. Our interactive analytics will accelerate decisions and increase analyst efficiency and accuracy with advanced analysis capabilities that perform at speed and scale.

Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government or any agency thereof. The authors thank Kevin Cronk and Kate Gibb for their invaluable advice on the use cases, mission assurance, and mission alignment.

References

[1] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, "Content-based features predict social media influence operations," *Sci. Adv.*, vol. 6, no. 30, Jul. 2020, Art. no. eabb5824.

[2] L. Vargas, P. Emami, and P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *Proc. ACM SIGSAC Conf. Cloud Comput. Secur. Workshop*, Nov. 2020, pp. 133–146.

[3] O. Varol, E. Ferrara, F. Menczer, and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ Data Sci.*, vol. 6, pp. 1–19, 2017.

[4] C. Zhan, C. K. Tse, Y. Fu, Z. Lai, and H. Zhang, "Modeling and prediction of the 2019 coronavirus disease spreading in China incorporating human migration data," *PLoS ONE*, vol. 15, no. 10, Oct. 2020, Art. no. e0241171.

[5] S. Volkova, E. Ayton, K. Porterfield, and C. D. Corley, "Forecasting influenza-like illness dynamics for military populations using neural networks and social media," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0188941.

[6] G. Renda, L. Kim, R. Jungwirth, F. Pabian, E. Wolfart, and G. Cojazzi, "The potential of open source information in supporting acquisition pathway analysis to design iaea state level approaches," in *Proc. IAEA Int. Safeguards Symp., Linking Strategy, Implement. People*, 2014, pp. 1–10.

[7] F. Pabian, G. Renda, R. Jungwirth, L. Kim, E. Wolfart, and G. Cojazzi, "Open source analysis in support to non-proliferation monitoring and verification activities: Using the new media to derive unknown new information," in *Proc. Symp. Int. Safeguards, Linking Strategy, Implement. People*, vol. 312, 2014, pp. 1–10.

[8] M. Kas *et al.*, "Analyzing scientific networks for nuclear capabilities assessment," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 7, pp. 1294–1312, Jul. 2012.

[9] J. DIAB, P. BURR, and R. Stohr, "Using machine learning and natural language processing to enhance uranium mining and milling safegaurds," IAEA, Vienna, Austria Tech. Rep. IAEA-CN–267, 2018.

[10] J. Arterburn, E. D. Dumbacher, and P. O. Stoutland, "Preventing nuclear proliferation with machine learning and publicly available information," Nucl. Threat Initiative (NTI), Washington, DC, USA, Tech. Rep., 2021.

[11] S. Fortunato *et al.*, "Science of science," *Science*, vol. 359, no. 6379, 2018, Art. no. eaao0185.

[12] A. K. Khakimova, O. V. Zolotarev, and M. A. Berberova, "Visualization of bibliometric networks of scientific publications on the study of the human factor in the operation of nuclear power plants based on the bibliographic database dimensions," *Sci. Vis.*, vol. 12, no. 2, pp. 127–138, 2020.

[13] E. A. Agyeman and A. Bilson, "Research focus and trends in nuclear science and technology in Ghana: A bibliometric study based on the INIS database," *Library Philosophy Pract.*, vol. 4, pp. 1–45, Oct. 2015.

[14] K. Akbari and A. Bozorgi, "Citation analysis of articles indexed by atomic energy organization of Iran for INIS database during the years 2002-2006," Sci. Inf. Database, Tech. Rep., 2009.

[15] P. Shrestha, S. Maharjan, D. Arendt, and S. Volkova, "Learning from dynamic user interaction graphs to forecast diverse social behavior," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2033–2042.

[16] S. Volkova *et al.*, "Explaining and predicting human behavior and social dynamics in simulated virtual worlds: Reproducibility, generalizability, and robustness of causal discovery methods," *Comput. Math. Org. Theory*, pp. 1–22, 2021

[17] E. Saldanha *et al.*, "Evaluation of algorithm selection and ensemble methods for causal discovery," in *Proc. Workshop Causal Discovery Causality-Inspired Mach. Learn., Colocated NeurIPS*, 2020.

[18] D. Angelov, "Top2 Vec: Distributed representations of topics," 2020, *arXiv:2008.09470*.

[19] M. Gardner *et al.*, "AllenNLP: A deep semantic natural language processing platform," 2017, *arXiv:1803.07640*.

[20] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 188–197.

[21] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep semantic role labeling: What works and What's next," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 473–483.

[22] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3219–3232.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[24] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1–14.

[8] https://bdc.lbl.gov/

[9] https://www.json.org/json-en.html

[25] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 1–8. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[26] D. Larionov *et al.*, "Semantic role labeling with pretrained language models for known and unknown predicates," in *Proc. Natural Lang. Process. Deep Learn. World*, Oct. 2019, pp. 619–628.

[27] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Comput. Linguistics*, vol. 31, no. 1, pp. 71–106, Mar. 2005.

[28] A. Aloraini, J. Yu, and M. Poesio, "Neural coreference resolution for Arabic," 2020, *arXiv:2011.00286*.

[29] K. Ak, C. Toprak, V. Esgel, and O. T. Yildiz, "Construction of a Turkish proposition bank," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 26, pp. 570–581, 2018.

[30] V. Likic, "The needleman-wunsch algorithm for sequence alignment," in *Proc. 7th Melbourne Bioinf. Course*, 2008, pp. 1–46.

[31] F. Morillo, I. Santabárbara, and J. Aparicio, "The automatic normalisation challenge: Detailed addresses identification," *Scientometrics*, vol. 95, no. 3, pp. 953–966, Jun. 2013.

[32] C. A. D'Angelo and N. J. van Eck, "Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation," *Scientometrics*, vol. 123, no. 2, pp. 883–907, May 2020.

[33] C. Schulz, "Exploiting citation networks for large-scale author name disambiguation," *EPJ Data Sci.*, vol. 3, no. 11, pp. 1–14, Sep. 2014.

[34] S. Huang, B. Yang, S. Yan, and R. Rousseau, "Institution name disambiguation for research assessment," *Scientometrics*, vol. 99, no. 3, pp. 823–838, Jun. 2014.

[35] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, 2004, Art. no. 066111.

[36] M. Newman, *Networking*. London, U.K.: Oxford Univ. Press, 2018.

[37] D. Arendt and S. Volkova, "Esteem: A novel framework for qualitatively evaluating and visualizing spatiotemporal embeddings in social media," in *Proc. ACL*, 2017, pp. 25–30.

[38] C. Bischof and D. Wilfinger, "Big data-enhanced risk management," *Trans. FAMENA*, vol. 43, no. 2, pp. 73–84, Jul. 2019.

[39] C. J. Unger, A. M. Lechner, J. Kenway, V. Glenn, and A. Walton, "A jurisdictional maturity model for risk management, accountability and continual improvement of abandoned mine remediation programs," *Resour. Policy*, vol. 43, pp. 1–10, Mar. 2015.

[40] Y. Badr, S. Hariri, A.-N. Youssif, and E. Blasch, "Resilient and trustworthy dynamic data-driven application systems (DDDAS) services for crisis management environments," *Proc. Comput. Sci.*, vol. 51, pp. 2623–2637, Dec. 2015.

[41] S. García-Herrero, M. A. Mariscal, J. M. Gutiérrez, and A. Toca-Otero, "Bayesian network analysis of safety culture and organizational culture in a nuclear power plant," *Saf. Sci.*, vol. 53, pp. 82–95, Mar. 2013.

[42] W. Jin, M. Qu, X. Jin, and X. Ren, "Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6669–6683.

[43] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," 2020, *arXiv:2006.10637*.

[44] M. Glenski and S. Volkova, "Identifying causal influences on publication trends and behavior: A case study of the computational linguistics community," in *Proc. 1st Workshop Causal Inference NLP*, 2021, pp. 83–94.

[45] C. Fiesler and N. Proferes, "'Participant' perceptions of Twitter research ethics," *Social Media Soc.*, vol. 4, no. 1, 2018, Art. no. 2056305118763366.

[46] K. Beninger, A. Fry, N. Jago, H. Lepps, L. Nass, and H. Silvester, "Research using social media; users' views," *NatCen Social Res.*, vol. 4, pp. 1–40, Feb. 2014.