# 2024 the 9th International Conference on Big Data Analytics

# ICBDA 2024

## March 16-18, 2024

### Tokyo, Japan

**Sponsored by:**

**Hosted by:**

**Patrons:**

# Towards Understanding Low Energy Nuclear Reactions: Structuring the Literature and Applying Data Analytics

Anasse Bari*
*Computer Science Dept*
*New York University*
*New York, USA*
*abari@nyu.edu*

Sneha Singh*
*Computer Science Dept*
*New York University*
*New York, USA*
*ss15639@nyu.edu*

Yvonne Wu
*Computer Science Dept*
*New York University*
*New York, USA*
*yw4142@nyu.edu*

Tanya Pushkin Garg
*Computer Science Dept*
*New York University*
*New York, USA*
*tg2520@nyu.edu*

Emos Ker
*Computer Science Dept*
*New York University*
*New York, USA*
*ck3189@nyu.edu*

Charles Wang
*Computer Science Dept*
*New York University*
*New York, USA*
*mw4899@nyu.edu*

Benjamin Kang
*Computer Science Dept*
*New York University*
*New York, USA*
*bhk9599@nyu.edu*

Adelina Simpson
*Computer Science Dept*
*New York University*
*New York, USA*
*asg683@nyu.edu*

Dongjoo Lee
*Computer Science Dept*
*New York University*
*New York, USA*
*dl4273@nyu.edu*

David Nagel*
*George Washington University*
*Washington D.C., USA*
*nagel@gwu.edu*

*Corresponding authors

***Abstract*— After more than three decades of research, Low Energy Nuclear Reactions (LENR) still pose a challenge to comprehensive understanding. In this study, we introduce a preliminary framework of analytical tools that could assist the LENR community in accessing literature and applying machine learning for mining insights. We first collected and structured a dataset of over 4,500 LENR publications. Following that, we designed and deployed a descriptive analytics tool to search and draw insights by using a platform that allows data slicing based on keywords, authors, publication dates, and other metadata. Additionally, we applied unsupervised machine learning algorithms to the data to generate clusters of publications based on semantics and other features. Through interactive interfaces, we enable targeted investigation of specific reported phenomena. Findings provide data-driven insights into the connections between concepts like heat, helium, transmutation, and other measured effects. The analysis also identifies key contributing authors, organizations, and publication venues. By moving beyond isolated results to higher level knowledge, this work aims to advance the field by providing revealing relationships in the collective evidence. Our preliminary results and the first version of our tools will be useful for scientists already in the LENR field, as well as for those considering research on the topic. Additionally, they will benefit other scientists and policymakers. We conclude that data science approaches show promise for demystifying and advancing LENR, and merit further application to this complex scientific challenge.**

**Keywords—Low Energy Nuclear Reactions (LENR), Natural Language Processing, Predictive Analytics, Artificial Intelligence**

## I. INTRODUCTION

The world needs new sources of clean energy due to relentless global population growth, increasing per-capita uses of energy for development, and the numerous and evident effects of climate change. One potential new source might do for energy what cell phones did for communications.

Low energy nuclear reactions (LENR), also known as cold fusion, represent a controversial field first reported in 1989 by Fleischmann and Pons. The claims of nuclear reactions occurring at modest temperatures and energies have drawn significant interest and research over the past 30+ years. However, despite substantial efforts, the mechanisms and outputs of LENR remain poorly understood and hotly debated.

A key challenge is that LENR research exists as a massive, fragmented collection of findings across thousands of papers, presentations, and news reports generated over decades [1]. This robust but disorganized corpus contains observational data, experimental designs, proposed theories, analytic methods, and more. However, with content scattered across so many sources, there has been limited consolidation or unified analysis. Nonetheless, there is substantial information about LENR available on the internet [2].

The unstructured nature of the collective LENR literature limits researchers' ability to draw connections, patterns, and insights through data-driven techniques. As an emerging interdisciplinary field, impactful knowledge discovery requires synthesis across findings from physics, chemistry, materials science, nuclear engineering, and other domains.

This research aims to address these gaps by applying text mining, natural language processing, and data visualization methods to extract insights from the expansive LENR textual corpus. By leveraging data analytics tools and algorithms to index and structure content, interactive interfaces can enable targeted investigation of specific concepts, authors, and relationship hidden within decades of accumulated information.

## II. LENR Research Data Landscape

Low energy nuclear reaction research represents an extensive yet disorganized data corpus distributed across thousands of papers, presentations, and reports. This work focuses on leveraging the robust LENR bibliography compiled by Jed Rothwell as the primary data source [1].

Rothwell's archive contains over 4,741 total entries, with many available for downloads [1]. Each entry includes metadata such as title, author list, publication year, place of publication source, and abstract. Additionally, direct PDF links are provided for 2,174 documents.

This corpus covers a wide range of publication types including journal articles, conference presentations, white papers, books, newsletters, and more. The time span runs from the origins of cold fusion reports in the late 1980s through state-of-the-art papers in recent years. Content covers experimental and theoretical results using tools from physics, chemistry, materials science, engineering, and other relevant domains.

In total, Rothwell's archive gives access to over 50,000 pages of structured LENR research content. While not exhaustive, this constitutes the largest known consolidated body of LENR literature. The corpus therefore provides a substantial foundation for applying text mining and visualization techniques to derive insights.

### *A. Description of the Data Hosted by Rothwell:*

Out of the 4,741 total entries in the Rothwell bibliography, 2,174 include direct links to PDF documents hosted on the LENR-CANR servers [1]. We programmatically downloaded these linked PDFs for further analysis and processing. A preliminary examination of the sources reveals a heterogeneous mix of publication types and styles. To understand the composition of the corpus, we classified the documents into the following categories:

TABLE I
CLASSIFICATION FOR PAPER TYPES

| PAPER TYPE | COUNT |
|---|---|
| Research Papers | 2292 |
| White Papers/ News Articles | 2291 |
| Presentation Slides | 158 |
| **Total:** | **4741** |

This categorization provides useful context on the wide range of document types present, from formal academic publications to informal updates. In later sections, we will explore how the paper type may influence things like length, tone, and topics covered. This metadata helps guide deeper content analysis.

### *B. Searching and Indexing Missing LENR Papers*

The initial 2,174 PDFs from LENR-CANR served as our core dataset. However, to expand the corpus, we manually searched and indexed documents for the remaining 2,567 bibliography entries without direct links. Leveraging Google Scholar, ScienceDirect, Nature, IEEE Xplore, and public preprint repositories, we retrieved links for an additional 1,250 papers and reports.

In total, this exhaustive process yielded 3,424 total LENR documents indexed into a centralized database we designed. This expanded corpus provides broader coverage of the field by incorporating related papers beyond only those hosted by LENR-CANR. With over 200,000 additional pages of content, the wider collection facilitates more robust analysis of publication trends, topic modelling, and conceptual connections across the scattered literature.

### *C. New Energy Times: Additional Documentation*

While our current corpus provides extensive coverage of LENR research over decades, additional relevant literature exists that could further enrich the analysis. As documented by resources like the LENR Reference Site [2] supplementary content spans books, preprints, international publications, and more. For example, key collections identified for prospective inclusion cover:

- 7 books focused specifically on LENR phenomena and energy production.
- 22 experimental papers from the New Energy Times, categorized by method like electrolytic, gas, and unique systems.
- 56 publications in Elsevier journals and 19 Oxford papers related to LENR and low energy reactions.
- Literature in Russian, Japanese, and Spanish collected by researchers like Alexandre Prosvirnov, Japan CF and Spanish Symposium.
- 11 articles authored by noted physicists Steven Jones and Richard Garwin.

Integrating these additional papers and reports in their native languages could significantly diversify the LENR corpus. Advanced machine translation techniques could allow analysing non-English content.

While outside the scope of this initial study, augmenting the literature with these external sources presents a valuable opportunity for future work. The augmented multilingual corpus could provide a more complete landscape of global LENR research over the decades. We hope the methods and frameworks introduced here will serve as foundations for continually expanding this collective data resource.

## III. Motivations: Taming an Unruly Literature

In recent years, vector databases and new large language models have allowed the directly querying information buried in mountains of unstructured text. However, readily applying these powerful techniques requires structured sources, something lacking in the scattered LENR literature.

Without integrated metadata, a vector similarity search cannot connect related across ideas decades of research; language models cannot uncover insights from specific papers without indexes to retrieve them. By wrangling this corpus into a coherent resource, we enable using modern AI to tackle long-standing knowledge gaps.

Our tools extract key fields like title, authors, text, and keywords to power semantic search and unsupervised learning. Structured metadata allows recommending papers based on novel experimental reports. Our granular keyword data enables querying all papers related to the searched term over time.

By organizing, indexing, and visualizing this complex landscape, we aim to construct a framework where the research community can intuitively explore both details and the bigger picture. Researchers can rapidly find evidence to confirm or refute claims by reviewing abstracts and quantified trends. With these tools, collective progress need not be impeded or stalled due to difficulties in assimilating and using thousands of pages about methods and data.

The urgency of advancing low-carbon energy technologies further motivates consolidating the findings of this controversial but promising field. We hope by taming this unruly literature, our work helps researchers wield modern techniques to bridge knowledge gaps. The insights unlocked could hasten basic understanding of LENR and speed its commercialization to solve urgent global energy-related challenges.

## IV. Pre-processing an Unstructured Corpus

To enable structured analysis, the corpus of LENR documents required pre-processing into a standardized format. Our extracted metadata fields for each paper included mainly:

- Title
- Publication year
- Authors
- Abstract
- Publisher
- Keywords
- Full-text content

For documents with existing metadata (e.g., BibTeX data), we indexed and parsed them directly. For PDFs and other raw content, we utilized techniques like optical character recognition and natural language processing to extract metadata heuristically.

The content was organized into a Pandas DataFrame - a table-like data structure commonly used for data analysis in Python. A unique ID was also added to identify each document. The text columns were handled as string data types for ease of manipulation versus numerical data.

This organized data frame allowed the corpus to be programmatically analysed row-by-row using the capabilities of the Pandas library in Python. Pandas provide powerful functions for slicing, dicing, filtering, and aggregating large datasets. For example, the publication years could be summarized to identify trends over time. And keywords and key phrases frequency counts could be generated to reveal research themes.

By structuring the collection of documents into a consistent DataFrame object, we enabled the application of Pandas' versatile data analysis tools to mine insights. The tabular representation facilitated quantitative analysis like statistical summaries, joining data, visualization, and machine learning. Converting the corpus into a structured format was a crucial preparatory step before data mining could occur.

In addition to text data, many of the LENR documents contained relevant data in the form of images, figures, and tables. To fully leverage the corpus, we also extracted these non-textual elements for each LENR paper.

## V. Apache Solr Indexing to Wrangle the Corpus

To facilitate robust analysis of the heterogeneous LENR literature, we need the ability to rapidly search, filter, and aggregate across thousands of documents and metadata fields. For this critical information retrieval task, we leverage Apache Solr, a highly scalable open-source enterprise search platform built on Lucene [3].

Solr provides powerful full-text search and indexing capabilities that are well-suited for mining large collections of text-heavy data. We initially indexed the LENR corpus into Solr using unique document IDs which can then be queried through Solr's search APIs (Application Programming Interfaces). APIs act as an interface to allow software programs to communicate with each other; in this case, it enables our search engine frontend to retrieve data from the Solr index.

Each document is tokenized and processed to extract various textual features and metadata. Solr's schema flexibility allows ingesting and indexing all document metadata.

Overall, Solr provides the foundational data retrieval and exploration functionality needed for a corpus of this scale and heterogeneity. The indexed LENR collection enables fast keyword search, filtering of metadata fields, and aggregation for analysis.

As can be seen in Fig. 1, the Solr index is configured through two key files - schema.xml and solrconfig.xml.

The schema.xml file defines the structure and fields of the documents being indexed into Solr. For example, for a news article document, schema.xml would define fields like title, body, tags, publish date etc. along with their data types like string, int, date etc. It essentially configures the schema of the index - the document fields and their datatypes.

The solrconfig.xml handles all the configuration for request handlers and other settings in Solr. Request handlers are URLs that execute specific plugins or Java code in Solr. solrconfig.xml defines and configures these request handlers. For example, it may configure a request handler to execute a spellcheck plugin.

In summary, schema.xml structures the indexed documents and solrconfig.xml configures the request handlers and settings for the Solr server. Together, they allow customizing and optimizing the Solr index as per the application requirements. The diagram provides a high-level overview of how these two key configuration files interact with the Solr server and documents being indexed.

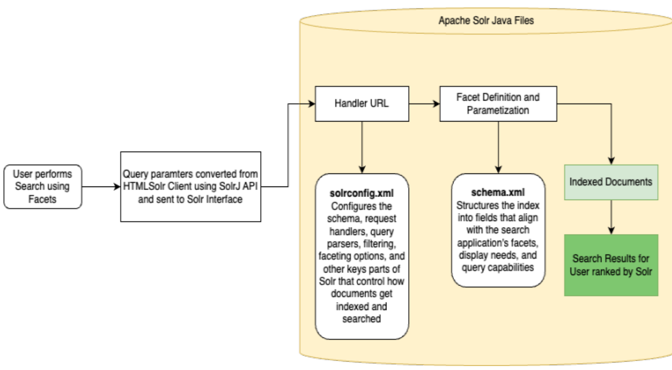



Fig. 1. Solr Search Implementation Architecture

## VI. Solr Schema Specifications

While Apache Solr provides extensive search functionality out-of-the-box, we customized key components to develop an optimized search engine for investigating the LENR literature. Tuning Solr's configuration and features to the corpus domain and interface requirements was critical for delivering an intuitive experience.

In this section, we detail the customizations made across various areas of Solr to enable performant indexing, relevant search, and responsive faceted browsing of the collected papers. Performant indexing refers to the speed and efficiency at which Solr can ingest content, build index shards, and replicate data. This determines how fast documents can be made searchable, which is crucial for very large collections.

### *A. Storing Publicly Available PDFs: Apache Tika Implemented Using Solr Cell*

A key challenge was ingesting content from the thousands of academic unstructured PDF papers in our LENR corpus. Unlike plain text documents, Solr cannot directly extract the raw text and metadata from PDFs to enable search and analysis.

The first step comes to defining the keys for our data. We specify our schema in the schema.xml file as:

```
<field name="all_authors" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="year_published" type="tint" indexed="true" stored="true"/>
<field name="title" type="text_general" indexed="true" stored="true"/>
<field name="publisher" type="text_general" indexed="true" stored="true"/>
<field name="document_link" type="text_general" indexed="true" stored="true"/>
<field name="abstract" type="text_general" indexed="true" stored="true"/>
<field name="keywords" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="text_from_pdf" type="text_general" indexed="true" stored="true"/>
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false"/>
```

We then integrated Apache Tika with Solr to leverage its PDF parsing capabilities. Tika provides Java libraries packaged as JAR (Java ARchive) files that contain classes and code for parsing different file formats.

For Solr to access the Tika Java code, we first needed to specify the JAR files containing those libraries in the solrconfig.xml configuration. JARs allow reusable Java code modules like Tika to be distributed and imported into projects as self-contained packages. Listing the Tika JAR file paths in solrconfig.xml effectively lets Solr know where to find the Java classes it needs to use the Tika parsing functionality. At runtime, Solr will load those libraries dynamically, so the Tika parsers are accessible. Specifying just the tika-app JAR itself in our configuration enabled all of Tika's parsing capabilities for Solr, including PDF, Word, Excel, PowerPoint, and many more formats. The JARs contain the Java implementations powering Tika's robust file content extraction features.

By integrating the Tika libraries in this way, we were able to leverage Tika's mature PDF parsing capabilities out-of-the-box within Solr, greatly simplifying the ingestion of literature PDFs into our search index. The power of Tika was available just by including the JARs containing the parser code modules.

```
<lib dir="${solr.install.dir:../../..}/contrib/extraction/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../..}/dist/" regex="solr-cell-\d.*\.jar" />
```

This is followed by us specifying the request handler /update/extract as:

```
<requestHandler name="/update/extract" startup="lazy" class="solr.extraction.ExtractingRequestHandler">
  <lst name="defaults">
    <str name="fmap.content">text_from_pdf</str>
    <str name="uprefix">ignored_</str>
    <str name="captureAttr">true</str>
  </lst>
</requestHandler>
```

## VII. Design and Deployment of a Descriptive Analytics Framework of LENR Data: [https://lenrdashboard.com/](https://lenrdashboard.com/)

To complement the indexed search and retrieval capabilities provided by Solr, we designed and developed interactive visualizations to reveal insights and trends within the corpus metadata that is publicly available at: [https://lenrdashboard.com/](https://lenrdashboard.com/). Effective data visualization enables us to analyse the aggregated LENR literature in new ways. The framework provides a search capability, interactive graphs, as well as insights from applying unsupervised learning to LENR data. Following are the visualizations we found to be key insights into our data:

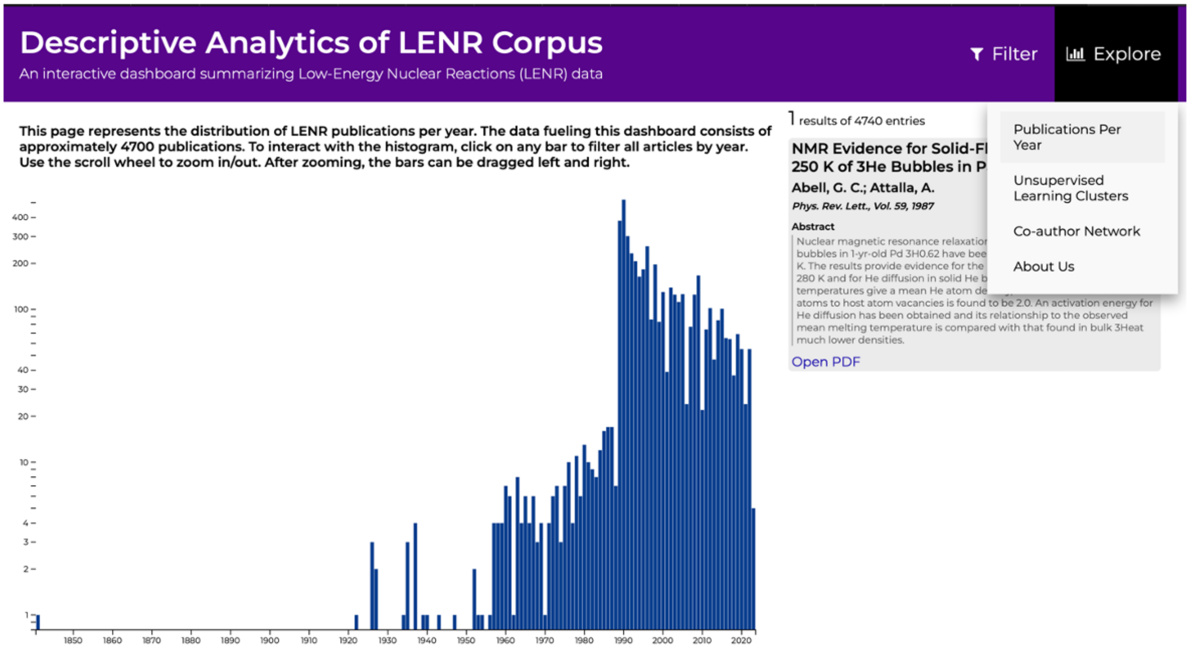



Fig. 2. Bar Graph showing yearly distribution of papers published

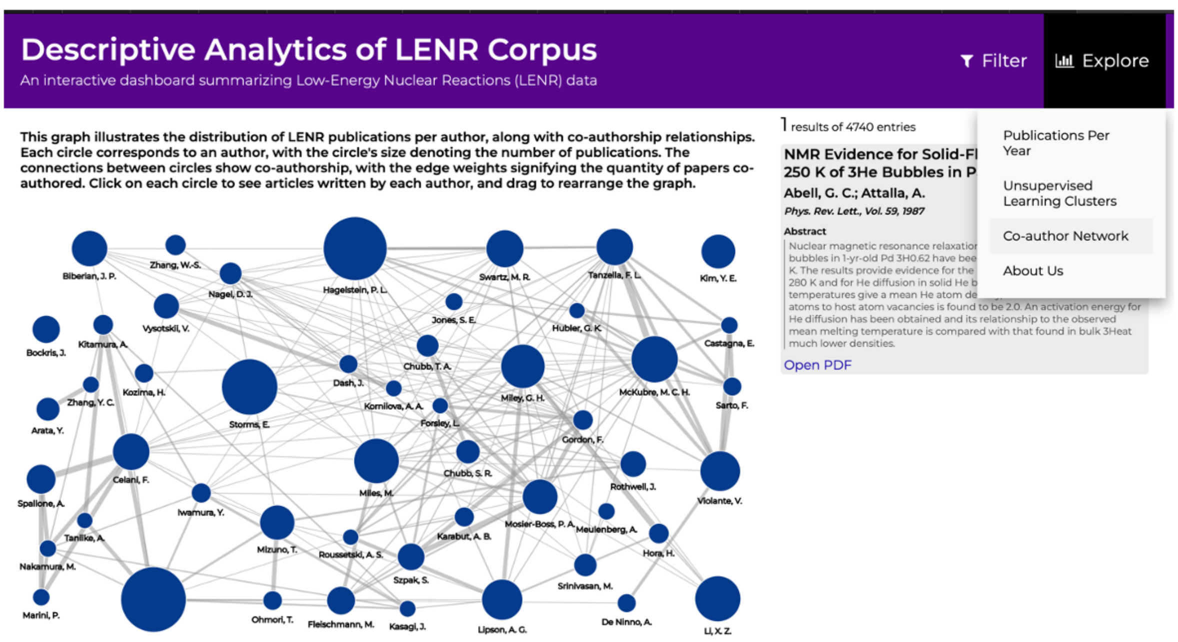



Fig. 3. Co-Author network to show top collaborators and their respective works

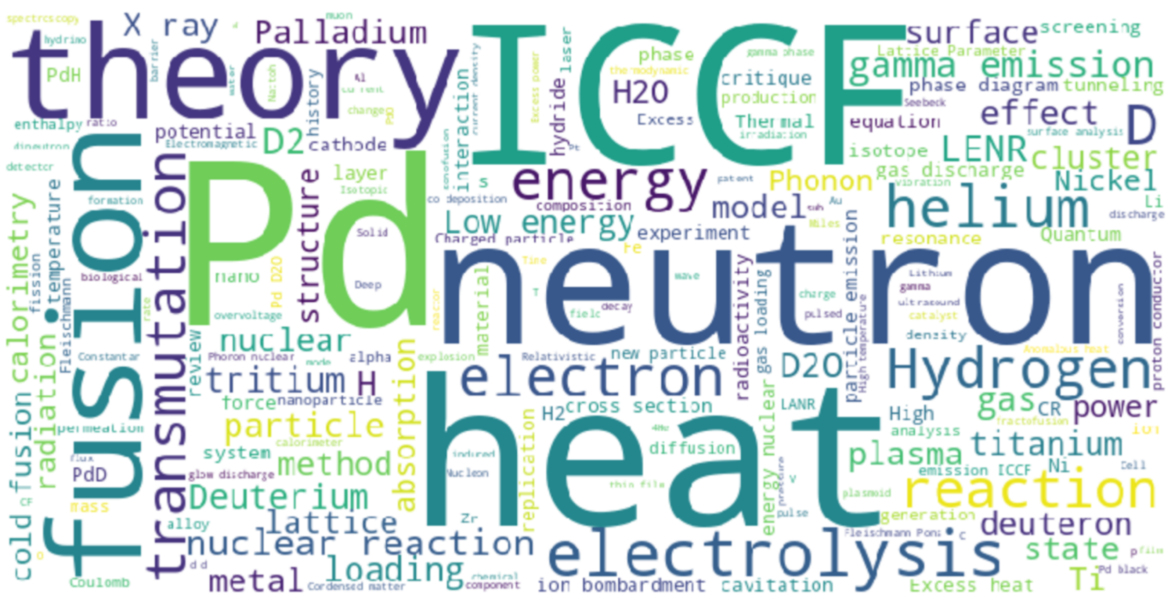



Fig. 4. Keyword Cloud to depict main focus areas of the research

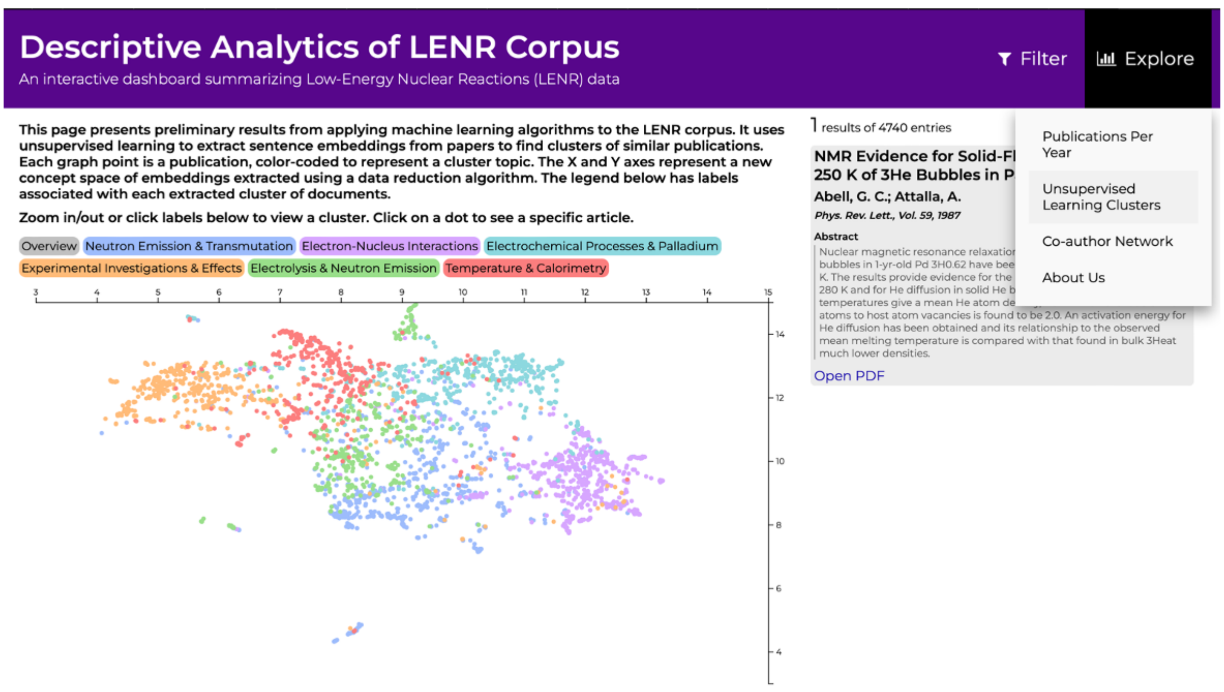



Fig. 5. Document Clustering based on research topics using Unsupervised Learning

### *A. Distribution of Yearly Publication Volumes Demonstrating the Timeline and Popularity of LENR Research*

Figure 2 presents a time series visualization of the number of LENR-related papers published annually within the corpus. Interactive clicking on each bar allows drilling down into the specific papers from a given year. The overall trendline can be seen in Figure 8 in supplementary material.

The trend shows a dramatic spike in publications around 1989-1990, when the controversial concept of cold fusion was first reported by Fleischmann and Pons. This intense research interest reached over 500 papers in the peak year of 1990.

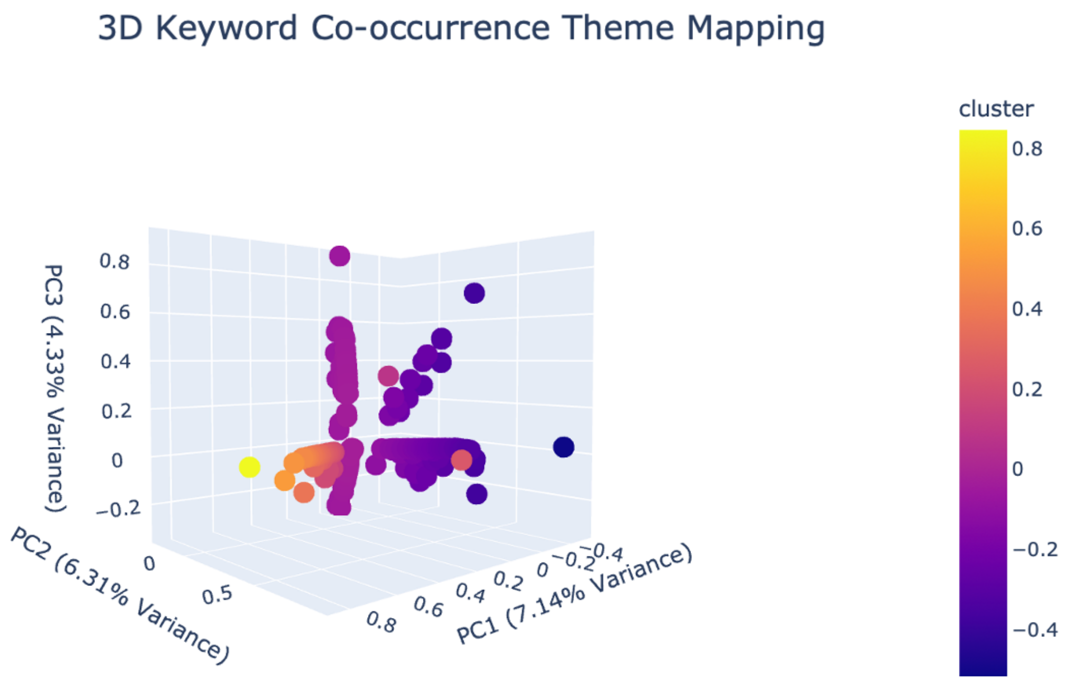



Fig. 6. 3D Scatter plot of keyword co-occurrence

Subsequent years showed a downward trend, with publications declining to under 100 papers in the late 1990s as enthusiasm waned. However, the early 2000s marked a resurgence, with yearly output rising back to over 200 papers by 2002. This suggests LENR continues to attract significant research attention despite lingering scepticism. The ebb and flow highlights periods of heightened activity around major developments like specific cold fusion claims and emerging new experimental evidence.

Further temporal analysis on a granular level could reveal additional patterns tied to funding, global events, energy prices, and other potential drivers. The timeline provides a high-level overview of the shifting landscape across decades of research.

### B. LENR Co-Author Network

Figure 3 provides a network visualization of co-author relationships within the LENR corpus, revealing collaboration patterns. Each node represents an individual author, with node size proportional to their publication volume. Links between nodes indicate authors have co-authored one or more papers, with edge thickness indicating the number of collaborations.

Clicking on a node lists all papers published by that author for inspection. Clicking on an edge between two nodes brings up the specific papers they co-authored. This interactivity enables drilling down into individual collaborations.

The network topology highlights a number of dense clusters, suggesting tight-knit research groups focused on experimental cold fusion work. We also see key prolific authors acting as bridges between clusters. This co-authorship visualization provides unique insights into the social dynamics that underpin this research domain.

### C. Keyword Cloud Visualizations of the Corpus Highlighting the Main Focus Areas of Research

Figure 4 illustrates a keyword cloud visualization to showcase the most prominent topics and themes within the LENR corpus. The size of each word indicates its relative frequency, revealing the central concepts and terms discussed frequently across publications.

The most dominant terms were "transmutation," "heat," and "theory", which are core areas of investigation related to nuclear reactions and proposed cold fusion mechanisms. Additional keywords like "palladium," "deuterium," "reaction," "fusion," and "electrolysis" highlight commonly studied materials and experimental configurations.

By distilling thousands of publications into salient terms, we can rapidly grasp the thematic landscape. The keyword frequencies effectively summarize the core experimental and theoretical foci within LENR work.

### D. Documents Plotted as Clusters Derived on Semantic Similarity Based on Research Topics

To reveal underlying thematic structures within the corpus, a subset of documents was manually annotated with key conceptual topics discussed in each paper. We manually scouted a list of 50 concepts to categorize these documents. These topics are listed in the supplement material in Table 2.

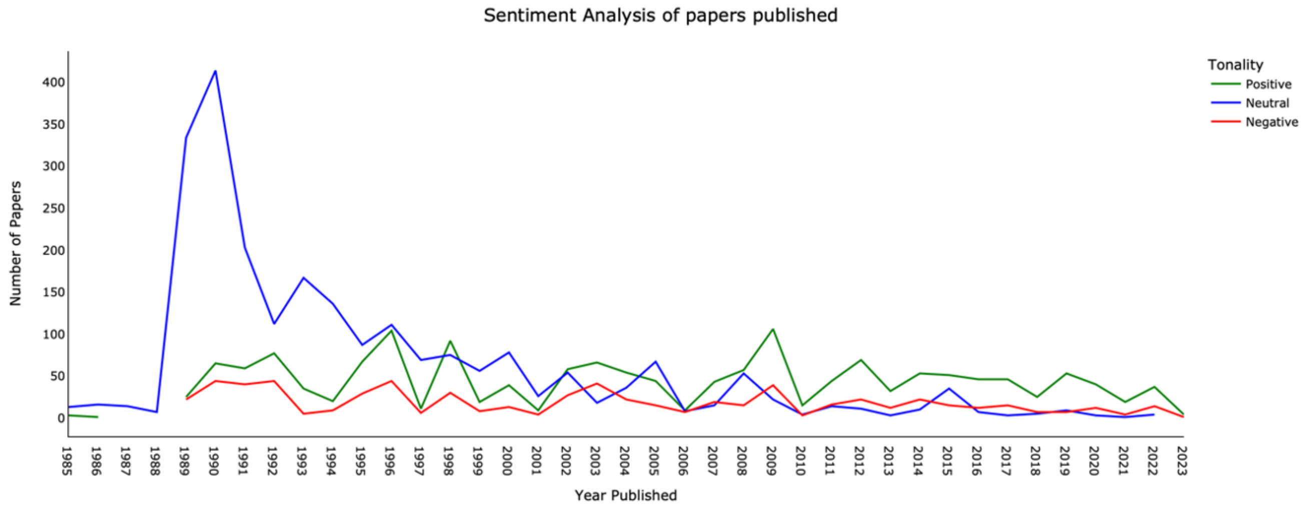



Fig. 7. Sentiment analysis of papers published grouped by year to understand the tonality of research discussions over the year

The scattering of documents in the 2D space shows their relative similarity based on content, with clusters forming for documents on similar topics by applying TF-IDF vectorization, truncated SVD dimensionality reduction, and t-SNE, documents were projected into a 2D embedding space representing semantic similarity based on their textual content in Figure 5. Visual inspection shows a distinct clustering of documents around the assigned topics. Papers related to condensed matter group together, separate from those focused on nuclear renaissance. This provides quantitative validation that our manual topical annotations effectively capture meaningful semantic differences.

While limited in scale, this supervised mapping technique highlights the potential of unsupervised methods like LDA topic modelling to systematically discover latent topics and cluster documents across the broader corpus. The embeddings provide features for focused searches or recommendation systems. With further refinement, projected topic maps could bring structure to a disjointed domain, revealing connections between ideas otherwise buried within decades of text.

### *E. Keyword Co-occurrence to Uncover Latent Semantic Relationships*

Dimensionality reduction and clustering techniques were applied to analyse semantic relationships within the corpus keywords. The pre-processing involved converting keyword lists into consistent strings, then applying TF-IDF vectorization to encode the keywords [10]. Principal component analysis (PCA) was utilized to reduce the high-dimensional keyword vectors down to the most significant components [11]. A three-component PCA model was trained on the TF-IDF keyword matrix to obtain condensed semantic theme vectors for each keyword.

Each keyword is positioned based on its values on the three identified semantic themes. Hovering shows the theme values along with the keyword itself. By colouring based on density-based clustering, we can see keywords grouped into semantic clusters, as illustrated in Figure 6. This provides a method to uncover latent semantic relationships between terms in a corpus.

### *F. Tonality of Discussion of LENR Research*

To assess changing attitudes and tone over the course of LENR research history, we conducted sentiment analysis on the abstracts of papers published in each year. As visualized in Figure 7, sentiment was classified into positive, negative, and neutral categories based on keyword spotting and natural language processing techniques.

The results show a decline in positive sentiment from the initial spike around 1989-1990 when cold fusion was first announced. This corresponds to the initial enthusiasm that soon gave way to scepticism. Negative sentiment rises through the mid-1990s as criticism mounted.

However, from the 2000s onward, sentiment has stabilized with relatively balanced positive and negative tones, suggesting the field is established but contentious. There are opportunities for a more nuanced analysis of factors influencing sentiment shifts, but the high-level timeline provides insight into the evolving reception of LENR claims over decades.

Capturing this qualitative narrative arc complemented empirical analysis, painting a richer portrait of the ups and downs within this controversial domain. The multi-faceted elucidation demonstrates how text mining can uncover insights beyond just trends and patterns.

## VIII. TARGETED INVESTIGATIONS THROUGH FACETED SOLR QUERIES

Extending the visualization-based insights, we also leveraged the power of Solr indexing to build an interactive search engine for targeted corpus investigation. Specifically, we develop a faceted search interface that allows drilling down to papers of interest across multiple metadata dimensions.

Faceted search involves applying filters to progressively narrow results based on category, attributes, or taxonomy. For the LENR corpus, key facets included author names, publication year, source title, keywords, and full-text search.

We implemented the frontend search interface using Solr's web UI components like the Facet Pivot which enables dynamic filtering and aggregation. The responsive interface allows quickly reviewing and refining results through applied facets.

Key features include:

- Search by author name
- Filter by publication year
- Refine by source titles like journals
- Multi-select keyword and title

For the search results, we summarize each paper with title, authors, year, source, and abstract. A "Read More" link provides one-click access to the full text.

By combining Solr's capabilities with an intuitive interface, researchers can now intelligently probe the corpus to find papers and connections relevant to their interests. The faceted exploration encourages serendipitous discovery. Screenshots of the search platform showing available filters can be found in Supplementary Figures 8-13. Search operations by "Author", "Lead Author" and "Date" have been visualized and presented in Supplementary Figures 14-15.

## IX. Exploring the Search Interface for LENR SOLR Search

To enable readers to directly experience the interactive search capabilities developed, this section provides a walkthrough of the system functionality using illustrative step-by-step instructions.

First, the user navigates to the search portal as seen in Fig. 8. To filter, author names can be searched under the "Search by Author" facet. They can filter based on key concepts and titles through the facet "Search by title/keyword" as in Fig. 9. The users can also select all the publications from a given journal which can be accessed through the dropdown as can be seen in Fig. 10 and Fig. 11. The "Year range" slider lets the user filter their search by year range as shown on Fig. 12.

This walkthrough introduces the intuitive search interface that was developed, empowering both targeted known-item investigation and exploratory discovery within the literature.

## X. Conclusion

This research demonstrates the potential of applying modern data science techniques to synthesize insights from the expansive and scattered literature on low energy nuclear reactions. By collecting, organizing, and analyzing the corpus of LENR literature, we were able to illuminate key trends, themes, and relationships buried within decades of accumulated research results.

Our multi-faceted framework integrating text mining, visualization, and interactive search solutions provides a toolkit to make sense of this complex scientific domain. Both high-level knowledge discovery and targeted investigation are enabled through our processed dataset and analytics platforms.

Key findings include unraveling the timeline of research enthusiasm, uncovering prominent authors and collaborators, revealing central topical focuses, and profiling the evolving community sentiment over 30+ years of activity. Both confirmatory and unexpected insights emerged from harmonizing disparate data sources.

There remain ample opportunities to further refine, extend, and apply our frameworks to expand the collective understanding of LENR. Additional papers can be integrated, more advanced models deployed, and new interface features added. We hope this work provides a foundation for the community to build upon in demystifying and advancing LENR science.

Ultimately, the solutions developed here could be adapted to making sense of other controversial, fragmented, or rapidly evolving fields. Taming the scientific literature to extract knowledge is crucial for everything from fusion to physics beyond the standard model. We believe the techniques pioneered in this analysis can accelerate discovery across disciplines.

There are several exciting opportunities to build upon this research through additional data types, models, and system capabilities. One aim is incorporating semantic similarity search to retrieve related papers based on document embeddings. We plan to train neural ranking models to further enhance the keyword search accuracy and relevance.

Another goal is expanding the analysis beyond just text to include the graphs, images, and tabular data extracted from the papers. Techniques like computer vision could enable image similarity comparisons to connect related experimental setups. Analysis of tables could reveal relationships between experimental parameters and observed results.

Overall, we hope to evolve the system into a cutting-edge AI research assistant that scientists can interact with to explore papers, data, and ideas. By combining multiple modalities of data with learned connections between concepts, the scope and impact of knowledge could significantly expand.

## Acknowledgments

The work presented in this paper constitutes a minor segment of an extensive research initiative spearheaded by Prof. David Nagel from George Washington University and Prof. Anasse Bari from New York University. The project, sponsored by the Anthropocene Institute, is dedicated to advancing AI and

Predictive Analytics tools to facilitate the commercialization of LENR.

Sneha Singh played a pivotal leading role in the Solr development efforts and also in crafting the initial version of this paper that was later edited by Prof. Bari and Prof. Nagel. The comprehensive data collection and indexing were carried out by Sneha Singh, Tanya Pushkin Garg, Yiyi Wu, Emos Ker, Charles Wang, Adelina Simpson, Gurmehr Sohi, Saiteja Siddana, and Dongjoo Lee.

The conceptualization and design of the dashboard (https://lenrdashboard.com/) were attributed to Yvonne Wu. Noteworthy contributions to the dashboard were made by Tanya Pushkin Garg, Sneha Singh, and Ben Kang, among other team members. All contributors are esteemed members of the Predictive Analytics and AI group at New York University's Courant Institute, led by Prof. Anasse Bari.

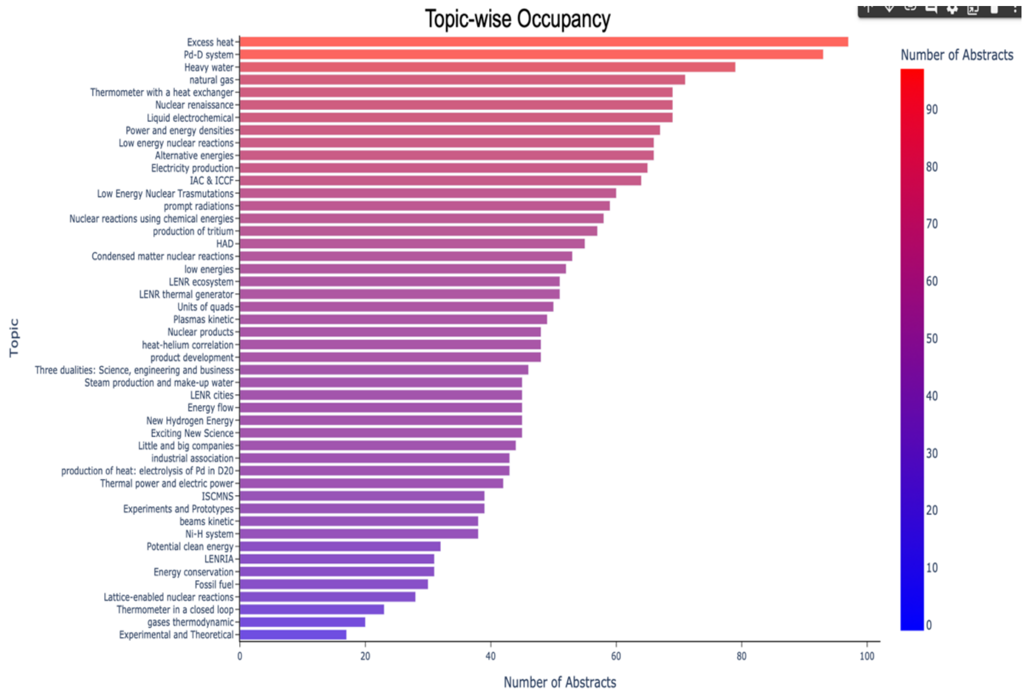
## References

[1] "LENR-CANR.org — A library of papers about cold fusion," lenr-canr.org. https://lenr-canr.org
[2] S. B. Krivit, LENR Reference Site, https://www.newenergytimes.com/
[3] "Welcome to Apache Lucene," *lucene.apache.org*. https://lucene.apache.org/
[4] T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, Apr. 2016
[5] Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. *Extensions of recurrent neural network language model*, 2011, IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5528-5531). IEEE.
[6] S. Li et al., "Geospatial big data handling theory and methods: A review and research challenges," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 115, pp. 119–133, May 2016, doi: https://doi.org/10.1016/j.isprsjprs.2015.10.012.
[7] J. Pound, P. Mika, and H. Zaragoza, "Ad-hoc Object Retrieval in the Web of Data."
[8] Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual Analytics: Definition, Process, and Challenges." https://d-nb.info/1098134664/34
[9] O. Noran and P. Bernus, "Towards an Evaluation Framework for Ubiquitous, Self-Evolving Patient Identification Solutions in Health Information Systems," Procedia Computer Science, vol. 196, pp. 550–560, 2022, doi: https://doi.org/10.1016/j.procs.2021.12.048.
[10] K. Guntupally, K. Dumas, W. Darnell, M. C. Crow, R. Devarakonda, and G. Prakash, "Automated Indexing of Structured Scientific Metadata Using Apache Solr," www.osti.gov, Dec. 01, 2020
[11] B. Kules, R. Capra, M. Banta, and T. Sierra, "What do exploratory searchers look at in a faceted search interface?," Jun. 2009, doi: https://doi.org/10.1145/1555400.1555452.
[12] D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: https://doi.org/10.1038/44565.

## Supplementary Material

TABLE III
IDENTIFIED TOPICS AND ASSOCIATED DOCUMENT FREQUENCIES

| TOPIC | DOCUMENT FREQUENCY |
|---|---|
| Electricity production | 90 |
| Experimental and Theoretical | 89 |
| Excess heat | 86 |
| product development | 76 |
| Lattice-enabled nuclear reactions | 75 |
| Nuclear renaissance | 75 |
| LENRIA | 70 |
| Alternative energies | 69 |
| Exciting New Science | 67 |
| prompt radiations | 67 |
| Low energy nuclear reactions | 65 |
| Ni-H system | 63 |
| Potential clean energy | 60 |
| Thermometer with a heat exchanger | 57 |
| IAC & ICCF | 54 |
| production of tritium | 54 |
| low energies | 54 |
| Liquid electrochemical | 52 |
| Energy conservation | 52 |
| Steam production and make-up water | 51 |
| Experiments and Prototypes | 51 |
| Thermometer in a closed loop | 51 |
| gases thermodynamic | 50 |
| Heavy water | 50 |
| ISCMNS | 49 |
| Energy flow | 48 |
| beams kinetic | 48 |
| Nuclear products | 48 |
| Pd-D system | 47 |
| Nuclear reactions using chemical energies | 46 |
| heat-helium correlation | 44 |
| natural gas | 44 |
| LENR thermal generator | 43 |
| production of heat: electrolysis of Pd in D20 | 41 |
| Power and energy densities | 40 |
| Thermal power and electric power | 38 |
| Condensed matter nuclear reactions | 38 |
| Little and big companies | 37 |
| Three dualities: Science, engineering and business | 36 |
| Plasmas kinetic | 35 |
| Units of quads | 34 |
| industrial association | 31 |
| New Hydrogen Energy | 30 |
| Fossil fuel | 28 |
| Low Energy Nuclear Trasmutations | 28 |
| LENR cities | 24 |
| HAD | 19 |
| LENR ecosystem | 19 |



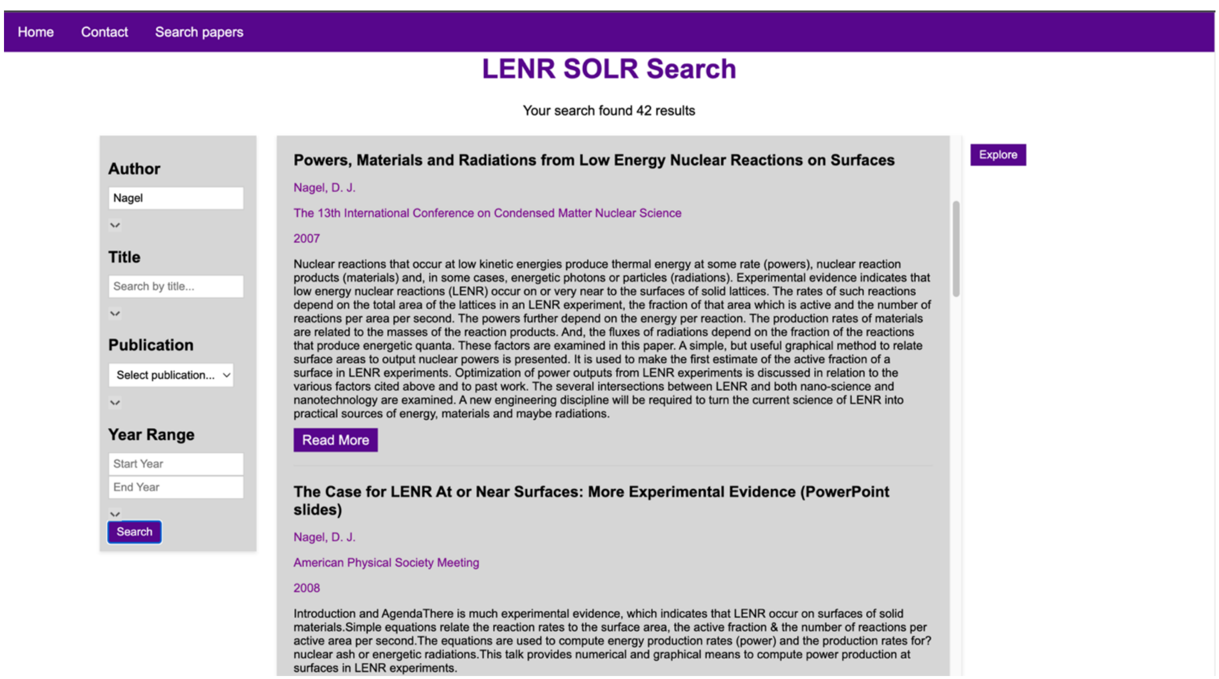




Fig. 8. Performing keyword-based search using Solr on the first filter "Search by Author" for input "Nagel"

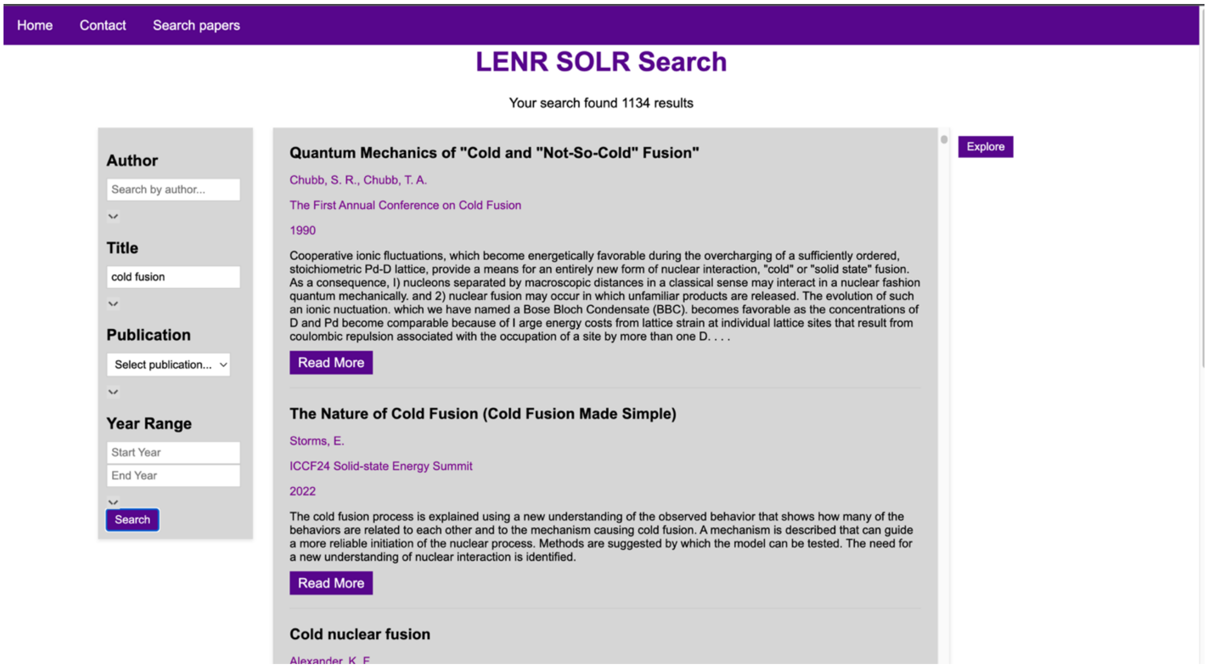



Fig. 9. Performing keyword-based search using Solr on the second filter "Search by Title" for input "cold fusion"

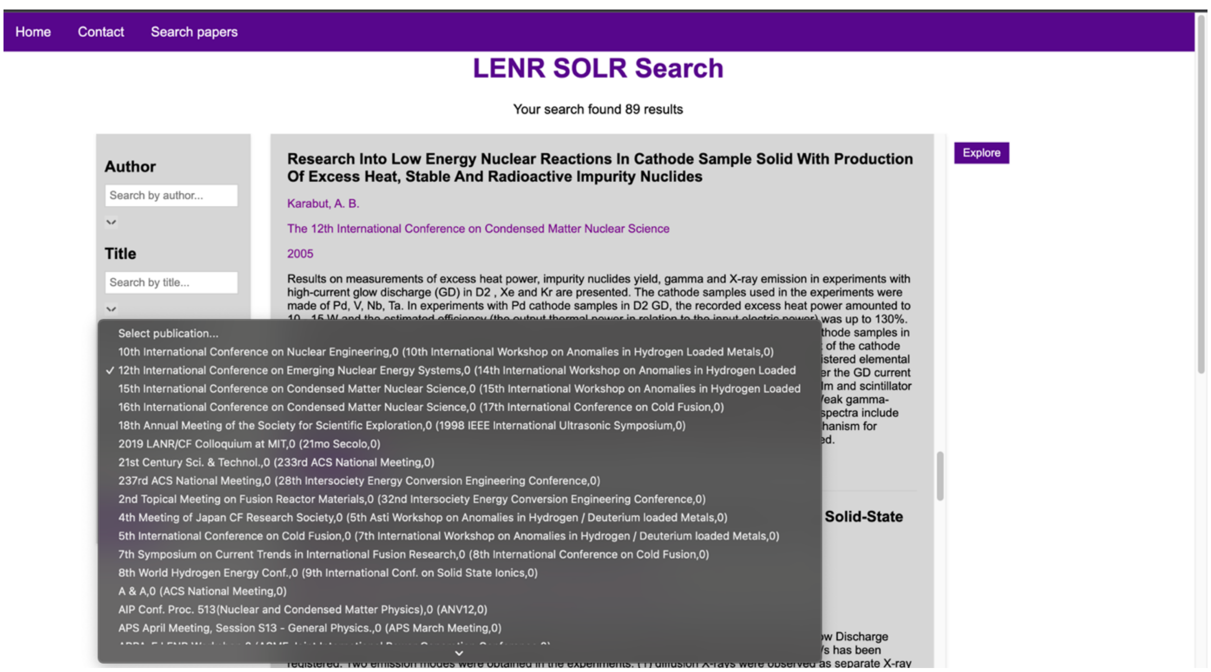



Fig. 10. Selecting the filter with the dropdown menu to "Search by Publication" and obtaining results in next figure

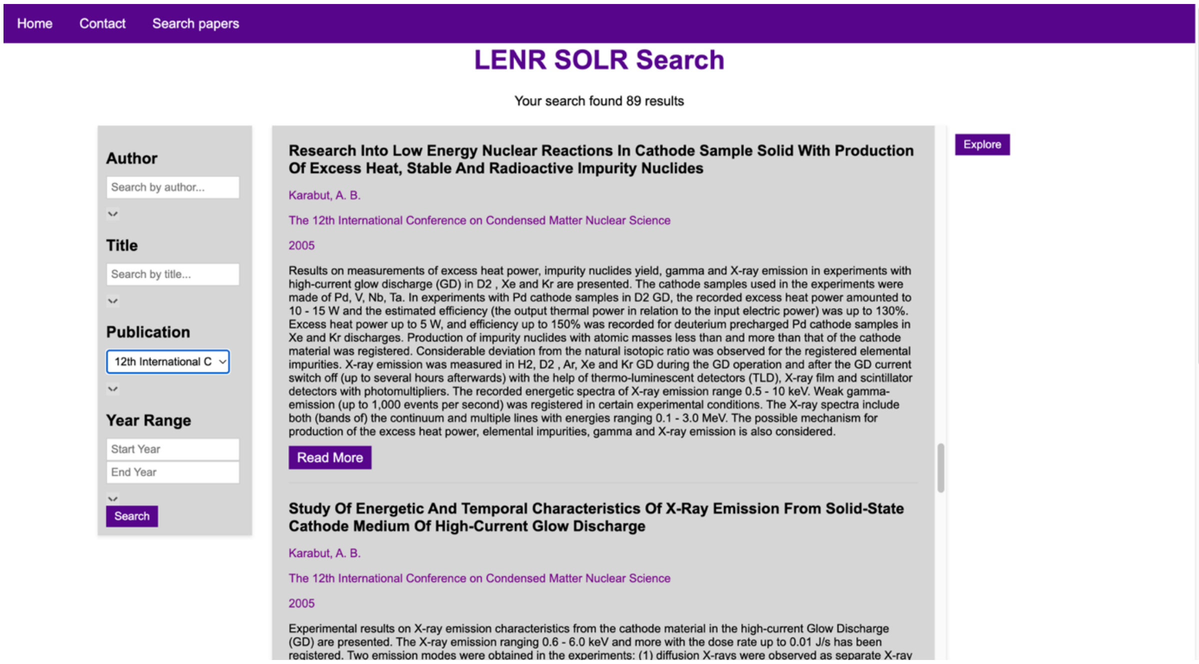



Fig. 11. Selecting the option “The 12th International Conference on Condensed Matter Nuclear Science”

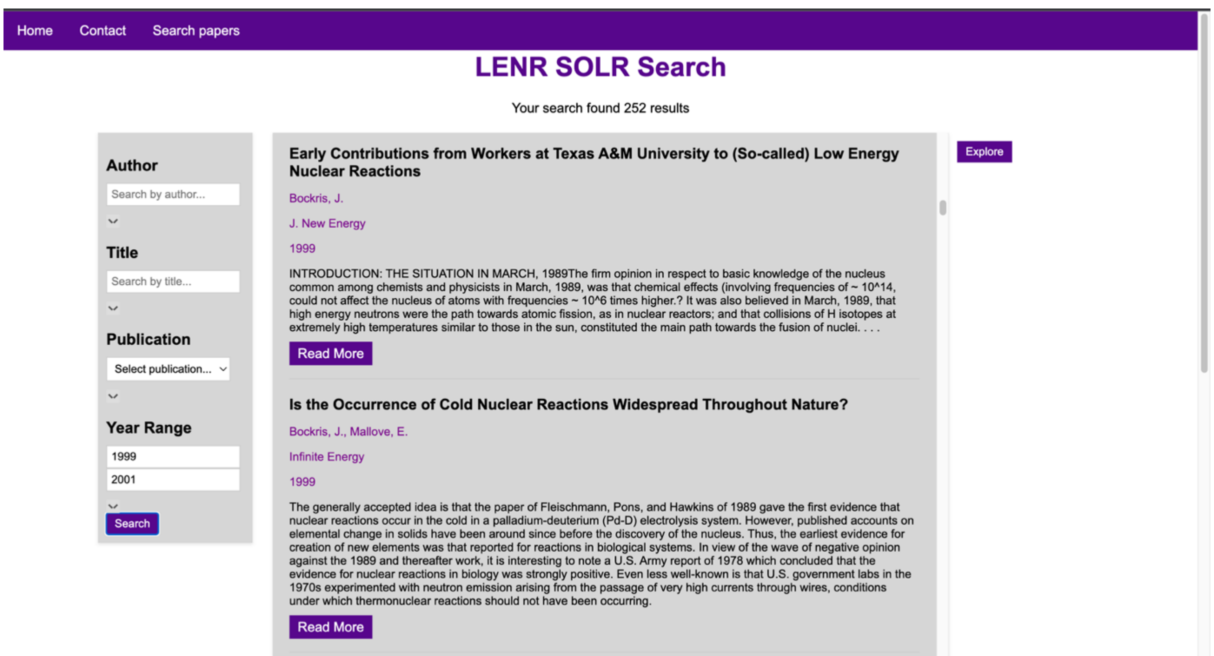



Fig. 12. Specifying a year range of 1999-2001 and obtaining the results

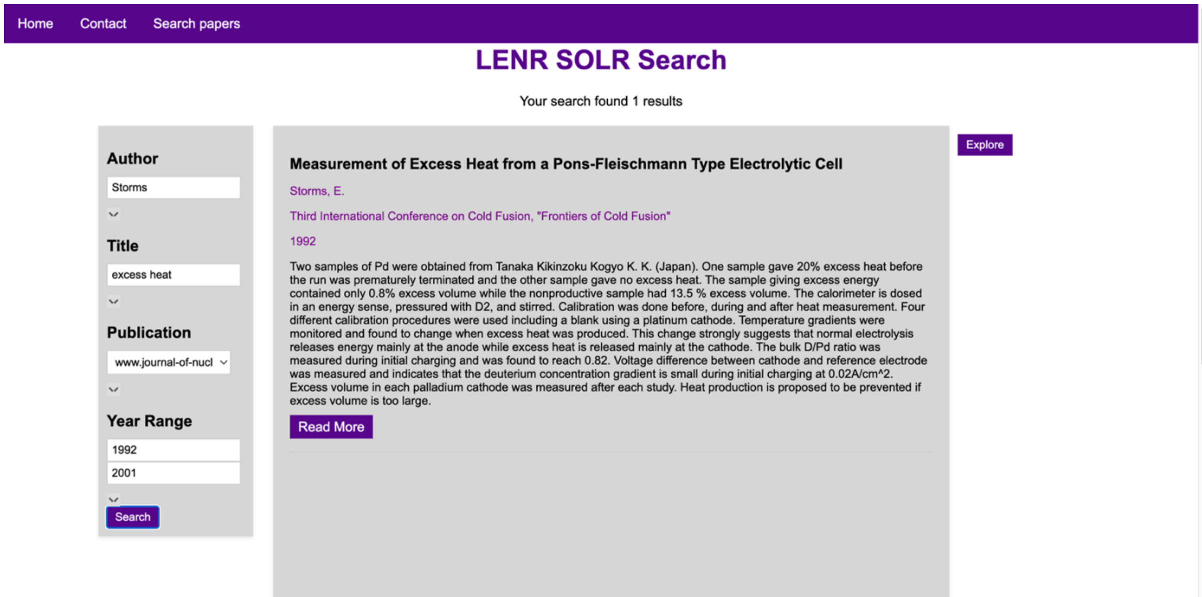



Fig. 13. Using all the filters simultaneously to obtain a specific paper by author "Storms", having title "excess heat" published in the journal "Frontiers of Cold Fusion" between 1992-2001

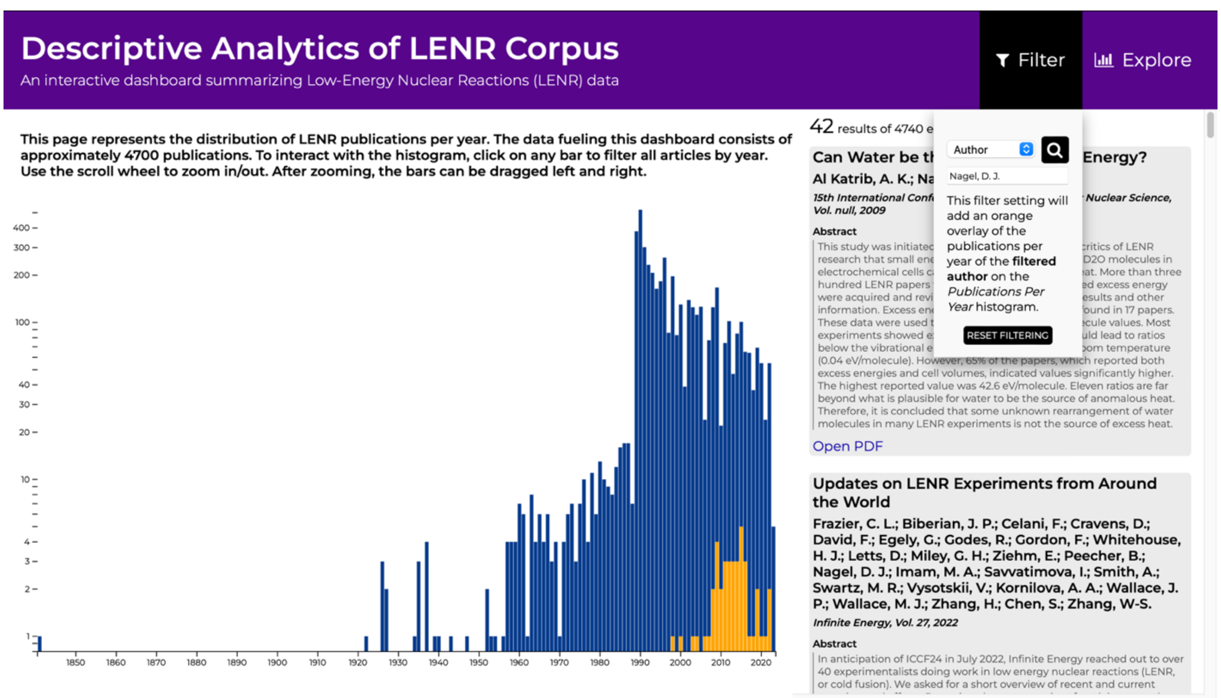



Fig. 14. Search Operation with visualizations based on Author Name for Input "Nagel D. J."

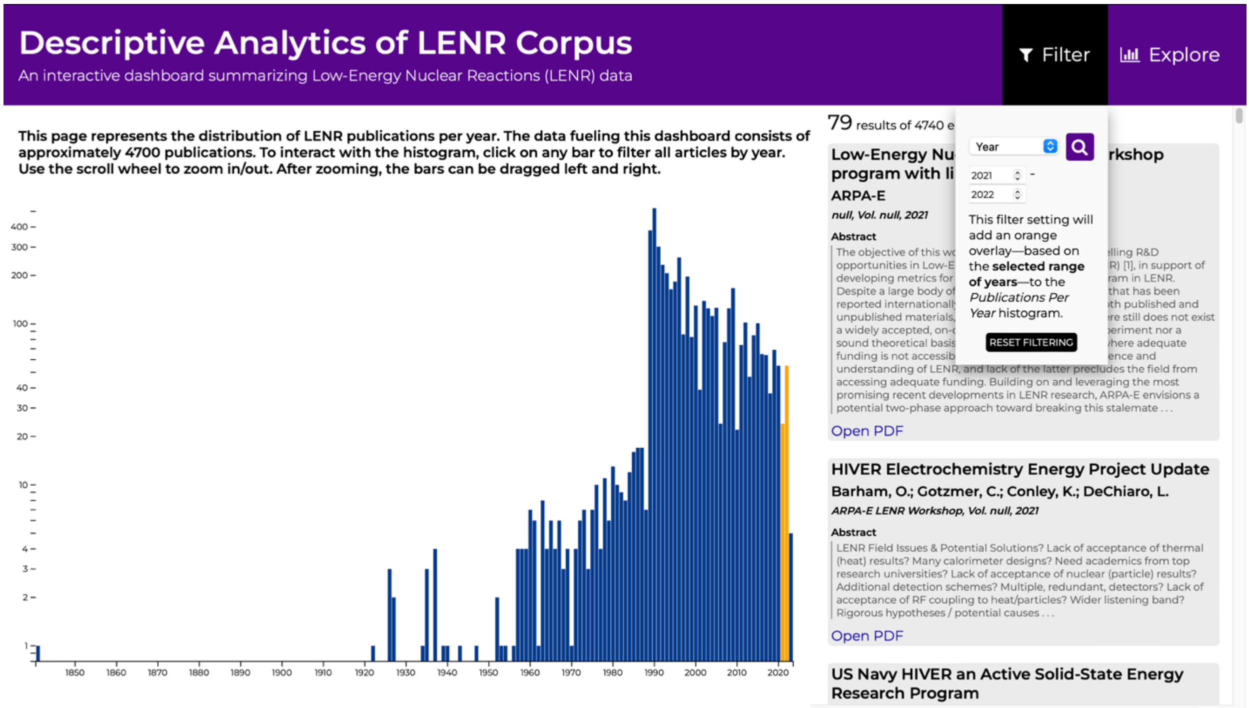



Fig. 15. Search Operation with visualizations based on Year for Input “2021-2022”