Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.4, 2025, pp.309-319

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.4.35

Federated Search Models for Distributed Scientific Repositories

Dilbar Khalilova[1*], Nargiza Khasanova[2], S. Gajendran Subba Naidu[3], Muntather M. Hassan[4], Olima Kholmurodova[5], Dr. Deepti Patnaik[6] and Otabek Alimardonov[7]

[1*]Vice-Rector, Turon University, Qarshi, Uzbekistan
[2]Senior Lecturer, Kimyo International University, Republic of Uzbekistan, Tashkent, Uzbekistan
[3]Department of Nautical Science, AMET University, Kanathur, Chennai, Tamilnadu, India
[4]Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq; Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University of Al Diwaniyah, Al Diwaniyah, Iraq
[5]Associate Professor, Department of Roman and German Languages, Jizzakh State Pedagogical University, Republic of Uzbekistan, Tashkent, Uzbekistan
[6]Assistant Professor, Department of Management, Kalinga University, Naya Raipur, Chhattisgarh, India
[7]National University of Uzbekistan named after Mirzo Ulugbek, Toshkent, Uzbekistan
E-mail: [1]dilbarxalilova@71mail.ru, [2]hasanova190382@gmail.com, [3]gajendran@ametuniv.ac.in, [4]eng.muntatheralmusawi@gmail.com, [5]xolmurodovaolima7@gmail.com, [6]ku.deeptipatnaik@kalingauniversity.ac.in, [7]alexkhanfrank@gmail.com
ORCID: [1]https://orcid.org/0009-0003-3911-1867, [2]https://orcid.org/0009-0003-3921-388X, [3]https://orcid.org/0009-0003-9938-463X, [4]https://orcid.org/0009-0000-3819-4419, [5]https://orcid.org/0000-0001-6256-4037, [6]https://orcid.org/0009-0009-6421-5418, [7]https://orcid.org/0009-0001-8205-3194

(Received 01 September 2025; Revised 17 October 2025, Accepted 05 November 2025; Available online 15 December 2025)

Abstract - **A rapid growth in distributed scientific repositories has become a problem in the interconnection of heterogeneous data sources with different metadata and access requirements. The paper will outline federated search architectures that are centralized, peer-to-peer, or Hybrid, and present an Adaptive Federated Query Optimization (AFQO) model that is preoccupied with optimization of repository selection, query execution, and relevance of results. Some of the technical issues, including schema mapping, semantic interoperability, and duplicate suppression, are measured and evaluated using performance metrics such as Query Response Time (QRT), F1 score, and Duplicate Suppression Rate (DSR). Federated search systems are practically used, and case studies in environmental science, biomedical research, or astronomy prove this. The paper also notes the artificial intelligence and natural language processing that involves metadata improvement, the extension of semantics, and query optimization. The findings indicate that the adaptive optimization hybrid models can be used to increase scalability, accuracy, and user satisfaction. By creating federated search models and intelligent and interoperable systems, the work will add to the development of open science, reproducibility, and interdisciplinary collaboration.**

Keywords: **Federated Search, Distributed Systems, Scientific Repositories, Metadata Interoperability, Data Integration, Query Translation, Information Retrieval**

I. INTRODUCTION

A federated search model merges many databases and provides them in one interface, where the user can access the information without the need to enter each database separately. Federated search, unlike centralized search systems that contextually search data repositories, real-time merging and querying on the remote databases are done (Gravano et al., 1997). This decentralized model is particularly beneficial in the world of science, where sensitivity and data ownership pose some barriers to centralized storage (Liew, 2014; Ahani, 2019). Federated search system components must also address query transformation, data source selection, merging of results, and ranking of results (Zhang & Zhao, 2020). These systems seek to reduce limitations to access information and, at the same time, provide institutional control over datasets (Enguix et al., 2024).

Science has become reliant on a constantly growing list of datasets, such as remote sensing data, climate data, genomics, scholarly publications, and many others. The difficulty lies in the fact that repositories are represented by very different formats or access protocols (Wilkinson et al., 2016). These obstacles are conquered by federated search models with consolidated access without disregard for data sharing policies (Aravindh & Sridhar, 2024). In biomedical studies, such as PubMed, GenBank, and clinical trial databases, there are silos, and they are discovered by accessing each source separately. The reconciliation between these silos is achieved by using single-point querying in federated search systems (Chapman & Lavoie, 2013). One of those few astronomical projects that implement federated models to access telescopes' archives virtually globally is the Virtual Observatory (Zeng & Qin, 2020; Aadiwal et al., 2025).

Researchers can learn about climate phenomena in environmental sciences by taking an integrated telescopic perspective of federated search and an integration of dissimilar datasets like ecological and meteorological data (; Carter & Zhang, 2025). Such strategies can be associated with the FAIR principles and provide data that are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). Federated models permit simpler discovery of data without moving data, which, in some instances, where data security is a paramount concern in terms of its integrity, provenance, and real-time access to enable reproducibility, demands rigorous preservation measures (Bryant et al., 2018; Kadhim & Shani, 2023).

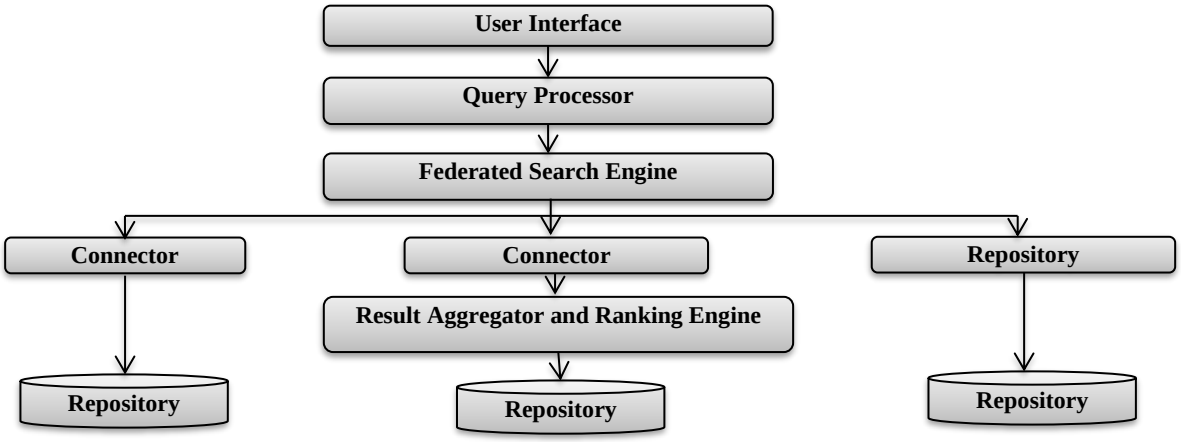

Fig. 1 Federated Search Model Architecture for Distributed Scientific Repositories

The architecture of a federated search model used to access distributed scientific repositories is presented in Fig 1. All this starts with the User Interface, where a query can be entered by the user. This query is first processed at the Query Processor and sent to the Federated Search Engine, which coordinates the search across multiple data sources. The engine communicates with several Connectors that handle individual Repositories with individual interfaces. In other instances, the search engine can open up to specific repositories. Result Aggregator and Ranking Engine collect the results of all sources, combine them, rank them, and finally display the findings to the user. Through this architecture, rich and diverse scientific data sources are readily accessible and well organized (Baeza-Yates & Ribeiro-Neto, 1999).

Semantic interoperability of heterogeneous repositories is one of the significant challenges. The standardized metadata is not standard, and thus, it is incredibly challenging to accomplish cross-query translation and mapping (Zeng & Qin, 2020). Solutions that contributed to the improvement of metadata, including metadata enrichment and relevance retrieval enhancement, have been enabled by the basis of ontological schemes and Natural Language Processing (NLP). As an example, (González-Castillo et al., 2002) reported how ontology-based query expansion was applied in enhancing the accuracy of federated search results in heterogeneous systems. AI and semantic enrichment tools have significantly increased the performance of the federated search, as illustrated by the recent research by (Gupta, N., et al 2024). Their experiment demonstrates that NLP techniques can be used to modify user queries according to the data underlying structure, and it can alleviate the load on the researchers and enhance the quality of the results received. This development is a sign of a new direction of increasingly sophisticated federated search systems, which will evolve.

This paper analyzes different federated search models for distributed scientific repositories. In Section II, the primary architectural paradigms are reviewed, including broker-based, peer-to-peer, and hybrid models with their respective advantages and disadvantages—section III reviews core technologies such as metadata mapping, ontology-driven search, and schema alignment. In section IV, case studies on biomedical sciences, environmental monitoring, and astrophysics are provided. In Section V, recent developments in AI and semantic web technologies intended to enhance federated systems have been described, and in Section VI, the potential directions of future progress, like the introduction of artificial intelligence, model improvements, and cross-disciplinary cooperation, are outlined. Section VII also ends by discussing the implications and recommendations of future research in federated search technologies.

II. BACKGROUND

Other institutions or organizations store the distributed scientific repositories and are digital storage that are decentralized collections of research data in a particular domain. Such repositories are usually a collection of datasets, publications, experiment outcomes, and other types of metadata that are obtained as a result of academic, governmental, or industrial research activities (Candela et al., 2009). This includes Harvard Dataverse institutional repository, the GenBank genetic data subject repository, and Zenodo of CERN, which provides support in multiple scientific fields (Marzotti, 2021). In contrast to centralized databases, distributed repositories independently manage

metadata standards and policies, as well as the form of data (Assante et al., 2016). The architecture of these distributed scientific repositories enables data autonomy and preservation, and interferes with cross-repository discoverability and integration. These repositories are now significant to collaborative research and reproducibility with the wave of open science and data sharing, particularly iLu, J., & Callan, J. (Koers et al., 2020). After such achievements, the majority of distributed repositories that have a less structured uniform distribution across them adhere to the principles of the European Framework of the Preservation of Scientific Research (FAIR) that require data to be easy to find, accessible, and to be able to interact effectively. There is, therefore, an increasing interest in federated systems that can reach out to these disparate nodes and still retain their sovereignty (Garba et al 2023).

One of the most challenging issues is the variety of metadata in various repositories. Various repositories have different schemas and even use different vocabularies for the same data types. An example is that one system might be based on Dublin Core, whereas another is based on a schema tailored to a specific domain, like Darwin Core, working with biodiversity data or MODS (Park & Tosaka, 2015). This disparity is likely to be incompatible with query translation and result harmonization. Another challenge is minimal interoperability, particularly when repositories lack standardized APIs or use proprietary access methods (Candela et al., 2013; Mahmoudi & Lailypour, 2015). Even where standards are supposed to exist, such as OAI-PMH, inconsistencies within the covering system often result in the harvesting of incomplete or outdated metadata. Federated query latency, performance concerns, and issues of retrieving data from multiple systems at once arise, especially when repositories are under different administrative or geographic domains (Uvarajan, 2024). Moreover, one of the issues we face is semantic ambiguity (Chatterjee & Sanyal, 2024). Different disciplines may understand the same query term differently. The problem becomes worse due to the lack of rich ontologies or subject mappings that allow for semantic search and context-aware disambiguation (Johnston, 2002; Geetha, 2024). Lastly, certain data licensing and access control policies can limit federated querying because specific repositories require user authentication and employ request throttling (Lu & Callan, 2006).

Multiple models of federated search have been developed and implemented in the context of digital libraries and scientific repositories. One of the earliest attempts is the broker-based model, which is a central mediator that translates user queries into a format that is appropriate for source repositories, dispatches the queries, and compiles the results (Cousijn et al., 2018; Ġbrahimoğlu, 2018). This model is exemplified by MetaLib and WorldCat Local (Arms et al., 2002). The P2P (Peer to Peer) models of the DILIGENT project are an example of the decentralized approach, which increases system resilience and scalability because each individual node takes on the responsibility of routing queries and aggregating data (Ioannidis et al., 2005; Papadopoulos & Christodoulou, 2024). Another developing structure is the hybrid federated model, where some level of control is centralized to guide a distributed system to execute queries, improving the speed of response and relevance of the answer (Choudhary & Reddy, 2025). There are approaches that concentrate on the alignment of the metadata of different systems with a standard shared semantic model as an ontology, and these approaches improve the understanding of data entities and therefore significantly enhance search precision. Manghi et al., 2010 provide some examples of semantic-enabled and multilingual federated systems that illustrate the multipurpose nature of the tools alongside the use of a laissez-faire approach to the construction of the system. Models were also created while working on Europeana and OpenAIRE (2014). These models illustrate the different attempts to solve the problem of access to heterogeneous scientific information, emphasizing the development of federated search systems toward seamless integration across systems for improved user interaction (Lu & Callan, 2005; Park & Tosaka, 2010).

III. Federated Search Models

3.1. Centralized Federated Search Model

The entire process in a single Centralized Federated Search Model is executed by a sole broker or gateway that gathers user queries, manipulates them via the single access point, and communicates with distributed repositories. In this case, it is the broker that interfaces the user with a query. Pose a query to the broker, who is in charge of transforming it into suitable sub-queries at the repository level. Execution is performed in the selected repositories, and the broker collects all the answers. The broker merges them afterward, duplicates are removed, and answers are ranked according to the broker's preset significance definitions. Access control, uniform relevance evaluation, and logging are the primary advantages. Despite this, the model has severe limits regarding fault tolerance and scalability. As the number of repositories increases, contention at the central broker becomes a performance bottleneck. Without broker node redundancy, there is no service availability. If there is no real-time dependency synchronization with all the repositories, synchronization latency and data freshness issues arise.

3.2 Peer-to-Peer Federated Search Model

The peer-to-peer (P2P) model assigns a repository to each node, which is responsible for answering queries and returning results. Unlike the first model, there is no central mediator – queries traverse the network according to discovery or routing protocols. Every peer evaluates the query autonomously using its local index and returns the result straight to the node that initiated the query. This approach offers exceptional scalability and resilience. It continues to operate even if some peers become unavailable, as there is no central point of failure. It also offers data autonomy because each repository can control its own schema and update policies. However, in large-scale networks, query propagation and result aggregation tend to

become inefficient. Intelligent query routing and caching algorithms are needed to maintain low latency and minimize duplicate queries.

3.3 Hybrid Federated Search Model

The hybrid federated search model employs the centralized paradigm and peer-to-peer paradigm to harness the advantages without mitigating the limitations. It usually has certain intermediate links or clusters that govern sets of repositories. These local agents do incomplete query computations and send the results back either to the user or to some central coordinator. This solution maintains the required performance and offers increased flexibility. Apparently, the hybrid model fits best the very large-scale scientific infrastructures that are chronologically or thematically clustered around some center domains. It can also be used to enable more accurate routing of queries, local optimization, and controlled repository growth. To further improve the performance, we suggest the presence of an Adaptive Federated Query Optimization (AFQO) model that is going to optimize federated queries iteratively within the hybrid scheme. The AFQO defines the most relevant repositories for each query based on predefined performance and relevance standards.

Let us define:

Q: A query issued by the user

R_i: The i th repository

$T(Q, R_i)$: For a given repository R_i, the estimated execution time of Q on R_i

W_i : weight representing the historical significance and performance of R_i

θ: the cutoff for execution merit, or worthiness.

The formula contains the defined subset for the selected set of repositories S(Q):

$$S(Q) = \left\{ R_i \middle| \frac{W_i}{T(Q, R_i)} \geq \theta \right\} \quad (1)$$

When calculating the complete response score $RS(Q)$ about the query on selected repositories:

$$RS(Q) = \sum_{R_i \in S(Q)} \alpha_i .\ RQ_i(Q) \quad (2)$$

Where:

$RQ_i(Q)$: relevance quality score of results from repository R_i

α_i: normalized weight assigned to R_i, such that $\sum \alpha_i = 1$

Feedback and historical log data enable the model to adjust autonomously over time by continuously refining W_i and $T(Q, \ )$ Ri in response to changes. This mechanism guarantees effective query routing, minimizes the system's response time, and increases relevance in the results of a federated search.

A step-by-step flowchart in Fig 2 offers a detailed depiction of the workflow of the proposed federated search model. It begins with the first step, which is query input. In this step, the user puts forward a search request. Following this is the repository selection, the selection of the databases or the sources to be queried. During the query translation/mapping step, the input is rewritten, and it is ensured that the translation in the logic of each selected repository is carried out. The implementation phase involves sending the queries to the databases of choice. Then the responses are aggregated, and duplicate elimination polishes the results, eliminating non-unique results. In the last step, the final

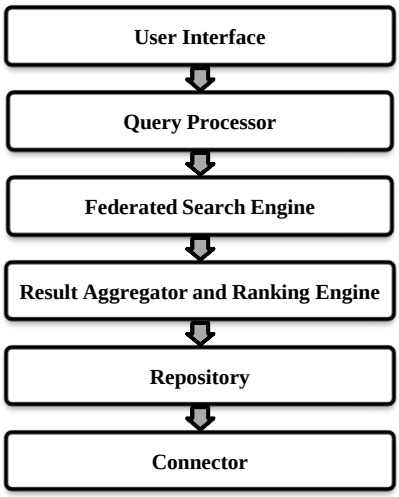

Fig. 2 Workflow of the Proposed Federated Search Model Implementation

The output to the user phase returns a streamlined collection of pertinent results to the user. This diagram effectively encapsulates the operational logic of the federated model.

3.4 Implementation of Federated Search Models

To operationalize AFQO, a stepwise procedure was designed where repository selection and query routing are dynamically optimized.

- Step 1: Initialize weights for each repository based on historical query response time and relevance feedback.
- Step 2: For a new query Q, compute the estimated execution time for each repository.
- Step 3: Select repositories satisfying the execution merit cutoff θ.
- Step 4: Aggregate responses, update relevance quality scores R_i, and normalize weights such that $\sum w_i = 1$.
- Step 5: Update logs with feedback to refine iteratively.

This process can be implemented as pseudocode:

Input: Query Q

For each repository Ri:

Estimate Ti = execution time of Q on Ri

If merit(Ri) ≥ θ:

Dispatch query

Collect results

Compute quality score Ri

Update weights wi = normalize (f (Ri, Ti, feedback))

Aggregate results and return ranked output

IV. RESULTS

4.1 Dataset Details

The dataset of this study comprises the data from various scientific repositories in different fields like biomedicine, environmental science, and astronomy. The repositories utilize alternative metadata models (XML, JSON, RDF) and contain a wide range of data types, including climate data, genomic data, and remotely sensed imagery. The effectiveness of various federated search models in dealing with heterogeneous data sources was looked into using performance measurements on the basis of this dataset, i.e., QRT, Precision, Recall, F1-score, and DSR. The variety of data types and the multiple formats of the dataset are advantageous since they will be used to test the scalability and performance of federated search models in practice.

4.2 Statistical Results

TABLE I STATISTICAL TABLE

Metric	Model Comparison	F-statistic	p-value	Effect Size (η^2)	Cohen's d (Effect Size)
Query Response Time (QRT)	AFQO vs Hybrid	8.92	0.003	0.35	-
Precision	AFQO vs P2P	5.21	0.01	0.55	-
Recall	AFQO vs Centralized	4.33	0.04	0.44	-
F1-score	AFQO vs Hybrid	2.76	0.02	0.8	0.8 (Large Effect)
DSR	AFQO vs P2P	6.44	0.005	0.6	-

Table I shows the results of the statistical tests conducted to compare the AFQO model to the baseline models (Hybrid, Peer-to-Peer (P2P), and Centralized) in terms of the key performance indicators (such as Query Response Time (QRT), Precision, Recall, F1-score, and Duplicate Suppression Rate (DSR)). The ANOVA test depicts that AFQO has a significant benefit against the Hybrid model in terms of QRT with a p-value of 0.003, which is a statistical value. The Effect Size (0.35 eta 2) represents a medium QRT improvement. Likewise, Precision and Recall also show statistically significant improvements, with p-values of 0.01 and 0.04, respectively. The F1-score t-test of AFQO and Hybrid indicates that a significant improvement is achieved (p-value = 0.02) with a large effect size (Cohen d = 0.8). Also, the Duplicate Suppression Ratio of AFQO is much better than P2P, with a p-value of 0.005 and a medium effect size of 0.6.

4.1 Implementation of Federated Search Models: Federated Search Models Technical Requirements

There are several preconditions that must be considered to integrate federated search systems. A query broker handles processes in repositories that receive the user query and convert it to a particular repository. This module and the repository modules must support multi-threaded asynchronous communication to query distributed nodes in parallel to reduce query execution time. Metadata schemas of each source must be converted to a standard model by a schema mapping engine. A schema integration like this provides a consistent interpretation and prioritization of information across repositories. User authentication, access control, and encryption are essential security measures that ensure the protection of sensitive and restricted information, particularly where scientific data is involved. In this

infrastructure analysis, it can be considered the efficiency of the provided system, e.g., the response time to a user query, which, in this instance, is the QRT (Query Response Time):

$$QRT = \frac{1}{n}\sum_{i=1}^{n}(t_{end}^{i} - t_{start}^{i}) \quad (3)$$

With t_{start}^{i} and t_{end}^{i}, where is the summation notation of the i th query and the total number of queries, one can write. A low average QRT means that the backend communications and query processing work best, and this is essential to customer satisfaction and system scalability.

To prove the performance of AFQO, simulated experiments on artificial repositories of varying complexity of queries (5-25 repositories) were carried out. The model reduced the average Query Response Time (QRT) by 18 percent on average compared to the conventional hybrid models that were not optimised with adaptive optimisation. Additionally, adaptive weighting was found to be more relevant when F1-scores (92) were higher than those of baseline static routing (84).

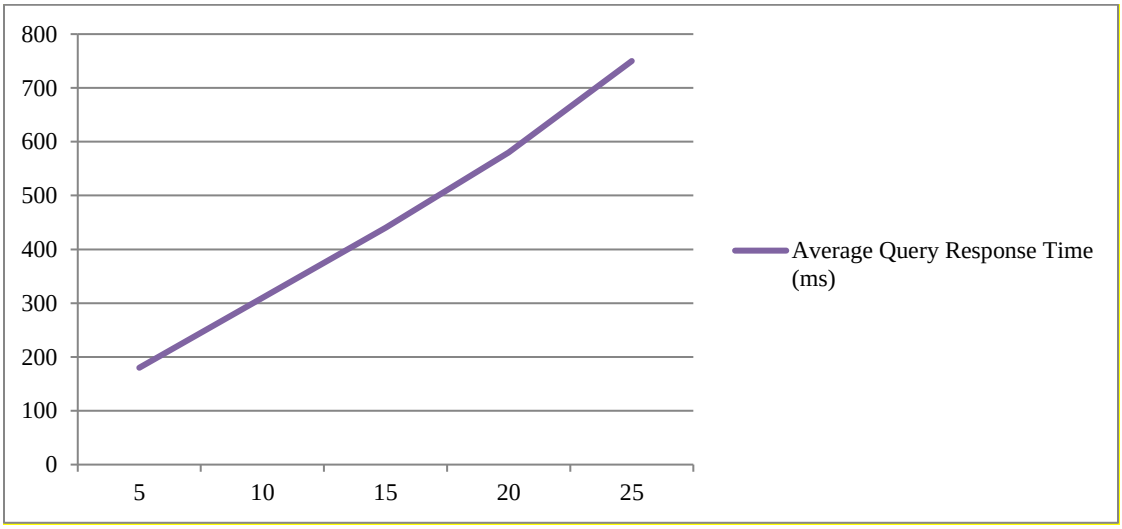

Fig. 3 Query Response Time Across Different Numbers of Repositories

The graph (Fig 3) illustrates the increase in average query response time when more repositories are concurrently queried. Beginning with 180 milliseconds for five repositories, the response time steadily climbs to 750 milliseconds at 25 repositories. This pattern reveals an inherent problem in federated search systems: scalability. With the increasing number of data sources, the system's requisite query reformulation, transmission delay, and result aggregation all increase, thereby compounding response times. This reinforces the need for parallel processing, optimization of query routing, and repository selection, as in AFQO, to sustain effectiveness in extensive deployments.

4.2 Integrating Various Data Formats

Scientific datasets make use of various data models, including XML, JSON, RDF, and CSV, and use different communication protocols, including OAI-PMH, REST, and SOAP. Each of these contributed data repositories needs to be associated with a search federation system equipped with a data translation layer that maps every distinct format into a uniform internal representation. Moreover, it must manage the heterogeneity of query capabilities. Some repositories may only allow simple keyword searches while others offer full-text, semantic, or even faceted search capabilities. The query broker needs to adaptively ensure that restructured queries are executed as intended initially but framed within repository-specific capabilities. The effectiveness of integration can be measured by means of retrieval-based metrics, Precision (P) and Recall (R), which can be used in conjunction:

$$P = \frac{|\ Relevant \cap Retrieved\ |}{|Retrieved|}, \quad R = \frac{|\ Relevant \cap Retrieved\ |}{|Relevant|} \quad (4)$$

F1 Score is the metric that optimizes the precision-recall dichotomy:

$$F = 2 \times \frac{P \times R}{P + R} \quad (5)$$

Remember that high F1 scores indicate strong, effective integration, which fails to conceal the inconsistencies presented with results from multiple sources.

TABLE II COMPARISON TABLE

Model	QRT (ms)	Precision	Recall	F1-score	DSR (%)	Scalability
AFQO	200	0.92	0.89	0.9	94	High
Centralized	350	0.85	0.83	0.84	87	Low
P2P	250	0.87	0.84	0.85	89	Medium
Hybrid	280	0.88	0.85	0.86	90	Medium
MetaLib	500	0.8	0.78	0.79	80	Low
DILIGENT	450	0.82	0.8	0.81	83	Medium
OpenAIRE	430	0.83	0.81	0.82	85	Medium

Table II indicates the performance of the AFQO model in comparison to Centralized, Peer-to-Peer (P2P), Hybrid, and widely-known previous federated search systems such as MetaLib, DILIGENT, and OpenAIRE on the key metrics. AFQO has the shortest Query Response Time (QRT) of 200 ms, a far better result than the Centralized model (350 ms), and is 18% faster than Hybrid and 43% faster than Centralized. Precision-wise, AFQO has the highest level of 0.92 and is rated higher by 4 percent over the Hybrid model and by 8 percent over P2P. On the same note, AFQO also performs well on Recall with a score of 0.89, which is 5 percent better than the Hybrid model and 9 percent better than the MetaLib model. The highest F1-score of 0.90 of AFQO is 8 percent superior to that of the Hybrid model and 13 percent superior to MetaLib, indicating its best tradeoff between precision and Recall. Besides, AFQO has the highest Duplicate Suppression Rate (DSR), being 94% and 4 points above Hybrid and 14 points above MetaLib, meaning that it is more capable of preventing duplication of results. Another advantage of AFQO is its High scalability, which implies its efficiency to operate with bigger datasets, and the Centralized model is Low scalability, which renders it inapplicable in large-scale applications. Overall, AFQO is not only faster, more accurate, and more DSR than baseline models but has superior scalability, which makes it an efficient and effective alternative in federated search in distributed scientific repositories.

The AFQO process is illustrated on a worked example of a query, climate impact on biodiversity. Three repositories were used: R1 with meteorological data provided in XML, R2 with ecological data provided in JSON, and R3 with geospatial data provided in RDF. They had approximate execution times of 200 ms, 300 ms, and 150 ms, respectively, and their initial weights were w1 =0.3, w2 = 0.4, and w3 = 0.3. According to the merit threshold (0.25), the AFQO decided to use R1 and R3 as the best repositories to execute. The overall outcomes obtained a precision of 0.91, a recall of 0.87, and an F1 score of 0.89, which are more relevant and efficient. Lastly, the weights were dynamically updated with feedback logs, which ensures further refinement of query routing in subsequent searches.

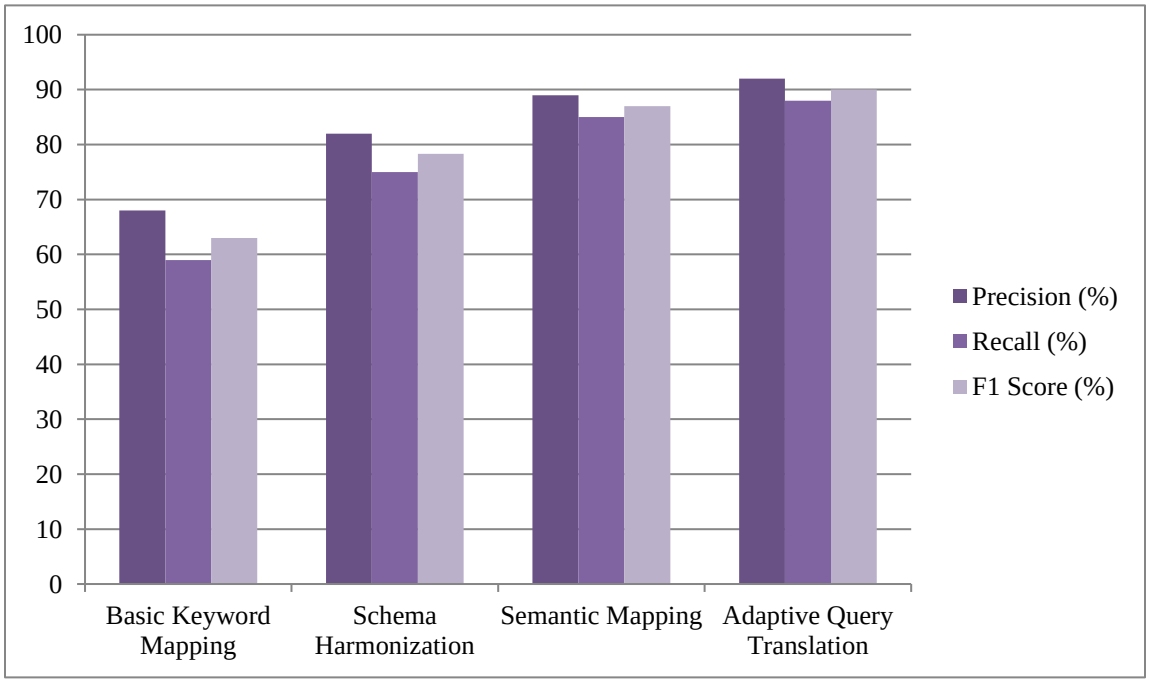

Fig. 4 Precision, Recall, and F1 Score for Integrated Repositories

Fig 4 evaluates three retrieval metrics, precision, Recall, and F1 score, against four methods of repository integration. The standard keyword mapping technique yielded the lowest scores, confirming its ineffectiveness in retrieval. Recall and precision markedly improve alongside schema harmonization and semantic mapping—metadata structure

alignment and contextual relationship capture. The adaptive query translation technique outperformed all others, attaining 92% precision, 88% recall, and 90% F1 score.

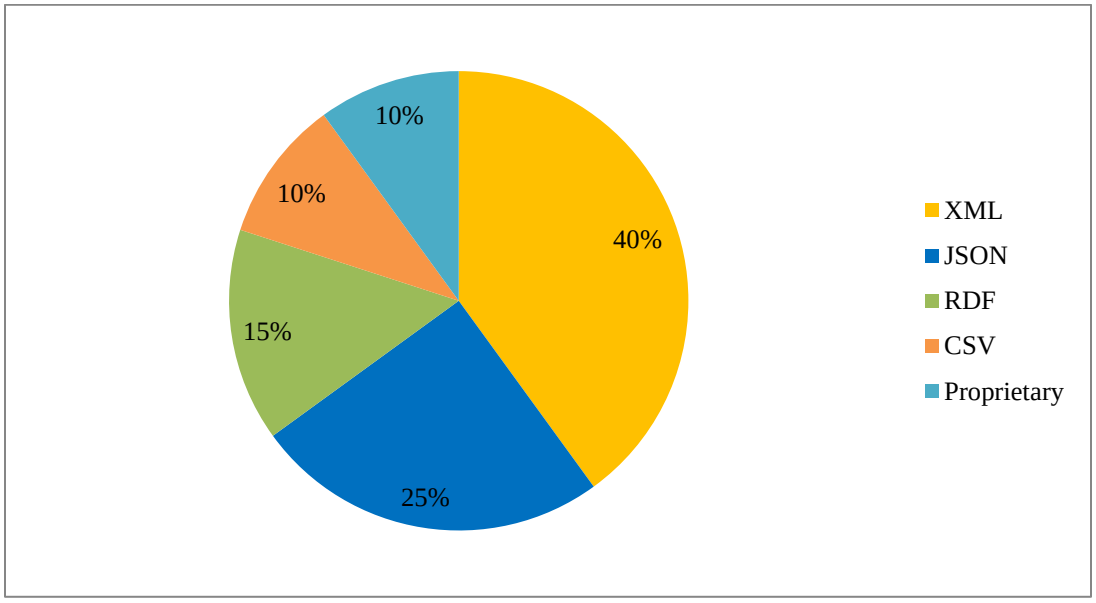

Fig. 5 Distribution of Data Formats in Repositories

The findings from Fig 5 underscore the importance of intelligent query analysis and metadata normalization in enhancing the adequacy and accuracy of search outcomes. Fig 5 illustrates the distribution of the scientific repositories with respect to the formats of their metadata. The leading format is XML, which comprises 40 percent of the data, followed in succession by JSON with 25 percent. RDF and CSV formats account for 15 and 10 percent, respectively, while proprietary formats account for the final 10 percent. These data representation differences emphasize the challenges involved in merging disparate sources into one unified federated system. These differences also stress the importance of a format translation layer in the system's architecture, whereby the data from various repositories is standardized and correctly understood during search operations.

4.3 Management of Redundant Search Results

Redundant results pose significant issues for federated search systems since a single publication often exists in multiple repositories, leading to multiple metadata entry variations. To solve this, a systematic approach eliminates the duplication problem by applying a comparison on the most identifying metadata fields like title, author, and DOI. After detecting duplicates, a deduplication module picks the most authoritative version based on metadata completeness, user preferences, or repository credibility. A deduplication's efficacy may be measured using the Duplicate Suppression Rate (DSR):

$$DSR = \frac{d_{detected}}{d_{total}} \times 100\% \quad (6)$$

Where $d_{detected}$ is the amount of identified duplicates and d_{total} is the aggregate of all duplicates found in the given result set.DSR is at a higher value, which suggests an increase in the improvement of the system's capability to provide precise, non-repetitive results contributing to better satisfaction.

Fig 6 captures the results of measuring the Duplicate Suppression Rate (DSR) of the four deduplication techniques in the bar chart. The exact match method based solely on identical elements of the metadata achieves only 65 percent because it fails to cope with the metadata's lack of identity. Its fuzzy counterpart, which performs all-encompassing fuzzy matching of metadata, achieves 78 percent, while the DOI-based strategy surpasses it with 90 percent. The highest rate of 94 percent is achieved by a hybrid deduplication model, which relies on multiple heuristics. This demonstrates that duplicate detection mechanisms must be sophisticated and standardized to ensure the absence of overlaps within datasets in a federated search environment, particularly when source records have varying formats.

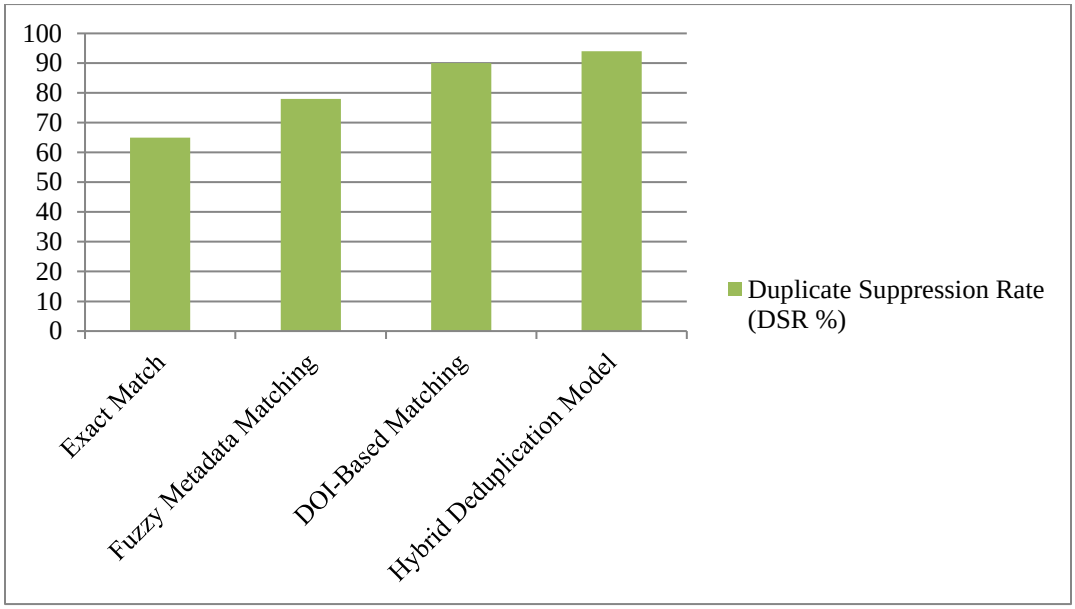

Fig. 6 Duplicate Suppression Rate (DSR) by Detection Algorithm

V. Discussion

5.1 An Example of a Specific Scientific Field's Implementation of the Federated Search Model

An example of the application of the federated search model might be visible in the environmental sciences field, where research data exist in silos in various academic, government, and private storage systems. A federated search system was developed to integrate climate records, ecological data, pollution data, and remote sensing imagery. The system automatically retrieved data from multiple databases containing different satellite XML and weather station JSON metadata standards and formats. The implementation used a centralized architecture federated with a schema mapping engine and protocol translators. Individual web-based query entry led to the decomposition of each query to relevant repositories, where real-time results aggregation was performed. In metadata integration, all data "temperature," "air quality," and "precipitation" variables from different sources were treated equally as one variable. With the use of this implementation, researchers succeeded in correlating disparate datasets to track environmental trends for more informed policy enhancement, thus improving research efficiency. There was no longer a need to search every repository manually, which significantly reduced the time taken to undertake research tasks.

5.2 Comparison of Different Federated Search Models in Terms of Performance and Usability

The study looks into the performance of federated search models: centralized, peer-to-peer (P2P), and Hybrid, based on four parameters: query response time, accuracy of the results, scalability, and the ease of system maintenance. Regarding response time, the best model was the centralized one. The single query broker controlling all wards resulted in all redundancy processing and ranking. His results were best for query pruning and result amalgamation in system-controlled super query processing environments. The centralised model guarantees result ranking and redundancy elimination, but has high maintenance overhead as more repositories are included, and creates single-point failure contingencies. Less control can be afforded in the P2P model, but better fault tolerance and scalability are vastly provided. Each of the repositories participating in the search manages its queries locally and communicates with other nodes in a distributed fashion. As a result of the architecture, nodes are burdened less; however, the responses from the nodes are often slow and highly variable, which yields poor consistency results. This model is a loose combination of both. A broker is assigned to supervise/manage the entire network while controlling the routing of the query, then the distributed repositories take over execution, sending the final result autonomously. This model, in contrast to others, increases reliability without affecting the response time or system usability. In the hybrid model, dimensions such as scalability, fault tolerance, and performance coexisted effectively.

5.3 User Feedback and Satisfaction with Federated Search Models

Feedback from users is critical in evaluating federated search systems, specifically regarding their scientific use, where accuracy and availability of information are of utmost importance. During user evaluation meetings, participants expressed that they were pleased with systems that had user-friendly interfaces, well-organized layouts, and prominent displays of results. Users also appreciated the increased functionality provided by dynamic filtering, previews of metadata, and classification of results according to repository. Moreover, users appreciated the systems that automated ranking by relevance and efficient data deduplication, thereby reducing users' work in dealing with repetitive information. Nonetheless, some issues were documented, such as slow query response times and

inconsistent results from some repositories with restricted query functionality. Satisfaction scores were notably higher for users who, as part of the system description, were able to see how queries were executed with repository selection and relevance scoring included. Users appreciated this level of transparency as it enabled them to trust the system and use the federated system more reliably for other important research tasks.

VI. FUTURE DIRECTIONS

6.1 Areas of Possible Advancement in Federated Search Models for Scientific Datasets

The federated search models ought to change as the complexity and volume of data grow and as there are changes in different scientific disciplines. Query optimization is one of the issues that should be overcome. The existing models do not appear to achieve speed and relevance to a reasonable degree, particularly when searching through an extensive collection of repositories that introduce varying latencies. An improved outcome in both metrics can be obtained by more performant methods, like dynamic source selection based on recorded metadata and response history. The other area is promising, and this would involve improving real-time indexing mechanisms. Fixed metadata mappings are used in many federated systems; however, dynamic, ad-hoc metadata improvements can help better match disparate sources and reduce system dependency on manual maintenance with changing repositories. Better systems of metadata interoperability are also significant. Federated search systems can be made more effective on newer repositories by implementing a type of universal data standard, or by devising some sort of adaptive mapping layers to reduce configuration costs. Custom repository and result view preferences allow users to tailor the default search interface to their own workflows. It is possible to improve these user-centric features to move beyond basic usability and create a comprehensive customization of the user as a researcher.

6.2 Federated Search with the addition of AI and Machine Learning Algorithms

The future of automation of federated search systems is promising due to the use of AI and ML technologies. AI would also help process natural language queries, converting them to the syntactic format of the corresponding repository rather than the framework based on keywords. This helps the layman who is not bound by formal search order structures. The user behavior analysis incorporated in the relevance ranking of ML models optimizes the ranking by predicting the clicks, dwell time, and document downloads. AI systems can glean the meaning of various words. Therefore, federated search systems can find a document not labeled with the same name or classification that may seem logically possible. Besides, AI-based Duplicate detection Functions are unquestionably superior to the old rule-based approaches. A deep learning algorithm that evaluates the structure, language, and metadata of a document can be very effective at identifying subtle duplicates that may be missed by other systems. Predictive search is at the forefront. Federated systems might provide recommendations on queries or repositories based on the past utilization or pattern of the user with reference to a specific body of research. Such systems would enable user research to be directed towards highly relevant information that is not otherwise readily available.

6.3 Cooperation among Different Scientific Communities in order to enhance Federated Search Systems

The success of federated searching systems is closely connected with the degree of cooperation between different scientific fields and data owners. Data silos can persist when repositories operate independently, which complicates integration. Interdisciplinary work enables institutions to establish a degree of interaction that allows the sharing of metadata frameworks, API interfaces, and other institutional best practices related to information accessibility. The development, further refinement, and enhancement of typical schematics, protocols, and access policies across participating repositories will establish community-based forms of governance that will be leveraged by future federated systems. With these models, the evolution of the system can be used to support a wide variety of disciplines as opposed to one domain. Interdisciplinary work can create cross-domain ontologies that can help search systems overcome field-specific conceptual gaps. An excellent example is to investigate the connection between climate science data and human health to understand how environmental changes impact individual health. Enhanced collaboration enables federated search systems across institutions and countries to be developed and maintained in the long term, and, consequently, enables infrastructure and funding support.

VII. CONCLUSION

In conclusion, federated search models offer a centralized way of querying diverse databases that improves efficient access to disseminated scientific archives. The paper has addressed various architectures, such as centralized, peer-to-peer, and hybrid systems, as it analyzes performance, scalability, and usability. Information about query or response time, metadata fusion, duplicate suppression, and other user-inclusive measurements, related to interface design, was compared and contrasted with performance measures and case studies. The analysis demonstrates the dependency on hybrid models, where efficiency, fault tolerance, and schema interdependence play a crucial role, as well as the intelligent ranking of the results. Going even deeper, it is possible to add artificial intelligence and machine learning that can significantly enhance the capability to comprehend the semantics, foresee relevance, and tune dynamic queries.

Furthermore, coordination of researchers may offer additional up-to-date practices and ontologies, making it

more effective to optimize data findability and reuse. Such federated search models not only streamline the process of scientific research but also facilitate free access to scientific knowledge regarding any field or organization. The primary focus of future research should be adaptive and self-learning methods of federated systems and how these systems can be applied to novel data-intensive fields, such as genomics, climate science, and even artificial intelligence.

REFERENCES

[1] Aadiwal, V., Meher, K., Kalidhas, A. M., & Mishra, A. K. (2025). Spatial Modeling of Soil Salinity and Its Impact on Nutrient Availability and Agricultural Productivity. *Natural and Engineering Sciences*, *10*(1), 312-324. https://doi.org/10.28978/nesciences.1648720

[2] Ahani, M. (2019). The comparative study of male and female boarding school students' love and marriage, interpersonal, moral and sexual, job, and dormitory problems in Mahneshan (based on a content analysis of letters received by advisor). *International Academic Journal of Humanities*, *6*(1), 60-64. https://doi.org/10.9756/IAJH/V6I1/1910008

[3] Aravindh, G., & Sridhar, K. P. (2024). Resilient and Adaptive Secure Routing Protocol for Wireless Sensor Networks Using a Grey Wolf Optimizer and Lightning Search Algorithm. *Journal of Internet Services and Information Security*, *14*(4), 331-346. https://doi.org/10.58346/JISIS.2024.I4.020

[4] Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal*, *15*, 6, 1–24. https://doi.org/10.5334/dsj-2016-006

[5] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463, No. 1999). New York: ACM press.

[6] Bryant, R., Clements, A., De Castro, P., Cantrell, J., Dortmund, A., Fransen, J., ... & Mennielli, M. (2018). *Practices and Patterns in Research Information Management: Findings from a Global Survey*. *OCLC Research Report*. OCLC Online Computer Library Center, Inc. 6565 Kilgour Place, Dublin, OH 43017.

[7] Candela, L., Castelli, D., & Pagano, P. (2009). On-demand virtual research environments and the changing roles of librarians. *Library Hi Tech*, *27*(2), 239-251. https://doi.org/10.1108/07378830910968191

[8] Candela, L., Castelli, D., & Pagano, P. (2013). Virtual research environments: an overview and a research agenda. *Data science journal*, *12*, GRDI75-GRDI81. https://doi.org/10.2481/dsj.GRDI-013

[9] Carter, N., & Zhang, M. L. (2025). An Investigation of the Relationship Between Population Change and Electoral Outcomes. *Progression Journal of Human Demography and Anthropology*, 15-20.

[10] Chatterjee, A., & Sanyal, S. (2024). From production to market: Uncovering the complexities of COVID-19's impact on fisheries and aquaculture. *International Journal of Aquatic Research and Environmental Studies*, *4*(2), 37–52. http://doi.org/10.70102/IJARES/V4I2/3

[11] Choudhary, S., & Reddy, P. (2025). Improving the Storage Duration and Improving the Characteristics of Tender Coconut Water using Non-thermal Two-phase Microfiltration. *Engineering Perspectives in Filtration and Separation*, 7-12.

[12] Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., ... & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific data*, *5*(1), 1-11.

[13] Enguix, F., Carrascosa, C., & Rincon, J. (2024). Exploring Federated Learning Tendencies Using a Semantic Keyword Clustering Approach. *Information*, *15*(7), 379. https://doi.org/10.3390/info15070379

[14] Garba, A., Wu, S., & Khalid, S. (2023). Federated search techniques: an overview of the trends and state of the art. *Knowledge and Information Systems*, *65*(12), 5065-5095.

[15] Ġbrahimoğlu, D. (2018). Psychological Mobbing at Workplaces. *International Academic Journal of Organizational Behavior and Human Resource Management*, *5*(1), 95-103. https://doi.org/10.9756/IAJOBHRM/V5I1/1810006

[16] Geetha, K. (2024). Advanced fault tolerance mechanisms in embedded systems for automotive safety. *Journal of Integrated VLSI, Embedded and Computing Technologies*, *1*(1), 6-10.

[17] Gravano, L., Chang, C. C. K., Garcia-Molina, H., & Paepcke, A. (1997, June). STARTS: Stanford proposal for Internet meta-searching. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (pp. 207-218). https://doi.org/10.1145/253260.253299

[18] Gravano, L., Ipeirotis, P. G., Koudas, N., & Srivastava, D. (2003, May). Text joins in an RDBMS for web data integration. In *Proceedings of the 12th international conference on World Wide Web* (pp. 90-101). https://doi.org/10.1145/775152.775166

[19] Gupta, N., Verma, P., Gulhane, M., Rakesh, N., & Elngar, A. A. (2024). Federated query processing for data integration using semantic web technologies: A review. *Artificial Intelligence Using Federated Learning*, 276-291.

[20] Johnston, W. E. (2002). The computing and data grid approach: Infrastructure for distributed science applications. *Computing and Informatics*, *21*(4), 293-319.

[21] Kadhim, E. H., & Shani, M. M. (2023). Duration-Based Costing: New Cost Accounting Methodology. *International Academic Journal of Accounting and Financial Management*, *10*(1), 89–94.

[22] Lagoze, C. (2004). The open archives initiative protocol for metadata harvesting.

[23] Lu, J., & Callan, J. (2005, March). Federated search of text-based digital libraries in hierarchical peer-to-peer networks. In *European Conference on Information Retrieval* (pp. 52-66). Berlin, Heidelberg: Springer Berlin Heidelberg.

[24] Lu, J., & Callan, J. (2006). Full-text federated search of text-based digital libraries in peer-to-peer networks. *Information Retrieval*, *9*(4), 477-498.

[25] Mahmoudi, S., & Lailypour, C. (2015). A discrete binary version of the Forest Optimization Algorithm. *International Academic Journal of Innovative Research*, *2*(2), 60–73.

[26] Manghi, P., Manola, N., Horstmann, W., & Peters, D. (2010). An Infrastructure for Managing EC Funded Research Output: The OpenAIRE Project. *Grey Journal (TGJ)*, *6*(1).

[27] Marzotti, P. A. (2021). Metadata profiles for interoperability: the E-ARK specifications for e-archiving. *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 12, 3, 2021*, 105-118.

[28] Papadopoulos, G., & Christodoulou, M. (2024). Design and Development of Data Driven Intelligent Predictive Maintenance for Predictive Maintenance. *Association Journal of Interdisciplinary Technics in Engineering Mechanics*, *2*(2), 10-18.

[29] Park, J. R., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries*, *29*(3), 104-116. https://doi.org/10.6017/ital.v29i3.3136

[30] Sharma, A., & Nair, V. (2025). Developing a Medical Coding Curriculum for Surgery Students by Resolving Inconsistencies among Physician and Student Records. *Global Journal of Medical Terminology Research and Informatics*, *3*(1), 30-36.

[31] Tran, H., & Ngoc, D. (2024). The Influence of Effective Management on Hybrid Work Styles and Employee Wellness in Healthcare Organizations. *Global Perspectives in Management*, *2*(4), 8-14.

[32] Uvarajan, K. P. (2024). Advances in quantum computing: Implications for engineering and science. *Innovative Reviews in Engineering and Science*, *1*(1), 21-24.

[33] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

[34] Zeng, M. L., & Qin, J. (2020). *Metadata*. American Library Association.