# The ESW of Wikidata: Exploratory Search Workflows on Knowledge Graphs

Matteo Lissandrini[a], Gianmarco Prando[b], Gianmaria Silvello[b]

[a]*Department of Foreign Languages and Literature, University of Verona, Italy*
[b]*Department of Information Engineering, University of Padua, Italy*

**Abstract**

Exploratory search on Knowledge Graphs (KGs) arises when a user needs to understand and extract insights from an unfamiliar KG. In these exploratory sessions, the users issue a series of queries to identify relevant portions of the KG that can answer their questions, with each query answer informing the formulation of the next query. Despite the widespread adoption of KGs, the needs of current KG exploration use cases are not well understood. This work presents the "Exploratory Search Workflows" (ESW) collection focusing on real-world exploration sessions of an open-domain KG, Wikidata, conducted by 57 MSc Computer Engineering students in two advanced Graph Database course editions. This resource includes 234 real exploratory workflows, each containing an average of 45 SPARQL queries and reference workflows that serve as gold-standard solutions to the proposed tasks. The ESW collection is also available as an RDF graph and accessible via a public SPARQL endpoint. It allows for analysis of real user sessions, understanding query evolution and complexity, and serves as the first query benchmark for KG management systems for exploratory search.

*Keywords:* Wikidata, Exploratory Search, SPARQL



## 1. Introduction

The adoption of Knowledge Graphs (KGs) has been widespread, both for representing and storing enterprise knowledge bases as well as open-domain encyclopedic knowledge and Linked Open Data (LOD) in various scientific domains [20, 23, 28, 7, 10]. However, the heterogeneity of KGs presents a challenge in their effective utilization [18]. Their contents have become less familiar even to domain experts and almost impenetrable to first-time users, leading to the need for exploratory methods for KGs [17, 11]. Thus, KG exploration [13] is the machine-assisted process of progressive analysis of a KG aiming at understanding the graph's structure and nature to identify portions satisfying an information need, and thus extract insights to aid in formulating new questions and hypotheses.

Exploring a KG often involves extended, interactive sessions where the user issues a series of queries to meet their information needs [17, 11, 13, 4, 30]. Each query result typically informs the user's subsequent question, and the combination of all queries used to satisfy the information need forms an *exploratory search workflow*. Understanding these workflows as a whole rather than individual queries in isolation is essential for creating effective search tools and assistants for KGs. In their search and exploration workflows, KG users interact iteratively with the database by repeatedly issuing and modifying SPARQL queries in a short time-span [30]. Previous work [4] working on query logs of Wikidata, used a heuristic definition called "streak analysis" to understand which queries were part of the same session, capturing a sequence of similar queries within close distance of each other. This approach currently represents the best effort to understand users' search behavior over a large KG; however, it is limited by the uncertainty of the information needs that triggered a given search process, making it only marginally useful for understanding exploratory search processes.

Our goal in this work is to advance the understanding of exploratory workflows over large KGs and to create a new shared resource, the Exploratory Search Workflows (ESW) collection, to aid researchers in understanding the key aspects of exploratory search. Therefore, given the prominence of SPARQL as the de-facto standard query language for KGs, and given the large open-domain nature of WikiData, along its accessibility, which has attracted much research in the past years, we focus on these technologies. To our knowledge, this is the first extensive field study collecting a set of exploratory workflows conducted by users interacting directly via SPARQL with a real-world, large-scale, open-domain KG in a controlled environment.

Specifically, we designed 45 high-level exploratory workflow specifications, such as *"Explore the information regarding the movies directed by Woody Allen and Quentin Tarantino."* Each workflow is centered around a main search topic representing the user's information need and is then divided into more specific search tasks to help the user progressively discover more about the topic.

Each workflow specification further describes a set of subtasks. For example, each workflow begins with tasks pertaining the understanding of how the data regarding this topic is stored in the KG, such as *"which BGP can retrieve the movies directed by Woody Allen,"*, and then leads to more complex questions, such as *"who are the workers that participated in movies directed by both the directors."*

Multiple participants were assigned the same workflows, and their queries were collected via Jupyter notebooks, which

tracked all interactions with the SPARQL endpoint. The use of any other Web resources was prohibited. One or more SPARQL queries can solve each task, yet each participant was instructed to report all the syntactically correct queries they performed. Thus, we also collected queries returning the wrong set of results or no result at all. As a result, we can identify, for example, how many different queries a user issued before identifying the correct SPARQL query that can answer a given information need, e.g., all queries that the user issued before finally formulating the correct query to "retrieve all movies by a specific director."

Our resource was collected during two editions of the Graph Database course at the master's degree in Computer Engineering of the University of Padua (Italy). Participants were 57 MSc students in Computer Engineering who were trained in the course on Semantic Web technologies, including SPARQL, RDF, and ontology design. Each student completed one or more search topics, organized in seven different macro topics, each composed of at least four sub-tasks. This resulted in 234 distinct workflows containing a total of 10,645 SPARQL queries, each one targeting the same snapshot of Wikidata (truthy version, with only English labels) that was stored in a commonly controlled triplestore. Students were not querying the online live version of Wikidata.

Moreover, to ease the analysis of the ESW collection, we extracted the queries from the notebooks and represented them as an RDF graph that extends the LSQ schema [24]. This representation facilitates observational analysis of the characteristics of each task. Additionally, our resource includes a reference workflow for each search topic, serving as a gold standard for addressing each task. This provides a means to estimate the completeness and accuracy of the answers retrieved by participants. We also make available the full details and materials of our field study, enabling researchers to replicate our study with new users and deploy similar studies in different domains or using different workload specifications.

We anticipate several important applications for this resource, such as:

- enabling studies on how real users approach query reformulation and data exploration by representing the first real-world recording of end-to-end full exploratory workflows on KGs;
- serving as a real-world query benchmark, where researchers can study the impact of query optimization techniques on entire user workflows and exploratory sessions;
- offering a resource for teachers and instructors to help students learn to formulate complex SPARQL queries and interact with real-world KGs effectively.

Overall, ESW is designed as a comprehensive resource consisting of multiple components, each offering distinct value. It is the first resource to provide both a methodology for obtaining real workloads and query logs for exploratory workflows on KGs (KGs), as well as an actual representative dataset of such workflows performed by users, along with the ground truth. Hence, ESW is proposed as the first-of-its-kind complete benchmark dataset, featuring real user queries organized into sessions and accompanied by ground truth solutions.

The structured design of the ESW, derived from well-defined search tasks, makes it highly versatile for various research domains beyond its initial use case with MSc students. The dataset's availability in RDF format and accessibility via a SPARQL endpoint ensures seamless integration into diverse research workflows. This allows researchers to study user interactions with KGs in a structured and reproducible manner. For example, researchers can now analytically identify and explore patterns of usage of diverse operators and correlate them with the characteristics of the data or of the user's skills. Structured and programmable access to this resource makes it accessible to novel synehtic query generators tools as well. Additionally, in line with the research direction already highlighted in previous works [30], the dataset provides a valuable foundation for developing and evaluating query optimization techniques, query refinement tools, and other advanced methodologies for data management of Semantic Web data, significantly extending its relevance and utility beyond the scope of this study.

The remaining of this work is structured as follows. Section 2 defines the important concepts for modeling the exploratory workflows and highlights why existing resources do not provide insights into real-world exploration sessions. Section 3 discusses the related works and Section 4 presents two running examples to understand the collection's contents. Section 5 describes the field study and the main design choices. Section 6 reports the main figures of the ESW collection we release and the schema of the dataset collected. Section 7 provides some analyses of the query log collected. Then, Section 8 presents the reference workflows and a first evaluation to measure the precision and recall of the real workflows. Finally, Section 9 draws some conclusions, and outlines future work.

## 2. Understanding Exploratory Search Workflows

Given a user information need, data exploration involves understanding the dataset's structure and nature, identifying and characterizing relevant data and insights, and formulating new research questions and hypotheses [9]. To categorize data exploration tasks, we generally consider three macro-categories: (i) data summarization and profiling, (ii) exploratory analytics, and (iii) exploratory search [13]. Additionally, analyzing data exploration involves looking at sequences of interactions between the data analyst/user and the dataset instead of a single interaction [9]. Thus, an exploratory session, or workflow, typically consists of a sequence of queries where the results of each query inform the formulation of the next one.

In this work, *we focus on exploratory search use cases* and aim to analyze how real users interact with large open-domain KGs in the context of an entire data exploration workflow. We begin by defining the important concepts for modeling these workflows and then highlight how existing resources, although they provide insights into real interactions with KG endpoints, cannot effectively support the understanding of real-world ex-

ploration sessions. Finally, we provide an example of a search topic with two search tasks within the "movies" macro-topic.

***Definition of terms***. When modeling an exploratory workflow, we distinguish between two distinct stages. The first stage involves defining the user's *information need*, which begins with a general *search topic* expressed in natural language and is then refined into specific *search tasks*. Each search task represents a more focused information need, articulated as a natural language question, but closely aligned with a structured query. The second stage focuses on identifying the operations required to address these information needs. This entails translating the search topic and tasks into the queries necessary to accomplish each task, referred to as *search jobs*. Together, these jobs form the entirety of the *exploratory workflow*.

In our context, we define a **macro topic** as a general domain of an information need (e.g., *movies*). A **search topic** is a specific information need within a given macro topic. For instance, a search topic under the "movies" macro topic could be to compare and contrast Woody Allen and Quentin Tarantino as movie directors, analyzing factors such as number of Academy Awards won, budget differences, and shared crew members. A search topic comprises multiple **search tasks**, i.e., generic or specific questions allowing the user to extract the required information gradually. The complexity of search tasks increases as the user better understands the available data while proceeding with the exploration.

A **search job** is a collection of SPARQL queries that is supposed to respond to a given search task. For example, when collecting the movies directed by Tarantino, a search job could involve exploring the predicates around the entity for Tarantino, filtering on labels containing the keyword "director", and then using the identified predicate to return the full list of movies. Finally, an **exploratory workflow** is a collection of jobs representing an exploratory session over a search topic.

## 3. Related Work

To date, multiple online resources contain large-scale SPARQL query logs. There exist many synthetic benchmarks, e.g., WatDIV and its variations [6, 2], as well as a collection of query logs recording real users (and bots) interacting with public endpoints, e.g., LSQ [24, 29], and generators of synthetic queries based on real logs [25]. These resources are effective benchmarks to test the performance of different triplestores or different query processing and execution systems. The ESW collection distinguishes itself by addressing the structured, task-oriented workflows necessary to capture exploratory search behavior, which is absent in other resources. For instance, while LSQ provides general query logs, it lacks the progressive, task-driven structure that is central to understanding exploratory behavior. Similarly, WatDiv, though valuable as a synthetic benchmark for evaluating query engines, does not capture the real-world, iterative nature of exploratory search tasks that the ESW dataset is designed to represent. By focusing on these aspects, ESW fills a critical gap in supporting the study and evaluation of exploratory searches on KGs.

Overall, there has always been great interest in understanding how real users interact with KGs and how systems can better cater to their needs [16, 24, 4, 3, 30]. Consequently, having a well-balanced set of queries to study that is also representative of real user needs is fundamental in identifying the advantages and limitations of specific systems to cater to typical real-world workloads [2, 26, 22]. The need to investigate how user queries evolved during their information-seeking workflows has already been highlighted in past works [30, 4]. Analyzing how users modify their SPARQL queries within a session and examining their structural query patterns, reveals important insights into user behavior, and this can be used, among others, to design new tools to support users as well as to design algorithms to improve SPARQL query performance [30]. One of the most recent and extensive works on query log analysis [4] analyzes a large corpus of queries derived from public SPARQL endpoints. The availability of such query logs allows us to study the structural characteristics of the queries. This type of field study can further inform the design of new query language features or bring attention to optimizing specific query execution processes. A novel type of analysis has been proposed in such analysis, i.e., the *streak* defined as a sequence of queries that appear as subsequent modifications of a seed query. Because of a lack of information regarding the authors of the queries and their intent, the streaks have been derived from a heuristic analysis combining the edit distance of the queries with their temporal proximity, which leads to possible inaccuracies and ambiguities. For example, when a streak is terminated, it is unclear whether that is because (a) the user had satisfied their information need, or (b) they realized that a completely different query was required, or even (c) they just abandoned the task. Such information would instead be fundamental in supporting, for example, the study of query suggestion systems [12].

Nonetheless, given the need to understand how queries are correlated with those preceding and following them within an exploratory session, we highlight the need for a novel resource where much more information is stored than the information that can be found in existing query logs. We see that in existing SPARQL query logs, two fundamental pieces of information are missing: (1) which queries are part of a given exploratory session, and (2) the intent (or information need) subsumed by the entire workflow. Hence, to the best of our knowledge, we are the first to design and conduct a field study where we collect in a controlled environment both a large-scale query log (corresponding to more than 10K queries) as well as the information on the grouping of each query within a specific exploratory workflow as interpreted and executed by a single user.

In this study, we present a novel resource that captures the exploratory workflows of real users with moderate to high proficiency in SPARQL. These workflows represent a collection of search tasks and associated queries on a given topic, hence search jobs. We provide a detailed understanding of users' strategies to achieve their goals during a search task. Our resource is enriched with reference workflows that we designed for each task. By comparing these reference workflows to the workflows generated by users, we can evaluate the quality of the user-generated workflows against a ground truth answer set.

While these analyses are not possible with existing query logs and benchmark generators.

It is worth noting that these exploratory workflows can be used also to evaluate existing methods that help users in writing SPARQL queries. One of the most known is Sparklis [8], a Semantic Web tool designed to assist users in exploring and querying SPARQL endpoints by interactively guiding them through the process of constructing questions and answers, ranging from simple to complex ones. Sparklis supports various SPARQL features, and the queries are verbalized in English or French, ensuring that users are not required to master the SPARQL syntax. In addition to Sparklis, other methods have been proposed to simplify SPARQL query construction, such as visual aids, natural language to SPARQL conversion tools, faceted search interfaces, and conversational systems [1, 19, 21]. These approaches aim to abstract the complexities of SPARQL syntax and enhance the understanding of the underlying KG schema. They are valuable tools to support the users in their exploration. The ESW collection proposed in this work thus offers a possible source of data to design semi-automatic tests for these tools. The ESW can thus be employed to evaluate whether these methods improve exploratory search tasks, if their use brings users to issue different queries to the database, and enables the assessment of the search performance offered by these tools.

## 4. Exploratory Search Use Cases

In this section, we present two sets of examples of common exploratory tasks. The first set pertains to the "Movie" macro topic, while the second set focuses on the "Sport" macro topic. The following description is based on the contents of the search logs and their associated notebook used by the students.

***Movies***. We select the "Movie" macro topic and focus the search topic on movie directors as the domain of interest. The high-level search topic given to the students conveyed the following general information need:

> *Investigate the results concerning the common aspects between movies directed by Woody Allen and Quentin Tarantino. We are interested in the people who worked for both directors, what are the differences in terms of their movies' budget, and who won more Academy Awards.*

Given this initial information need, we provided a series of more specific search tasks to allow the students to explore and learn about the two directors. It is assumed that the students may have limited prior knowledge on the topic, as the exploratory search aims to acquire new knowledge about a particular subject [15]. The students were restricted from using any web resource other than the Wikidata SPARQL endpoint and were provided with a small set of IRIs to initiate the exploratory process. The provided IRIs included basic properties such as *instance of*, *subclass*, *nationality*, as well as specific IRIs related to the directors, including the IRI of the Wikidata entities representing "Woody Allen" and "Quentin Tarantino". We now see two of the exploratory tasks we assigned within this topic.

*Movie Task 1: Identify the BGP for films.* The task requires identifying how directors and movies are described and connected. Initially, the search process usually involves multiple queries as the student seeks to understand how the main entities are described, e.g., how WikiData represents the fact that a person is a director. In terms of relationships, the student may want to retrieve which predicates describe relationships connecting directors and movies. This information would be needed later to formulate more complex queries.

Typically, the student begins by exploring the given IRI for the topic, in this case, say, "Woody Allen" as entity with IRI `wd:Q25089`. To gather more information about `wd:Q25089`, the student queries the KG to discover the associated properties and entities, e.g., all triples with that entity as the subject. Such first query may reveal his occupation as a "film director", and by examining the object of the *occupation* property, the student obtains the Wikidata IRI for this occupation (`wd:Q2526255`). The next step may be to uncover how Allen is connected to entities representing his movies. Querying the KG for triples where `wd:Q25089` is the object reveals then other properties, such as *director* (`wdt:P57`), which suggests the entities appearing as subjects in those triples are possibly movies. One of the subjects retrieved for the *director* property is, for example, "Midnight in Paris" (`wd:Q206124`). Subsequently, the students can delve into this specific region of the graph and investigate `wd:Q206124`, realizing that it is an instance of "Film" (`wd:Q11424`). Since the goal specified for this task suggests finding the BGP able to retrieve entities of type films, the student has now formulated a query that returns the Wikidata IRIs of the instances of this class. Hence, it can satisfy the specific information need the search task requires.

*Movie Task 2: Compare the workers among the films directed by Woody Allen and Quentin Tarantino.* The task involves conducting a comparative analysis and identifying the workers involved in the two directors' respective films. One way to approach this is, similarly to the above, by querying Wikidata to reveal worker-related properties, such as *cast member* and *composer*. Like the above, the student can inspect properties where a movie appears as a subject or object. This can provide information about the properties connecting, for example, a movie to actors and music composers involved in the production. Once this information is identified, a variety of statistics can be generated. For example, it is possible to determine the people who have worked both on films directed by Woody Allen and Quentin Tarantino. Another approach could be to identify only the cast members who appeared in films by both directors. Alternatively, it is possible to determine which composer was most frequently used by the two directors. It is worth noting that there are multiple valid answers to this task, given the range of possible analyses that can be conducted, compared to the previous task, here students will probably employ queries with more complex structures or aggregation functions. Fur-

ther, despite the scenarios focus initially on a few specific entities, e.g., find movies of a given director, the sessions involve (sometimes implicitly) broader questions both a the beginning and towards the end. For example, other scenarios require to answer general questions like finding the top-5 production companies for number of crime films produced.

***Sports***. We present a workflow example from the "Sport Workflow Series (Olympic Games Explorative Search)" using the Olympic Games as the domain of interest. The proposed search tasks are designed to help students explore and learn about the Olympic Games. For clarity, we provide details for three specific search tasks and their execution. As above, it's important to note that students may not have prior knowledge about the topic, as one of the goals of exploratory search is to gain new insights about a particular subject [15]. To begin the exploration, also in this case, students receive a core set of IRIs, including standard properties such as "instance of", "subclass", and "nationality", as well as specific IRIs related to the Olympic Games, including the IRI for "Usain Bolt".

*Sport Task 1: Identify the BGP for Olympic Games.* As for the previous workflow, the student starts by exploring the given IRI for the entities relevant to the topic, such as *Usain Bolt*. If unfamiliar with Usain Bolt, the student queries the KG to discover his properties and associated entities. For example, one of the first queries reveals the *occupation* property, identifying Usain Bolt as a sprint runner.

Next, the student explores Usain Bolt's connection to the Olympic Games, wondering whether it is possible to find direct or indirect connections between starting with Usain Bolt and leading to the entities describing specific Olympic Games events. By querying the *participant in* property, the student retrieves *athletics at the 2012 Summer Olympics - men's 100 metres*, recognizing it as related to the Olympic Games.

The student then investigates the entity representing the *athletics at the 2012 Summer Olympics - men's 100 meters*, leading to the *2012 Summer Olympics*, which is part of the *Summer Olympic Games*, and ultimately linked to the *Olympic Games*. The task is complete when the student formulates a query that produces the IRI representing the Olympic Games, meeting the Search Task's requirements. When investigating properties about Usain Bolt, the student learns also that Usain Bolt is a Jamaican sprint runner who has participated in multiple Olympic Games. While there are other aspects of his life, including a brief football career, this information is not relevant to the task.

*Sport Task 2: Return all the editions of the Summer Olympic Games (do not consider future Olympic Games) with the country where they were played.* By querying information about the Olympic Games, the student understands that the Olympic Games are divided into winter and summer editions and thus can identify the properties to use to filter for the Summer Olympics. However, the locations of these games are still unknown to the students.

Reusing the BGPs that connect Usain Bolt to editions of the games, the student examines the properties and objects associated with the games themselves. By investigating a random edition of the Summer Olympic Games, the student discovers the *country* property, which indicates where the games were held.

To provide an accurate answer, the student needs to construct a query that retrieves pairs of elements comprising the edition of the Olympic Games and the corresponding country.

*Sport Task 3: Return statistics for the 2008 Summer Olympics Games.* This task requires an in-depth analysis of a Summer Olympic Games edition, aiming to identify not only the entities involved but also those for which is meaningful to compute any statistics. One approach is to explore the edition's *has part* property, which details the sports included. Each sport also has the *has part* property, indicating the disciplines within that edition. Investigating the disciplines may reveal the *victory* property, identifying gold medal winners.

This analysis allows the generation of various statistics, such as the number of disciplines per sport or the number of gold medals won by each country in that edition of the Games. It is clear that structurally the queries answering this last task use aggregations and different attributes that do not appear in the queries to address the previous tasks, yet they are clearly informed and enabled by those.

## 5. Study Design

We conducted a field study in two separate instances, one in 2021 and the other in 2022, as part of the advanced Graph Databases course in the MSc Computer Engineering program at the University of Padua. Each edition resulted in a set of exploratory workflows, which we distinguish by referring to the first as the "2021 track" and the second as the "2022 track." The students involved had a background in relational database systems and search engines and had completed 25 hours of frontal lectures on RDF and SPARQL, two seminar lectures on exploratory search, and three on KG exploration and creation. The study was structured as a 45-day-long individual course project.

We crafted a set of search topics for each macro topic and included a series of search tasks with varying levels of complexity or depth. Each participating student received a Jupyter notebook specific to their assigned search topic. The notebook contained a unique ID identifying the user, the specific search topic, and the Python code, allowing them to submit queries to a shared SPARQL endpoint. The endpoint was used to query a local version of Wikidata populated with truthy data and English labels. In the notebooks, students were instructed to add a new cell for each query they executed, a textual comment describing their search intent, and to report all the syntactically correct queries. Finally, students were tasked to provide a comment based on the output of the queries they judged more significant, e.g., *"this query shows no triples connecting Woody Allen to a Movie, where the movie appears as an object"*.

In our study, students were required to use only IRIs retrieved from previous queries within their workflow, ensuring that each notebook remained self-contained regarding the information consumed. We provided a small set of pre-approved Wikidata IRIs to initiate the process. Notably, external services

were prohibited to maintain complete control over the search process. This restriction focused on utilizing SPARQL queries and understanding the challenges in query formulation and execution without introducing additional complexity from external tools. Students were allowed to use text matching in their queries when necessary, which they did.

Students were allotted 45 days to complete their assigned workflows. In both cohorts, students could seek feedback from professors within a designated timeframe – the first three weeks of their work. After this period, they could no longer request feedback. Furthermore, students were allowed to submit one workflow for review approximately halfway through their working timespan to receive constructive feedback on their progress. Finally, the students from the 2022 cohort received additional training, during which we demonstrated how to perform a workflow using an example from the 2021 cohort.

All student queries were recorded in the notebooks and a query log on the server hosting the endpoint. Each search topic was addressed by a minimum of four and a maximum of six students, ensuring redundancy in the search workflows and comparing different search strategies for the same information need. After the project, the teacher individually re-executed and evaluated the notebooks.

The 2021 and 2022 tracks have been organized similarly but differ in the number of topics and tasks assigned to the students. Moreover, the topic and task specifications have been adjusted from 2021 to 2022. In 2021 we defined six macro topics with four search topics each. Each participating student was assigned six search topics, one for each macro topic. The search tasks in the 2021 track had a broader informational intent and their formulation was intentionally vague compared to the tasks in the 2022 track. An example is *"Investigate the movies by Quentin Tarantino"*. In this case, there is a range of possible plausible answers as the number of movies directed by Tarantino, the titles of the movies with Tarantino as an actor, or the awards won by Tarantino. Thus, we refer to these more vague tasks in the analysis as *informative exploratory search tasks*. We asked the students to investigate the tasks in depth and to provide as much relevant information as possible. Yet, we recognize that we cannot expect all students to look exactly for the same answer.

In 2022, we defined three macro topics with seven search topics each, and every student was assigned three search topics. The search tasks in this edition were more specialized since, compared to the 2021 track, they specified more precisely the information need in each search task, including the format of the answers (e.g., we specified when a list of IRIs was requited or when instead only an aggregate number was expected) in a way that was possible also to evaluate the correctness and completeness of the answer obtained. We note that we evaluate completeness only for the query that outputs the answer requested by the task, but in the process, the students were formulating multiple intermediate queries, which we track and analyze but for which we do not have any predefined answers. An example is *"How many films were directed by Quentin Tarantino in the first decade of the 2000s?"*, for which students where instructed to show all queries needed to obtain the information that allowed them to formulate the necessary BGPs as well as the final queries that compute the desired number. Then, we evaluate correctness and completeness only for the answer of that last query.

Table 1: Statistics of the ESW collection across the two tracks.

| | **2021** | **2022** |
|---|---|---|
| **Macro topics** | 6 | 3 |
| **Search topics** | 24 | 21 |
| **Students** | 21 | 36 |
| **Workflows** | 126 | 108 |
| **Total queries** | 4,861 | 5,784 |

## 6. The ESW collection

Table 1 presents the key statistics for the ESW collection across the two tracks. In 2021, there were six macro topics, each with four search topics. However, in 2022, the number of macro topics was reduced to three, with seven search topics each. As a result, the 2021 track covered a larger portion of Wikidata, while the 2022 track focused on a narrower region of the KG. In both cases, the students were always querying the same data snapshot. To ensure redundancy and enable comparison of different exploratory approaches for the same information need, each search topic was assigned to at least four and at most six students. In 2021, the student cohort consisted of 18 males (16 from Italy, one from Iran, and one from Spain) and three female students (two from Italy and one from Iran). On the other hand, the 2022 cohort was more diverse, comprising 28 male and eight female students, 13 from Italy and the remaining from Bangladesh, France, India, Iran, Pakistan, and Spain. The resulting ESW collection comprises 234 Workflows, with over $10,000$ queries performed by 57 students.

To make the released resource FAIR (Findable, Accessible, Interoperable, and Reusable), we created an RDF Graph to explore and query the metadata and data about the search topics and workflows. Figure 1 provides a graphical overview of the ESW ontology we developed;[1] it maximizes the reuse of existing ontologies such as LSQ [24] [2] and SD [3]. To model specific concepts not mapped by those vocabularies, we introduce new classes, namely: `Track`, `SearchTopic`, `SearchTask`, `GroundTruth`, `Worker`, `ExploratoryWorkflow`, and `SearchJob`.

We can see that the `ExploratoryWorkflow` is at the center of the ontology; it is connected to the `Worker` (in this case, the student) who performed (wrote) it and to the `SearchTopic` it implements. In turn, a `SearchTopic` is part of a search `Track` to distinguish the 2021 and 2022 editions. The `ExploratoryWorkflow` comprises several `SearchJobs`. Each `SearchJob` performs one `SearchTask`, and is composed of an ordered *list* of queries authored by the same user. For each

[1] `http://w3id.org/esw/ontology#`
[2] `http://lsq.aksw.org/vocab#`
[3] `http://www.w3.org/ns/sparql-service-description#`

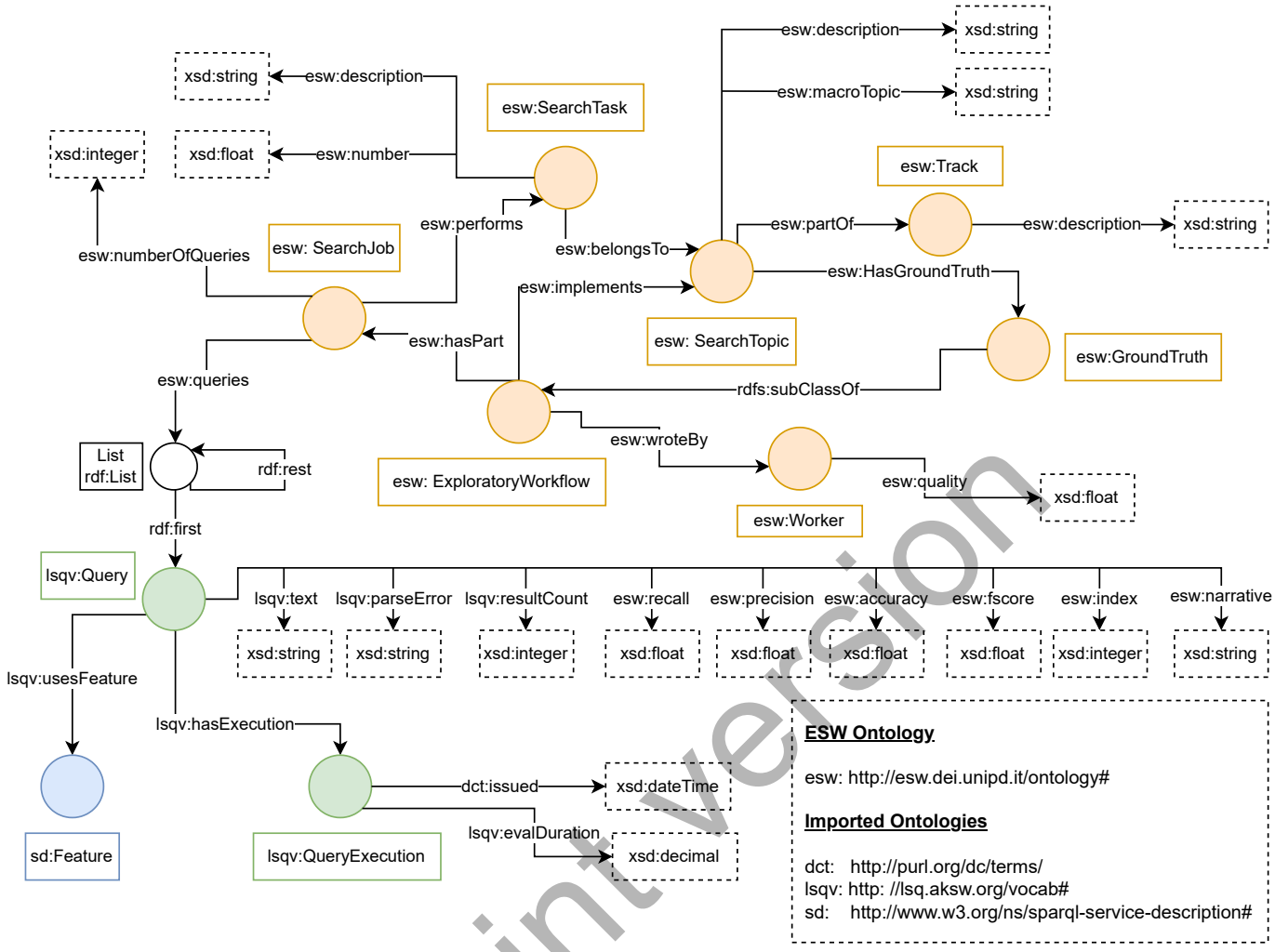


Figure 1: A graphical overview of the ESW Ontology.



`Query`, we model some core information as the text (i.e., the SPARQL query), the narrative (i.e., a textual comment explaining the goal of the query), if the query returns parsing errors, the size of the result list, and the index (i.e., the order of the query in the search workflow). Further, we also annotated queries with some effectiveness measures (e.g., precision, recall, and f-score) when possible.

The `Worker` class represents the user (in our case the student) who performed the `ExploratoryWorkflow`. We anonymized all information about students. Nevertheless, each `Worker` is annotated with a quality score based on the final mark obtained in the course. The student's quality measure is normalized in an [0,1] interval, where 0 means that the student did not pass the exam, and 1 means that the student passed the exam with the maximum score. This quality measure could be considered to exclude some students from downstream analyses.

## 7. Analysis and Statistics

In this section, we report the main statistics about the queries contained in the ESW collection we release. Table 2 displays the distribution of queries across tracks and macro topics. In the 2021 track, the six macro topics are "Movies", "Geography", "Politics", "Sports", "Companies" and "Books". As we can see, they cover diverse areas of Wikidata and there are 600 or more queries available for each macro topic. In the 2022 track, the macro topics are restricted to "Movies", "Sports" and "History" presenting more than 1600 queries per macro-topic. Two macro topics overlap between the two tracks, but the search topics within each are distinct.

As mentioned above, each of the 24 search topics from 2021 and the 21 from 2022 is executed independently by at least four students, for a total of 126 and 108 distinct search workflows, respectively.

Table 3 presents a breakdown of the number of queries available for each macrotopic and specific topic within the 2021 track of the Exploratory Search Workflows. The table highlights the aggregated statistics for each macrotopic in bold, with detailed data for individual topics listed beneath. Each macrotopic, such as Books, Companies, Geography, Movies, Politics, and Sport, represents the total number of workflows, the average number of queries per workflow, and the cumulative

Table 2: Query statics in the ESW collection divided by track and macro topic.

**2021 track**

| Macro Topic | #queries |
|---|---|
| Movies | 986 |
| Geography | 592 |
| Politics | 759 |
| Sports | 882 |
| Companies | 711 |
| Books | 931 |
| Total | 4,861 |

**2022 track**

| Macro Topic | #queries |
|---|---|
| Movies | 1,636 |
| Sports | 2,318 |
| History | 1,830 |
| Total | 5,784 |

number of queries. For example, the Books macrotopic encompasses 21 workflows with an average of 44 queries each, resulting in 931 queries overall. In contrast, the Geography macrotopic has the same number of workflows but a lower average of 28 queries per workflow, leading to 592 queries.

The table further details the distribution of queries among specific topics within each macrotopic. For instance, under the Movies macrotopic, the individual topics like TV series, Directors, and Horror Franchises exhibit a broad range of query counts and averages, with the total number of queries for Movies reaching 986. Similarly, within the Sport macrotopic, the individual topics such as F1 pilots and Olympic events show varying query statistics, culminating in 882 queries for the category. This structure not only emphasizes the aggregate statistics for each macrotopic but also provides insight into the diversity of query distributions among the various specific topics, illustrating the varying levels of focus and detail across different aspects of the 2021 exploratory search workflows.

Likewise, in Table 4, we provide a detailed overview of the statistics for the 2022 track of the ESW. The table breaks down the query statistics for specific topics within each macrotopic, providing insights into the focus and scope of the 2022 workflows. Within the Sport macrotopic, for instance, topics such as Association Football Players and Tennis show significant variation in query counts, with Tennis having the highest average of 104 queries per workflow and a total of 626 queries. Similarly, the Movies macrotopic reveals a diverse set of topics, from Tv series Without a Trace with 56 queries on average per workflow to Film Genre and directors with 50 queries on average, contributing to a total of 1,636 queries. This detailed breakdown not only highlights the overall query distribution across macro topics but also sheds light on the varying levels of detail and focus across different topics within each category.

### *7.1. Frequency of operators*

Table 5 shows the frequency of the various SPARQL keywords across queries. This analysis presents intriguing insights into user behavior. For example, we can observe that SELECT DISTINCT is present in most queries. While, given the exploratory nature of many queries, the presence of the LIMIT keyword signals the user's intent of just getting a sample of the output, e.g., to discard queries that would return an empty result set quickly. This strategy allows to get quick insights into the structure of the data, hence their fast execution is a crucial aspect of effective exploratory search. Nonetheless, a close inspection of the queries reveals that users often utilized the DISTINCT clause to obtain also unique properties or entities of interest and they use LIMIT to restrict the result to a manageable size and then verify whether the information they were looking for could possibly be retrieved by the query they just formulated, especially when querying large portions of the graph during the initial stages of the workflow.

Further, the query process's iterative and adaptive nature can explain the frequency distribution of keyword usage. Users often build on previous queries, making adjustments and refinements. At the same time, users may also retain `LIMIT` or `DISTINCT` keywords from previous queries even when not needed. For instance, users may keep the LIMIT keyword of an earlier query, even if the expected result set is much smaller, or they may retain a DISTINCT keyword, even if the result set is not likely to have duplicate elements. Furthermore, during the initial stages of exploration, the FILTER keyword is more commonly used. We have also noticed the usage of REGEX, to selectively match only the labels of properties and entities of interest when learning how to formulate the necessary BGPs.

An interesting aspect of the exploration process occurs at the culmination of the workflow, where all of the knowledge acquired on the graph's structure during the previous exploration is exploited to inspect complex structures and connections. Data aggregations often accompany this to obtain the final results. Queries that employ various aggregation keywords, such as GROUP BY and COUNT, are frequently used to group and summarize the data. We also report (not showing in the table) that nested queries are used 416 times (8.55% of the queries) in the 2021 track and 243 times (4.19% of the queries) in the 2022 track.

In conclusion, the querying behavior reflects the predominant need for exploring the structure and content of the KG in use cases where the user has to submit frequent queries that may quickly be revealed not to retrieve the exact answers needed. The widespread use of DISTINCT and LIMIT underscores their significance in facilitating effective exploratory search and query refinement in KGs. This is reflected in users adapting their queries based on the evolving requirements of their understanding.

### *7.2. Execution Time*

Table 5 also reports the median and the maximum execution time (in milliseconds) of the queries using a given keyword. To obtain uniform statistics, the exploratory workflows, with all their queries, have been re-executed using Virtuoso (version 07.20.3236) running on a 72 CPUs Intel(R) Xeon(R) Gold 6140M (2.30 GHz) with 1538GiB of RAM and 4.5 TB SSD.

Table 3: Detailed statistics of the 2021 track of the ESW. The macrotopic is marked in bold along the aggregated statistics. We report the number of workflows available for each topic, the average and variance of the number of queries per topic, and the total number of queries summing up all the workflows.

| *Search Topic* | *#Workflows* | *AVG(queries)* | *VAR(queries)* | *# queries* |
|---|---|---|---|---|
| **Books** | 21 | 44 | 33 | 931 |
| Political Magazines | 4 | 36 | 714 | 144 |
| Nobel laureates | 6 | 45 | 239 | 274 |
| Authors comparison | 6 | 42 | 194 | 253 |
| Author comparison | 5 | 52 | 1410 | 260 |
| **Companies** | 21 | 33 | 21 | 711 |
| IT Companies | 6 | 30 | 139 | 185 |
| Economy of EU States | 6 | 40 | 512 | 242 |
| Trademarks across the world | 5 | 34 | 267 | 171 |
| Business People in Germany | 4 | 28 | 490 | 113 |
| **Geography** | 21 | 27 | 46 | 592 |
| American Architects | 5 | 33 | 212 | 166 |
| European Cathedrals | 4 | 16 | 36 | 65 |
| Archaeological sites | 6 | 27 | 172 | 164 |
| Place of Birth, Death, and Burial | 6 | 32 | 869 | 197 |
| **Movies** | 21 | 47 | 55 | 986 |
| TV series | 5 | 50 | 574 | 251 |
| Directors | 6 | 55 | 322 | 331 |
| The Batman movies | 4 | 48 | 1395 | 193 |
| Horror Franchises | 6 | 35 | 223 | 211 |
| **Politics** | 21 | 35 | 138 | 759 |
| International Treaties | 4 | 21 | 50 | 87 |
| Monarchies | 5 | 51 | 1009 | 257 |
| Politicians in E.U. | 6 | 27 | 126 | 164 |
| Presidents of countries | 6 | 41 | 131 | 251 |
| **Sport** | 21 | 42 | 26 | 882 |
| F1 pilots | 6 | 46 | 552 | 276 |
| Olympic | 6 | 35 | 357 | 213 |
| World Records | 4 | 48 | 2858 | 193 |
| FIFA World Cup events | 5 | 40 | 359 | 200 |



Overall, the total execution time of the 2021 Track's 4781 queries is 8976 seconds (2 hours 29 minutes), with an average of 1877 milliseconds per query. The total running time for the 2022 Track's 5733 queries is 4542 seconds (1 hour and 15 minutes), averaging 792 milliseconds per query. Yet, not every query has been successfully executed due to the execution time limit of 300 seconds that we imposed in the SPARQL endpoint. 103 queries (63 in 2021, 40 in 2022) did not complete their execution as they exceeded the time limit.

In Table 6, we report the distribution of queries runtime across workflows. Here we see that most queries have very short response times. Specifically, for the 2021 and 2022 Track, 4088 and 5457 queries, respectively, accounting for 90% of the total queries, were executed in less than 300 milliseconds. The abundant use of LIMIT keyword and the fact that most queries inspect only the neighborhood of some specific entity are the two main factors keeping the computational complexity of these queries to a minimum.

In 2021, the search tasks were less focused and more open to interpretation, resulting in broader explorations and more complex queries. Conversely, in 2022, the tasks were more specific, resulting in more focused queries that were slightly more efficient. As a result, as shown in Table 5, in 2021, 8.7% of the queries, and in 2022, 3% of the queries took more than one second to execute. Often, the maximum running time almost reaches the 300-second limit as well. This shows running time can quickly escalate, leading to potential bottlenecks in the exploratory process.

In Table 7, we analyze the evolution of the complexity of the queries by looking at the change in running times. We divided (quantized) the exploratory workflows into ten phases (i.e., bins); that is, in the first bin we put the first 10% of queries of every workflow; in the second bin, the next 10% until the 10th bin contains the last 10% of queries that chronologically appear within a workflow. The table shows that the median execution time moderately increases as the workflows proceed. This trend can be attributed to the later phases of the search, which contain more complex queries (e.g., aggregations, and complex BGPs). In contrast, the first bins contain more straightforward queries that significantly impact the KG but are less demanding in execution time. In this case, using the `Limit` keyword helps reduce the execution time.

Table 4: Detailed statistics of the 2022 track of the ESW. The macrotopic is marked in bold along the aggregated statistics. We report the number of workflows available for each topic, the average and variance of the number of queries per topic, and the total number of queries summing up all the workflows.

| *Search Topic* | *#Workflows* | *AVG(queries)* | *VAR(queries)* | *# queries* |
|---|---|---|---|---|
| **History** | 36 | 49 | 333 | 1830 |
| Literary Movements and Divine Comedy | 5 | 53 | 767 | 265 |
| Euro | 6 | 53 | 2214 | 323 |
| World Wide Web | 5 | 31 | 579 | 158 |
| Ancient Civilization | 6 | 89 | 3234 | 539 |
| Cultural Movements | 5 | 34 | 172 | 174 |
| Ancient Rome | 4 | 46 | 548 | 185 |
| Literary Movements and physicists | 5 | 37 | 48 | 186 |
| **Movies** | 36 | 44 | 105 | 1636 |
| Tv series Without a Trace | 5 | 56 | 1558 | 284 |
| Disney | 5 | 34 | 179 | 171 |
| Film Genre and directors | 5 | 50 | 709 | 253 |
| Production company | 5 | 36 | 56 | 184 |
| Tv series HIMYM | 6 | 61 | 1796 | 368 |
| Sherlock Holmes | 6 | 38 | 162 | 232 |
| Film Genre and composer | 4 | 36 | 203 | 144 |
| **Sport** | 36 | 62 | 440 | 2318 |
| Association Football Players | 5 | 39 | 149 | 195 |
| Olympic Games | 4 | 46 | 551 | 186 |
| Running | 5 | 44 | 1056 | 221 |
| Tennis | 6 | 104 | 11213 | 626 |
| Basketball and NBA seasons | 5 | 65 | 1753 | 326 |
| Association Football Club | 5 | 76 | 1028 | 382 |
| Basketball and NBA finals | 6 | 63 | 1798 | 382 |

Table 5: Distribution of the SPARQL keywords across the queries divided by track.

| | **2021 Track** | | | | **2022 Track** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Queries** | | **Exec. Time (ms)** | | **Queries** | | **Exec. Time (ms)** | |
| **Keyword** | **number** | **perc.** | **median** | **max** | **number** | **perc.** | **median** | **max** |
| ASK | 51 | 1.04% | 14 | 4 847 | - | - | - | - |
| AVG | 57 | 1.17% | 31 | 8 220 | 6 | 0.1% | 909 | 20 328 |
| COUNT | 1 242 | 25.55% | 34 | 190 666 | 881 | 15.23% | 17 | 232 515 |
| DISTINCT | 3 412 | 70.19% | 19 | 250 308 | 5 319 | 91.96% | 9 | 232 515 |
| EXISTS | 41 | 0.84% | 19 | 73 771 | 22 | 0.38% | 21 | 13 009 |
| FILTER | 1 482 | 30.48% | 41 | 292 234 | 1 557 | 26.91% | 14 | 170 077 |
| GROUP BY | 1 050 | 21.6% | 42 | 250 308 | 864 | 14.93% | 17 | 170 077 |
| GROUP_CONCAT | 201 | 4.13% | 46 | 250 308 | 71 | 1.22% | 12 | 16 056 |
| HAVING | 56 | 1.15% | 36 | 2 287 | 250 | 4.32% | 21 | 170 021 |
| LIMIT | 2 349 | 48.32% | 16 | 292 234 | 5 234 | 90.49% | 9 | 232 515 |
| MAX | 114 | 2.34% | 50 | 190 666 | 46 | 0.79% | 17 | 170 077 |
| MIN | 85 | 1.74% | 40 | 181 773 | 35 | 0.6% | 14 | 6 005 |
| MINUS | 8 | 0.16% | 67 | 151 062 | 42 | 0.72% | 19 | 110 405 |
| NOT EXISTS | 137 | 2.81% | 104 | 119 194 | 82 | 1.41% | 16 | 22 268 |
| OPTIONAL | 343 | 7.05% | 17 | 250 308 | 348 | 6.01% | 10 | 8 896 |
| ORDER BY | 1863 | 38.32% | 28 | 250 308 | 746 | 12.89% | 15 | 170 077 |
| REGEX | 485 | 9.97% | 42 | 282 559 | 671 | 11.6% | 14 | 170 077 |
| SELECT | 4 841 | 99.58% | 14 | 292 234 | 5783 | 99.98% | 9 | 232 515 |
| SUM | 28 | 0.57% | 17 | 47 918 | 34 | 0.58% | 76 | 75 939 |
| UNION | 259 | 5.32% | 56 | 90 124 | 199 | 3.44% | 18 | 92 937 |

Table 6: Distribution of queries runtime across workflows divided by track.

| | 2021 Track | 2022 Track |
|---|---|---|
| **Ranges (msec)** | **queries** | **queries** |
| t<300 | 4 088 | 5 457 |
| 300<=t<=1000 | 295 | 105 |
| t>1000 | 398 | 171 |

Furthermore, Table 7 also reports the mean execution time for each bin. We can see that the later phases of the exploration are generally more demanding than the earlier ones. This result can be explained by the fact that users often had to reformulate a query several times before getting the desired result. This process could be expensive and time-consuming, highlighting the potential utility of approximate query-answering methods to give users a fast answer before they finalize their query. Overall, the analysis of execution times divided into different phases of the exploration provides insights into the complexity of the exploratory search process. By understanding the bottlenecks and challenges users face when crafting exploratory queries, we can develop more effective and efficient methods to support their search for knowledge.

## 8. Analysis With The Reference Workflows

The ESW collection includes a manually created reference workflow (i.e., a form of ground truth) for each search task within a given topic, representing the set of queries computing the ideal answer to the information need expressed by each task. These reference workflows represent a *gold standard* set of queries for addressing the topic. Note that the 2021 track contains more general tasks for which is hard to establish clear desired answer. Thus, the recall for those tasks is often unknown. Instead, the 2022 track was designed explicitly to comprise more focused tasks where the correct answer can be clearly identified and this the completeness of the answer can be evaluated. Hence, we created the gold standards for all the search topics of the 2022 track, while for the 2021 track, we created them only for the search topic within the "Movies" macro topics since they were the only ones specific enough to measure recall; in this case, we could provide the gold standard for 23 out of 28 tasks.

Although the "Movie" macro topics are the most specific for the 2021 track, we decided not to include some tasks where the request was vague (e.g., *"Compare the workers between Allen and Tarantino"*) in the ground truth, as there are multiple possible solutions. In particular, in the 4 "Movies" topics, the number of tasks we did not consider is 5 out of 28, 3 out of 11 for *Directors*, 1 out of 5 for *The Batman*, 1 out of 7 for *Horror Franchises*, and none out of 5 for *Tv series*.

### *8.1. Informative oriented Exploratory Tasks.*

As mentioned earlier, in the 2021 track, some search tasks are described by vague wording and may require different queries to describe complementary questions of the information needed to be described. Thus, our analysis recognizes that Informative Search Tasks are inherently broad and often pose challenges in determining whether a Search Job has been completed correctly and comprehensively. This difficulty arises because such tasks do not clearly specify which entities are involved or how they should be analyzed.

For instance, consider *Movie Task 2* from our running example, which asks: "Compare the workers among the films directed by Woody Allen and Quentin Tarantino." When students engage with this task, they might retrieve various relevant statistics, such as the average size of the cast. While many valid statistics could be retrieved, the task's inherent ambiguity means the student cannot unambiguously determine which exact set of queries to produce.

Similarly, consider the task "Compare Cristiano Ronaldo with Lionel Messi." The student recognizes that both are footballers, but the term "compare" is vague and can encompass a wide range of valid approaches: comparing the clubs they have played for, the awards they have won, the trophies they have received, their international appearances, and so forth.

This type of task is akin to information retrieval because the student must identify and report as many relevant and accurate items as possible without a guarantee of completeness, as it is impossible to know when all relevant data have been captured. Due to these challenges, we did not establish a ground truth for such *informative oriented* exploratory search tasks.

### *8.2. Completeness Oriented Exploratory Tasks.*

Completeness-Oriented Exploratory Tasks are very well-defined and generally straightforward to evaluate for correctness. This is because the task requirements are providing clear instructions on which entities are involved and how to combine them to achieve the final answer.

For example, *Movie Task 1* from our running example exemplifies this type of task. This task requires: "Identify the BGP for films." In this case, the user understands the entities that should be included in the final solution and how to structure the answer. The task involves finding a path to identify what constitutes a "Film," so the correct solution must report the IRIs that represent Films.

We always establish a ground truth for such tasks because the expected outcome is clearly defined. The task requirements are clear, making it easy to determine whether a Search Job correctly addresses the task.

Consider the following example: "List the footballers who have won the FIFA Ballon d'Or. For each country, return the number of footballers of that nationality who have won the FIFA Ballon d'Or." In this task, the student knows from the outset that they need to find all footballers who have won the award, group them by nationality, and then return the count of players per country. The specific and unambiguous nature of the information need described by this task ensures that we can effectively verify the correctness of the Search Job.

### *8.3. Experimental setup*

As we have already said, the two tracks of Search Workflows have been kept separated due to nature of the Search Tasks. In

Table 7: Evolution of queries runtime through a 10-bin workflow quantization.

| | 2021 Track | | | | 2022 Track | | | |
|---|---|---|---|---|---|---|---|---|
| | Queries | Exec. Time (ms) | | | Queries | Exec. Time (ms) | | |
| Bin | number | mean | median | max | number | mean | median | max |
| 0 | 542 | 810 | 11 | 140 751 | 629 | 21 | 8 | 5 979 |
| 1 | 477 | 2 201 | 10 | 160 259 | 573 | 320 | 8 | 140 492 |
| 2 | 485 | 1 772 | 12 | 282 559 | 582 | 107 | 9 | 16 355 |
| 3 | 469 | 736 | 15 | 47 791 | 569 | 872 | 9 | 152 186 |
| 4 | 453 | 3 088 | 18 | 250 308 | 555 | 611 | 9 | 137 006 |
| 5 | 506 | 1 551 | 17 | 119 194 | 588 | 1 694 | 10 | 140 096 |
| 6 | 478 | 1 136 | 17 | 79 848 | 575 | 85 | 11 | 11 080 |
| 7 | 475 | 1 827 | 17 | 109 527 | 563 | 845 | 10 | 115 366 |
| 8 | 478 | 3 045 | 16 | 190 666 | 574 | 839 | 11 | 134 637 |
| 9 | 418 | 2 940 | 20 | 292 234 | 525 | 2 748 | 10 | 232 515 |

fact, the 2021 track is mostly Informative oriented, while the 2022 is more Completeness oriented. For the 2022 track we have built a ground truth for each Search Task of each Search Topic, while for the 2021 track we have built the Ground Truth only for the four Movie's Search Topics.

It is worth noting that, besides comparing the execution times of the queries involved, there is currently no established method to evaluate the effectiveness (i.e., the accuracy or the quality of the search process) of an exploratory search workflow over a KG. This is not surprising given the absence of a benchmark collection to assess and quantify the success of exploratory search across a KG. The ESW collection provides a promising opportunity to develop a practical approach to evaluating effectiveness in this domain. In the following, we provide an initial analysis while we acknowledge the need for a more appropriate and comprehensive evaluation framework for exploratory search and leave this topic for future research.

We assessed the efficacy of the workflows by comparing them to their corresponding gold standard. We compared each query's output in the workflow with the expected outcome from the gold standard for each search job. Whenever possible, we calculated the recall and precision for each query's result set by treating the outputs as a set of tuples or values, depending on the query. Perfect precision is attained when the search task's result set only contains results found in the gold standard result set. For example, for task 1, which required films directed by Woody Allen, the gold standard had 50 IRIs, each representing a directed film. A task that returns only a subset of this result set would achieve perfect precision but not perfect recall. Conversely, perfect recall is accomplished when a superset of the expected result set is returned.

For aggregation queries such as *"return the maximum budget of Woody Allen movies"*, the precision is perfect if the correct budget is returned, and the recall is perfect if the IRI of the film with the maximum budget is returned. We acknowledge that more nuanced evaluation measures can be designed, and this evaluation may excessively penalize some workflows. For instance, future work could consider incorporating similarity-based measures that account for cases where alternative but valid literals are retrieved from the graph, even if they differ from those in the ground truth. This would allow for more flexibility in assessing the correctness of queries, as semantically correct statements, but with slight variations, may still meet the user's information need. In addition, while our current evaluation relies on set-based measures, incorporating ranking-based metrics could further enhance the assessment process. This is important because some portions of the correct answers may be more relevant to the user than others. By using ranking measures, we could account for the relative importance of different answers rather than treating every statement in the result set as having an equal impact, providing a more user-centric evaluation of the workflow's effectiveness.

### *8.4. Experimental Evaluation*

Table 8 reports the precision and recall for all the exploratory workflows available for the "Movies" macro topic in 2021. Table 10 reports the overall average precision and recall values with the variance for the 2021 ESW cohort. We can see that the 2021 track presents four search topics: some are executed by four students (e.g., *the Batman movies*) and others by six.

The same goes for the 2022 track reported in Table 9, which presents seven search topics for the Movie macrotopic, with no overlap with those proposed in 2021. In Table 11, we report the average precision and recall values with the variance for the 2022 ESW cohort. We can see that the effectiveness (in terms of both precision and recall) of the 2022's workflows is generally higher than those of the 2021 track. Also, in this case, this is probably due to the nature of the proposed search tasks. In 2022, there are successful workflows (almost) matching the gold standard (e.g., W3 for the *Sherlock Holmes* topic), but also others where the exploratory search did not lead to satisfying the information need (e.g., W3 for *Tv series HIMYM*). For some search topics, the exploratory process consistently led to satisfactory results, for instance, *Sherlock Holmes* in 2022, whereas others are harder to satisfy, such as *Directors* in 2021.

Table 9 reports the evaluation measures for the workflows available for the macrotopics "History", "Movie", and "Sport" from the 2022 track. The "Movie" macrotopic has already been analyzed compared to the 2021 results. The results obtained for "History" and "Sport" are consistent with those described

Table 8: Precision and Recall for the exploratory workflow in the "Movies" macro topic in the 2021 track. "-" indicates a missing workflow. The best precision (red) and recall (blue) for each search topic are in bold. Note that every table's cell evaluates a different workflow for a specific search topic.

| | | W1 | | W2 | | W3 | | W4 | | W5 | | W6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Topic** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** |
| **Movie** | Tv series | **0.56** | **0.58** | 0.23 | 0.27 | 0.28 | 0.47 | 0.21 | 0.50 | 0.08 | 0.04 | - | - |
| | Directors | 0.14 | 0.03 | 0.28 | 0.24 | 0.20 | 0.32 | **0.39** | **0.50** | 0.23 | 0.47 | 0.10 | 0.09 |
| | The Batman movies | 0.27 | 0.17 | 0.33 | **0.37** | **0.45** | 0.22 | 0.35 | 0.35 | - | - | - | - |
| | Horror Franchises | 0.07 | 0.15 | 0.17 | 0.17 | **0.65** | **0.42** | 0.18 | 0.07 | 0.09 | 0.24 | 0.32 | 0.17 |

Table 9: Precision and Recall for the exploratory workflow in the "History", "Movies", "Sport" macro topic in the 2022 track. "-" indicates a missing workflow. The best precision (red) and recall (blue) for each search topic are in bold. Note that every table's cell evaluates a different workflow for a specific search topic.

| | | W1 | | W2 | | W3 | | W4 | | W5 | | W6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Topic** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** | **prec** | **rec** |
| **History** | Literary Movements and Divine Comedy | 0.86 | 0.93 | 0.67 | 0.94 | **0.93** | 0.85 | 0.79 | **0.95** | 0.52 | 0.62 | - | - |
| | Euro | **0.94** | 0.87 | 0.86 | **0.99** | 0.77 | 0.81 | 0.57 | 0.54 | 0.54 | 0.52 | 0.62 | 0.63 |
| | World Wide Web | 0.66 | 0.75 | 0.63 | **0.84** | 0.71 | 0.63 | 0.53 | 0.71 | **0.86** | 0.72 | - | - |
| | Ancient Civilization | **0.77** | **0.78** | 0.67 | 0.72 | 0.64 | 0.72 | 0.51 | 0.62 | 0.69 | **0.78** | 0.53 | 0.75 |
| | Cultural Movements | 0.48 | 0.34 | 0.72 | 0.62 | 0.55 | **0.74** | 0.09 | 0.53 | **0.79** | 0.74 | - | - |
| | Ancient Rome | 0.63 | 0.74 | **0.91** | 0.99 | 0.83 | **1.00** | 0.70 | 0.78 | - | - | - | - |
| | Literary Movements and physicists | 0.55 | 0.63 | 0.57 | 0.67 | 0.65 | 0.71 | **0.70** | **0.74** | 0.65 | 0.67 | - | - |
| **Movie** | Tv series Without a Trace | 0.70 | 0.78 | 0.53 | 0.86 | **0.86** | 0.77 | 0.62 | **0.98** | 0.54 | 0.73 | - | - |
| | Disney | 0.59 | 0.76 | 0.77 | 0.90 | 0.68 | 0.82 | 0.37 | 0.65 | **0.83** | **0.94** | - | - |
| | Film Genre and directors | 0.61 | **0.88** | 0.58 | 0.51 | **0.75** | 0.74 | 0.62 | 0.62 | 0.69 | 0.64 | - | - |
| | Production company | 0.43 | 0.79 | 0.68 | 0.88 | 0.18 | 0.42 | **0.92** | **0.89** | 0.36 | 0.78 | - | - |
| | Tv series HIMYM | 0.40 | 0.53 | 0.38 | **0.57** | 0.19 | 0.30 | 0.34 | 0.34 | **0.41** | 0.56 | 0.34 | 0.45 |
| | Sherlock Holmes | 0.40 | 0.78 | 0.75 | 0.95 | 0.95 | 0.90 | 0.69 | **1.00** | **1.00** | 0.92 | 0.63 | 0.83 |
| | Film Genre and composer | 0.19 | 0.46 | 0.18 | 0.33 | 0.70 | 0.77 | **0.79** | **1.00** | - | - | - | - |
| **Sport** | Association Football Players | 0.11 | 0.62 | 0.51 | 0.81 | **1.00** | **1.00** | 0.69 | 0.93 | 0.53 | 0.88 | - | - |
| | Olympic Games | **0.88** | **0.86** | 0.76 | 0.73 | 0.54 | 0.42 | 0.35 | 0.57 | - | - | - | - |
| | Running | 0.34 | **0.55** | 0.03 | 0.14 | 0.38 | 0.33 | 0.37 | 0.51 | **0.42** | 0.55 | - | - |
| | Tennis | 0.63 | **0.72** | **0.73** | 0.64 | 0.64 | 0.65 | 0.50 | 0.54 | 0.48 | 0.58 | 0.64 | 0.56 |
| | Basketball and NBA seasons | 0.70 | 0.91 | 0.80 | 0.99 | 0.57 | 0.57 | **0.94** | **1.00** | 0.75 | **1.00** | - | - |
| | Association Football Club | 0.62 | **0.98** | 0.71 | 0.90 | **0.91** | 0.97 | 0.74 | 0.93 | 0.42 | 0.57 | - | - |
| | Basketball and NBA finals | 0.65 | 0.96 | 0.53 | 0.78 | **1.00** | 0.99 | 0.66 | 0.76 | 0.82 | **1.00** | 0.74 | 0.95 |



for the "Movie" macrotopic. Notably, for the "Sport" macrotopic, there is a wider variation in performance between the workflows compared to the other macrotopics. For instance, the "Association Football Players" topic (first row of the Sport section) shows W1 with very low precision, followed by W2 and W5 with better, yet still low precision. In contrast, W3 exhibits perfect precision and recall. These results highlight the diverse nature of the workflows and the different approaches to exploratory search. The ESW resource links each workflow to the grade obtained by the student performing it, which aids in identifying the workflows conducted by the most or least proficient students in SPARQL querying. This can also help identify common mistakes or misunderstandings, which can help develop educational materials.

In general, we can see that the workflows frequently fall short of achieving perfect precision and recall, showing that there is much room for developing new systems and techniques to aid users in the exploratory search process. Often, we see that the user is unaware that their query is retrieving a partial set of answers because the Wikidata may not store information uniformly. For instance, 48 films may be connected to Woody Allen by the `is directed by` property and two by the inverse (`directed`) property; hence, a successful query retrieving the films by Woody Allen can be incomplete for this reason. For instance, some awards may be connected to Woody Allen by the `award received` property and others by the inverse (`winner`) property; hence, a successful query retrieving the awards by Woody Allen needs to accommodate both cases.

## 9. Conclusions

We present the ESW collection, a resource providing real-world exploratory search workflows over Wikidata and quantitative and qualitative data about users' interactions with KGs beyond single isolated queries. The ESW collection comprises more than 10K SPARQL queries, 234 exploratory workflows performed by 57 trained master students in computer engineering working on 45 diverse search topics over Wikidata, the largest available open-domain KG. Each workflow focuses on a particular search topic, broken down into search tasks that aid the user in thoroughly exploring the KG. ESW fills an important existing gap in the study of exploratory search since it is the first resource providing insights into real long-winded interactions with KG endpoints also supporting the understanding of exploration sessions. We can study the search process within its incremental evolution, with each query building upon the previous one in a continuous refinement process. The need for quick interactive query refinement can motivate, for example, the study of new approximate query answering methods. We envision how a more in-depth study of these interactions could

Table 10: Average precision and recall measures (and variance) for the 2021 ESW cohort.

| | Topic | avgPrec | varPrec | avgRec | varRec |
|---|---|---|---|---|---|
| Movie | Tv series | 0.27 | 0.03 | 0.37 | 0.04 |
| | Directors | 0.22 | 0.01 | 0.28 | 0.03 |
| | The Batman movies | 0.35 | 0.00 | 0.28 | 0.01 |
| | Horror Franchises | 0.25 | 0.04 | 0.20 | 0.01 |

Table 11: Average precision and recall measures (and variance) for the 2022 ESW cohort.

| | Topic | avgPrec | varPrec | avgRec | varRec |
|---|---|---|---|---|---|
| History | Literary Movements and Divine Comedy | 0.75 | 0.02 | 0.86 | 0.02 |
| | Euro | 0.72 | 0.02 | 0.73 | 0.03 |
| | World Wide Web | 0.68 | 0.01 | 0.73 | 0.00 |
| | Ancient Civilization | 0.63 | 0.01 | 0.73 | 0.00 |
| | Cultural Movements | 0.53 | 0.06 | 0.59 | 0.02 |
| | Ancient Rome | 0.77 | 0.01 | 0.88 | 0.01 |
| | Literary Movements and physicists | 0.62 | 0.00 | 0.68 | 0.00 |
| Movie | Tv series Without a Trace | 0.65 | 0.01 | 0.82 | 0.01 |
| | Disney | 0.65 | 0.03 | 0.81 | 0.01 |
| | Film Genre and directors | 0.65 | 0.00 | 0.68 | 0.02 |
| | Production company | 0.51 | 0.07 | 0.75 | 0.03 |
| | Tv series HIMYM | 0.34 | 0.01 | 0.46 | 0.01 |
| | Sherlock Holmes | 0.74 | 0.04 | 0.9 | 0.01 |
| | Film Genre and composer | 0.46 | 0.08 | 0.64 | 0.07 |
| Sport | Association Football Players | 0.57 | 0.08 | 0.85 | 0.02 |
| | Olympic Games | 0.63 | 0.04 | 0.64 | 0.03 |
| | Running | 0.31 | 0.02 | 0.42 | 0.03 |
| | Tennis | 0.60 | 0.01 | 0.62 | 0.00 |
| | Basketball and NBA seasons | 0.75 | 0.01 | 0.89 | 0.03 |
| | Association Football Club | 0.68 | 0.03 | 0.87 | 0.02 |
| | Basketball and NBA finals | 0.73 | 0.02 | 0.91 | 0.00 |



be useful in identifying the primary bottlenecks of the exploration process and lead to the development of innovative new tools to assist users' explorations. Further, this resource can be used as a workload-aware benchmark to test the performance of KG management systems.

Thus, the ESW can support the following research tasks.

- **Better Understanding of the Exploratory Process:** By examining how various users approach the same information need, we gain insights into different ways people think about and explore KGs (KGs). This understanding aids in designing more effective tools and systems for exploratory search, and potentially informing the development of SPARQL to simplify complex query tasks.
- **Training (Semi-)Automatic Tools:** Creating a training set for tools that generate SPARQL queries from natural language descriptions, such as large language models, is valuable. This benefits users who are unfamiliar with SPARQL or lack the time or expertise to write their own queries.
- **Evaluating Exploratory Search Methods and Systems:** By comparing the solutions provided by exploratory search methods and systems with human-created workflows, we can better understand their strengths and weaknesses. These workflows can be benchmarks for studying query processing performance for exploratory workloads.
- **Providing a Common Baseline:** Reference workflows enable comparing existing solutions, such as query synthesis from natural language or next-query suggestion. They serve as a fundamental resource for researchers who need to validate new methods for exploratory search.

ESW meets the FAIR data requirements by being published as open data with a persistent URI (w3id), modeled via an open ontology, and searchable with a SPARQL endpoint. The exploratory workflows are released in human- (JSON, Jupyter notebooks) and machine-readable (RDF) formats. We provide the source code to process the raw data and analyze the workflows and the query log. We also release the empty Jupyter notebooks that can be reused in new research projects investigating exploratory search processes.

The ESW collection also comprises reference workflows, which allow for evaluating user workflows to determine their quality and degree of success. The availability of real and reference exploratory workflows opens new research directions in evaluating exploratory searches. Indeed, as future works, we aim to study how exploratory search tools [12, 5, 27] can be evaluated beyond the efficiency perspective (i.e., time and space), but also on the effectiveness side determining how much they can help a user to improve the quality of their search and the relevance of the obtained results (i.e., precision, recall, accuracy). This goal requires new evaluation methods to determine an ideal answer to an information need and an ideal search process over a KG. This resource further opens to investigations in the impact of user proficiency levels on exploratory search workflows. We envision the possibility to re-execute the ESW dataset workflows with participants exhibiting varying levels of

SPARQL expertise. This will allow to analyze how user proficiency affects workflow execution, particularly in query quantity and quality. These studies will enhance the generalizability and applicability of the ESW dataset, offering a deeper understanding of user behavior and tool-assisted search in KGs.

*Benchmark availability and useful URLs:.*

- The ESW Ontology, the ESW collection, and the query log are available in Zenodo [14].
- A SPARQL endpoint is available: `https://w3id.org/esw/sparql`. (Accessible with dataset name `http://w3id.org/esw/`).
- The code to process the raw workflows, produce the RDF graphs, calculate statistics, and a list of sample SPARQL queries based on the ESW ontology are available: `https://github.com/prapalu/esw/`.
- The Wikidata dump where the workflows were executed is available on the Wikidata archive: `wikidata-20210922-truthy-BETA.nt.gz`, while the Docker image with the scripts to download, clean, and ingest the archive is available at `https://github.com/prapalu/esw/tree/main/AnalyticalWorkload`.



## References


[1] A. Akritidis and Y. Tzitzikas. Querying knowledge graphs through positive and negative examples and feedback. *J. Intell. Inf. Syst.*, 62(5):1165–1186, Feb. 2024. ISSN 0925-9902. doi: 10.1007/s10844-024-00846-z. URL `https://doi.org/10.1007/s10844-024-00846-z`.

[2] G. Aluç, O. Hartig, M. T. Özsu, and K. Daudjee. Diversified stress testing of rdf data management systems. In *Proc. of the 13th International Semantic Web Conference (ISWC 2014)*, volume 8796, pages 197–212. Springer, 2014. doi: 10.1007/978-3-319-11964-9_13.

[3] R. Angles, C. B. Aranda, A. Hogan, C. Rojas, and D. Vrgoc. Wdbench: A wikidata graph query benchmark. In *Proc. of the 21st International Semantic Web Conference (ISWC 2022)*, volume 13489, pages 714–731. Springer, 2022. doi: 10.1007/978-3-031-19433-7_41.

[4] A. Bonifati, W. Martens, and T. Timm. An analytical study of large SPARQL query logs. 29(2-3):655–679, 2020. doi: 10.1007/s00778-019-00558-9.

[5] S. Ferré. Sparklis: An expressive query builder for sparql endpoints with guidance in natural language. *Semantic Web*, 8(3):405–418, 2017. doi: 10.3233/SW-150208.

[6] L. Gao, L. Golab, M. T. Özsu, and G. Aluç. Stream watdiv: A streaming rdf benchmark. In *Proceedings of the International Workshop on Semantic Big Data*, SBD'18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357791. doi: 10.1145/3208352.3208355. URL `https://doi.org/10.1145/3208352.3208355`.

[7] A. Ghose, K. Hose, M. Lissandrini, and B. W. Pedersen. An open source dataset and ontology for product footprinting. In *ESWC Satellite Events*, volume 11762, pages 75–79. Springer, 2019. doi: 10.1007/978-3-030-32327-1_15.

[8] E. Hyvönen and S. Ferré. Sparklis: An expressive query builder for sparql endpoints with guidance in natural language. *Semant. Web*, 8(3):405–418, Jan. 2017. ISSN 1570-0844. doi: 10.3233/SW-150208. URL `https://doi.org/10.3233/SW-150208`.

[9] S. Idreos, O. Papaemmanouil, and S. Chaudhuri. Overview of data exploration techniques. In *Proc of the International Conference on Management of Data (ACM SIGMOD 2015)*, pages 277–281. ACM Press, 2015. doi: 10.1145/2723372.2731084.

[10] M. Kejriwal. Knowledge graphs and covid-19: Opportunities, challenges, and implementation. *Harvard Data Science Review*, 11:300, 2020.

[11] M. Lissandrini, D. Mottin, T. Palpanas, and Y. Velegrakis. *Data Exploration Using Example-Based Methods*. Number 4. Morgan & Claypool Publishers, 2018. doi: 10.2200/S00881ED1V01Y201810DTM053.

[12] M. Lissandrini, D. Mottin, T. Palpanas, and Y. Velegrakis. Graph-query suggestions for knowledge graph exploration. In *The Web Conference 2020*, page 2549–2555. ACM Press, 2020. doi: 10.1145/3366423.3380005.

[13] M. Lissandrini, D. Mottin, K. Hose, and T. B. Pedersen. Knowledge graph exploration systems: are we lost? In *Proc. of the 12th Conference on Innovative Data Systems Research, CIDR 2022*, volume 22, pages 10–13, 2022.

[14] M. Lissandrini, G. Prando, and G. Silvello. Exploratory search workflows (esw) collection, 2023.

[15] G. Marchionini. Exploratory search: from finding to understanding. 49 (4):41–46, 2006. doi: 10.1145/1121949.1121979.

[16] M. Morsey, J. Lehmann, S. Auer, and A.-C. N. Ngomo. Dbpedia SPARQL benchmark - performance assessment with real queries on real data. In *Proc. of the 10th International Semantic Web Conference (ISWC 2011)*, volume 7031, pages 454–469. Springer, 2011. doi: 10.1007/978-3-642-25073-6_29.

[17] D. Mottin and E. Müller. Graph exploration: From users to large graphs. In *Proc. of the 2017 ACM International Conference on Management of Data, SIGMOD*, pages 1737–1740. ACM Press, 2017. doi: 10.1145/3035918.3054778.

[18] D. Mottin, M. Lissandrini, S. B. Roy, and Y. Velegrakis, editors. *Proc. of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA-Data 2021)*, volume 2929, 2021. CEUR-WS.org.

[19] O. Mussa, O. Rana, B. Goossens, P. Orozco Ter wengel, and C. Perera. Forestqb: Enhancing linked data exploration through graphical and conversational uis integration. *ACM J. Comput. Sustain. Soc.*, 2(3), Sept. 2024. doi: 10.1145/3675759. URL `https://doi.org/10.1145/3675759`.

[20] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-scale knowledge graphs: Lessons and challenges. 17(2):48–75, 2019.

[21] M.-E. Papadaki and Y. Tzitzikas. Unifying faceted search and analytics over rdf knowledge graphs. *Knowl. Inf. Syst.*, 66(7):3921–3958, Mar. 2024. ISSN 0219-1377. doi: 10.1007/s10115-024-02076-9. URL `https://doi.org/10.1007/s10115-024-02076-9`.

[22] T. Sagi, M. Lissandrini, T. B. Pedersen, and K. Hose. A design space for RDF data representations. 31(2):347–373, 2022. doi: 10.1007/s00778-021-00725-x.

[23] S. Sahu, A. Mhedhbi, S. Salihoglu, J. Lin, and M. T. Özsu. The ubiquity of large graphs and surprising challenges of graph processing: extended survey. 29(2-3):595–618, 2020. doi: 10.1007/s00778-019-00548-x.

[24] M. Saleem, M. I. Ali, A. Hogan, Q. Mehmood, and A.-C. N. Ngomo. LSQ: the linked SPARQL queries dataset. In *Proc. of the 14th International Semantic Web Conference (ISWC 2015)*, volume 9367, pages 261–269. Springer, 2015. doi: 10.1007/978-3-319-25010-6_15.

[25] M. Saleem, Q. Mehmood, and A.-C. Ngonga Ngomo. Feasible: A feature-based sparql benchmark generation framework. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14*, pages 52–69. Springer, 2015.

[26] M. Saleem, G. Szárnyas, F. Conrads, S. A. C. Bukhari, Q. Mehmood, and A.-C. N. Ngomo. How representative is a SPARQL benchmark? an analysis of RDF triplestore benchmarks. In *The World Wide Web Conference*, page 1623–1633. ACM Press, 2019. doi: 10.1145/3308558.3313556.

[27] S. Scheider, A. Degbelo, R. Lemmens, C. van Elzakker, P. Zimmerhof, N. Kostic, J. Jones, and G. Banhatti. Exploratory querying of sparql endpoints in space and time. *Semantic Web*, 8(1):65–86, 2017. doi: 10.3233/SW-150211.

[28] J. Sequeda and O. Lassila. Designing and building enterprise knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge*, 11(1): 1–165, 2021. doi: 10.2200/S01105ED1V01Y202105DSK020.

[29] C. Stadler, M. Saleem, Q. Mehmood, C. Buil-Aranda, M. Dumontier, A. Hogan, and A.-C. Ngonga Ngomo. Lsq 2.0: A linked dataset of sparql query logs. *Semantic Web*, (Preprint):1–23, 2024.

[30] X. Zhang, M. Wang, M. Saleem, A. N. Ngomo, G. Qi, and H. Wang. Revealing secrets in SPARQL session level. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference*, volume 12506 of *Lecture Notes in Computer Science*, pages 672–690. Springer, 2020.