DataLens: Enhancing Dataset Discovery via Network Topologies

Anaïs Ollagnier, Aline Menin

DataLens: Enhancing Dataset Discovery via Network Topologies

Anaïs Ollagnier and Aline Menin
Université Côte d'Azur, CNRS, Inria, I3S
Sophia Antipolis, France
{anais.ollagnier,aline.menin}@inria.fr

Abstract

The rapid growth of publicly available textual resources, such as lexicons and domain-specific corpora, presents challenges in efficiently identifying relevant resources. While repositories are emerging, they often lack advanced search and exploration features. Most search methods rely on keyword queries and metadata filtering, which require prior knowledge and fail to reveal connections between resources. To address this, we present DataLens, a web-based platform that combines faceted search with advanced visualization techniques to enhance resource discovery. DataLens offers network-based visualizations, where the network structure can be adapted to suit the specific analysis task. It also supports a chained views approach, enabling users to explore data from multiple perspectives. A formative user study involving six data practitioners revealed that users highly value visualization tools—especially network-based exploration—and offered insights to help refine our approach to better support dataset search.

1 Introduction

Advances in artificial intelligence and the open data initiative have promote a exponential growth of publicly available datasets. Numerous efforts have been made to improve dataset search, navigation, and discovery (Maier et al., 2014; Chapman et al., 2020; Paton et al., 2023). Recent studies have treated this task as an information retrieval problem, improving query handling, data management, and ranking (Dimitrakis et al., 2020; Mountantonakis and Tzitzikas, 2020; Castelo et al., 2021). Despite ongoing efforts to improve findability, accessibility, interoperability, and reusability (e.g., *RDA Data Discovery Paradigm Interest Group* [1], *FAIRsharing*[2]), significant challenges remain. Metadata quality is a major challenge when searching for machine learning (ML) resources (Chapman et al., 2020; Paullada et al., 2021). Despite the advances of ML (Skluzacek et al., 2022), web semantic (Feddoul et al., 2019), and metadata aggregation techniques to generate, structure, and enrich metadata, the uncertainty of generated metadata and general lack of standardization limit the potential of aggregation techniques as they hinder data integration, reduce search accuracy, and obscure meaningful relationships between datasets. Combined with traditional search methods that rely on keywords and filters to refine results, the sheer volume of information often makes exploration challenging, as suggested resources remain too numerous and obscure potential related datasets. Data visualization is well-known to aid human reasoning by using interactive tools that visually highlight and reveal these connections. Despite its wide use to support visual analysis of individual datasets or data points, their usage to support dataset discovery remains unexplored (Liu et al., 2014).

In this paper, we propose *Datalens*[3], a dataset search tool that integrates network-based visualizations and chained views to support the discovery of ML resources. Similar to existing approaches, we employ faceted search techniques to deal with large volumes of data. Then, using networks, our approach uncovers hidden relationships between ML resources, such as commonly supported ML tasks and models. Additionally, with chained views, the tool allow the user to narrow the exploration through multiple interconnected visualization techniques, each presenting a different perspective to the data. We demonstrate our approach through a set of use case scenarios and a formative evaluation involving 6 data practitioners.

[1] https://www.rd-alliance.org/groups/data-discovery-paradigms-ig
[2] https://fairsharing.org/

[3] *DataLens* is available at https://dataviz.i3s.unice.fr/datalens/

2 *DataLens* Overview

The *DataLens* approach integrates three key components: (i) a network customization panel that allows users to configure the network topology based on relationships pertinent to their search, (ii) a filtering panel that enables users to refine their focus on relevant datasets, and (iii) a visualization tool (MGExplorer (Menin et al., 2021)) that facilitates multi-perspective exploration of ML resources.

2.1 The Data

In this paper, we explore data from the Hugging Face catalogue[4] (Lhoest et al., 2021), which includes over 212,000 datasets in more than 8,000 languages, sourced from their API[5]. Each data record features descriptive information, such as the dataset's `modality` (indicating the type of data, e.g. text, audio, video), `task category` (defining the broad types of ML tasks the dataset supports), and `language`. One may need to find data for training models, or conversely, to identify models trained on specific datasets. Thus, we augmented the datasets' metadata with pre-trained models information, available on Hugging Face (see Listing 1).

Listing 1: Example dataset metadata.

```
{ "id": "amirveyseh/acronym_identification",
"author": "amirveyseh",
"created_at": "2022-03-02T23:29:22+00:00",
"last_modified": "2024-01-09T11:39:57+00:00",
"downloads": 154,
"likes": 19,
"paperswithcode_id": "acronym-identification",
"tags": [
  "task_categories:token-classification",
  "annotations_creators:expert-generated",
  "language_creators:found",
  "multilinguality:monolingual",
  "source_datasets:original",
  "language:en",
  "license:mit",
  "size_categories:10K<n<100K",
  "format:parquet",
  "modality:text",
  "library:datasets",
  "library:pandas",
  "library:mlcroissant",
  "library:polars",
  "model:deberta-v3-base-tasksource-nli",
  "model:deberta-v3-large-tasksource-nli",
  "model:deberta-v3-xsmall-tasksource-nli",
], ... }  ,
"description": "[...]" }
```

2.2 Interactive Customization of Network

Users can discover meaningful relationships through the *Network Topology Editor* (Fig. 1A). For instance, users can define nodes as `task_categories` interconnected by the `datasets` that support them, or configure nodes as `datasets` linked by shared `license`. The editor facilitates the task through interactive *combobox* menus, enabling users to select the variables that define the `source` and `target` nodes, the `link` between nodes, and the `thematic` information that describes the connections between nodes. For example, in a network where the nodes represent `task_categories` and the links represent `datasets`, thematic information characterizes the latter through aspects such as `license`, `language`, `format`, `modality`, and `size`.

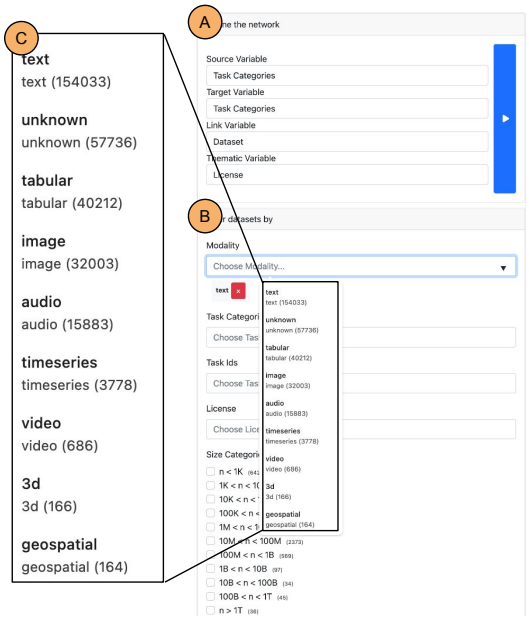

Figure 1: Network customization and faceted search. (A) a panel for customizing the graph topology, (B) a filtering panel for refining the search to focus on relevant datasets. Each search field in the filtering panel includes a list of potential values along with the corresponding number of associated datasets (C).

The *Filtering Panel* (Fig. 1B) helps users to focus on interesting information. For example, if a user is interested in exploring datasets that address at least a given ML task, such as *audio classification*, they can select this option from the combobox menus. The resulting network will then display only those datasets that support at least this task, while also illustrating the relationships with other tasks covered by the filtered datasets. In a more exploratory setting, where the user does not begin with a specific hypothesis, we assist users in discovering potentially relevant resources by displaying the count of datasets associated next to each value in the comboboxes (Fig. 1D). This allows users to get an overview of the significance of each feature,

[4] https://github.com/huggingface/datasets
[5] https://huggingface.co/docs/hub/api

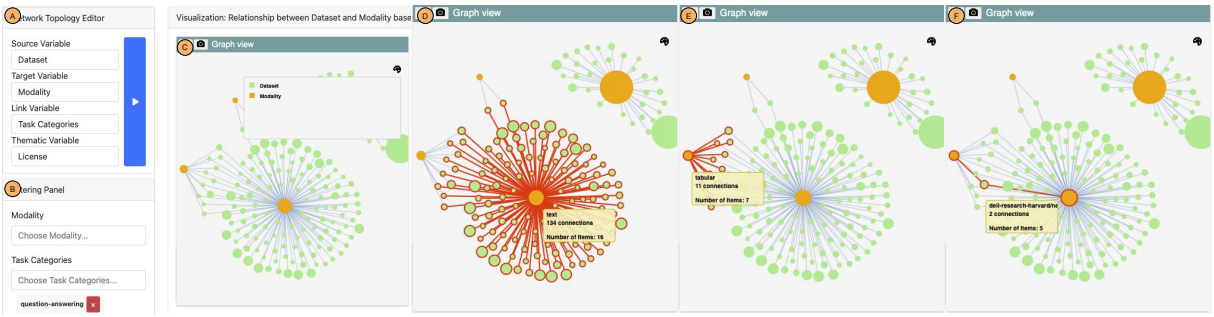

Figure 2: In Use Case Scenario 3.1: (A) *Dataset* serve as source variable and *Modality* as the target, linked together by shared *Task Categories*. (B) Data is filtered by selecting *question-answering* (QA) and dataset size as *1M<n<10M*. (C) The graph view depicts datasets and modalities as light-green and orange nodes, respectively. Node size reflects the count of associated tasks. (D-F) Hovering over the circles reveal detailed information such as the number of connections (datasets/modalities) and unique items (links/tasks).

based on how frequently it appears across datasets.

2.3 Multi-perspective Exploration

Our method builds upon MGExplorer's visualization capabilities (Menin et al., 2021). It facilitates data exploration using a variety of interconnected visualization techniques. These techniques include graph, pairwise relationship, temporal distribution, and listing views. Each view is interactively generated during exploration. The user right-click on an element of interest (e.g., a node in the graph view) and, through a contextual menu, selects and launches a new view to further explore the data. Each new view is then created using a filtered subset of the data, for example, only displaying information related to the selected node in the graph view. Relevant views will be presented and explained along with the use case scenarios hereafter.

3 Use Case Scenarios

3.1 Exploring Multimodal Datasets for Performing the Question-Answering Task

Alice is developing a new question-answering (QA) model for a chatbot application. To find relevant datasets for training, she aims to explore datasets that support the QA task while also considering their use across different modalities (e.g., text and tabular) to expand her model's applicability. She begins her exploration by defining a network topology that aligns with her task. She sets `dataset` and `modality` as nodes (i.e. source and target variables, respectively), with `task categories` as links (Fig.2A). Since her focus is on QA, she filters the data by selecting `question-answering` under *Task Categories* and further refines her selection by choosing datasets with a size between 1M and 10M in the filtering panel (Fig.2B). This configuration enables her to visualize the relationship between datasets and data modalities based on shared tasks, including at least QA.

Alice starts the visualization by clicking *play*, generating a bipartite graph (Fig. 2C) that links datasets (light green) to the modalities (orange) they support for QA or other related ML tasks. Using the hover feature, she explores the orange nodes to examine the associated modalities. She observes that the `text` modality has the highest number of connections in the network (Fig. 2D). The tooltip reveals that `text` is supported by 134 datasets (i.e., connected nodes) through 16 items (tasks), with 15 potentially related to `QA`. Additionally, she identifies a shared modality among these datasets. As illustrated in Fig. 2E, the `tabular` modality is linked to 11 datasets, all of which also connect to the `text` modality. The tooltip indicates that these datasets span 7 different tasks, supported in both `text` and `tabular` modalities. One example of a dataset facilitating QA-related tasks across both modalities is `dell-research-harvard/newswire` (Fig. 2F).

3.2 Expanding QA Applications by Exploring Datasets for Multiple Tasks

To explore additional QA-related tasks for extending her model, Alice returns to the network topology editor and sets `task categories` as nodes (i.e. both source and target variables) and `dataset` as links, with `license` as the thematic attribute (Fig. 4A). She keeps the `question-answering` filter for *Task Categories*, and deletes the size-based filter (Fig. 4B). This configuration allows her to visualize shared QA-related tasks based on the datasets that support them, regardless of their size.

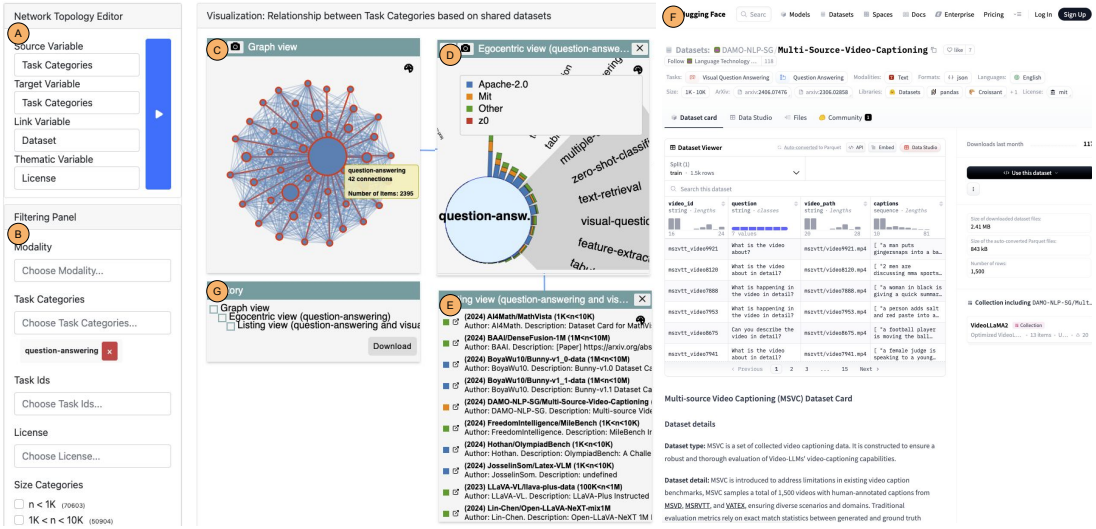

Figure 3: Overview of Use Case Scenario 3.2. (A) Task categories are represented as nodes, linked by datasets characterized by their distribution license. (B) Metadata is filtered to retain datasets covering at least the question-answering task. (C) The Graph View displays the network, with node size indicating the number of supporting datasets. (D) The Egocentric View enables focused, pairwise exploration of task relationships. (E) The Listing View presents datasets supporting both QA and visual-question-answering. (F) Direct dataset inspection at the source.

The generated network represents tasks as nodes, connected when at least one dataset supports them. Since the data is filtered by the QA task, this task appears as the largest node, with its size reflecting the number of supporting datasets. By hovering over the QA node, Alice discovers that these datasets also support 42 other distinct tasks (Fig. 4C). She explores this subset further by right-clicking on the QA node and opening the *Egocentric View* (Fig. 4D).

The *Egocentric View* places the selected node (`question-answering`) at the center, displaying its connected nodes along a peripheral arc and highlighting relationships in a pairwise manner. Between each pair of nodes, a bar represents the number of shared links (i.e., datasets), with its length indicating the strength of the connection and its color segments reflecting the selected thematic attribute (e.g., distribution license).

Alice is particularly interested in datasets that support both QA and `visual-question-answering` (VQA). To investigate further, she right-clicks on the bar linking these two tasks and opens the *Listing View*, which displays the datasets supporting both QA and VQA (Fig. 4E). Since she exclusively works with MIT-licensed datasets, she visually scans the list for a dataset marked with the corresponding color (5th in the list). Finally, by clicking the external link icon, she accesses the dataset's page on the source platform (e.g., Hugging Face) to inspect and download it (Fig. 4F).

3.3 Uncovering Dataset Relationships Through Shared Pretrained Models

Bob aims to train his own models for audio-based applications by identifying benchmark datasets based on their training usage across multiple pretrained models. He begins his exploration by defining a network topology that aligns with his objective. He sets `datasets` as (source and target) nodes and `models` as links (Fig.4A). To refine his analysis, Bob filters the data to include only datasets categorized under `audio` (Fig.4B).

Bob launches the visualization by clicking *play*, generating a graph that connects datasets based on shared models. He initially observes many datasets with no connections, indicating that their associated models were not trained using any other datasets. By observing the node sizes, Bob identifies promising datasets that have been used to train over 800 models—for example, `snousind/common_voice` (Fig. 4C), which has contributed to training 875 models. He inspects this dataset using the *Listing View* (Fig. 4D), which reveals the list of models trained using this dataset. By clicking the external link icon next to one of the models, Bob can access detailed information directly on Hugging Face, allowing him to verify metadata and gather additional insights (Fig.4E).

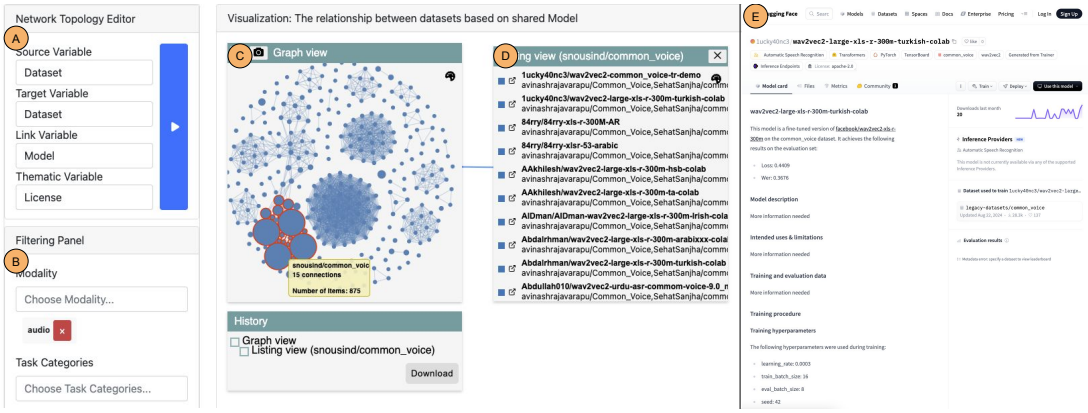

Figure 4: Use Case Scenario 3.2. (A) Datasets are nodes, models are links. (B) Filtering by *audio* modality. (C) Node size shows the number of models trained by each dataset. (D) Models trained with `snousind/common_voice`. (E) Direct model inspection at the source.

4 Formative Evaluation

We conducted a formative evaluation with six data practitioners specializing in NLP, Knowledge Engineering, and Software Engineering. Their common challenge was finding suitable datasets for research. Participants included those searching for ML datasets, open datasets, and coding datasets. Four had over seven years of experience on their research domain, while the others had one to three years. The evaluation consisted of three phases (see Appendix A): (i) a questionnaire on participants' needs, habits, and challenges in dataset search, (ii) hands-on use of Datalens to answer aforementioned scenario-based questions, such as identifying datasets supporting both text and tabular modalities, and (iii) a final questionnaire assessing usability, user satisfaction, and improvement suggestions.

User Needs, Habits and Challenges: Participants primarily search for datasets in repositories like Hugging Face and Kaggle, academic papers, and code repositories (e.g., GitHub). Key criteria for selection include domain, modality, license, size, supported tasks, language, and models, while aspects like related publications, source, popularity, and publication date are considered less important. Data quality and curation were also emphasized as important aspects when searching datasets. Participants commonly face difficulties in finding relevant datasets due to issues like label mismatches, noisy data, inconsistent definitions, partial availability, and dataset discrepancies. Assessing dataset quality often requires extensive manual inspection, and data cleansing can be time-consuming. Most participants were unfamiliar with using visualization tools in dataset searches, with mixed opinions on its potential usefulness. When asked about desired features in an exploration tool, participants expressed a need for standardized, comprehensive listings of dataset characteristics, precise search capabilities, dataset previews, metadata exploration, and insights into usage by others. They would also like to have dataset comparison, similarity search, and visualizations to identify gaps or patterns. Linking datasets to related papers, code, and models was considered valuable for context and validation. Datalens addresses some of these needs by revealing similarities between datasets and supporting comparisons through visualizations.

Datalens' Usability and Relevance: We used the well-known SUS questionnaire (Brooke, 1996) to assess the usability of our tool, which received an average score of 61.66/100. The positive aspects highlighted in the feedback included good feature integration and participants' willingness to use the tool regularly. However, improvements are needed to enhance accessibility and ease of use. Participants noted that the chosen visualization tool, MGExplorer, while useful, may be too complex for its intended purpose. Additionally, they felt a need to learn many aspects of the system before use, which increased the learning curve. Despite these challenges, participants were generally neutral about the difficulty of completing the tasks, with one finding the third task particularly difficult. All participants agreed on the usefulness of the tool's features (visualization, network editor, and faceted search), with the network topology editor being especially helpful for most tasks.

User Appreciation and Improvements: Par-

ticipants appreciated Datalens for its intuitive graph view, smooth transitions, customizable multi-window organization, and ability to explore the search space. They valued its responsiveness, powerful filters, and the network topology editor and egocentric view. The tool effectively visualizes links between datasets, models, and tasks, with cardinality estimation. Key strengths included overview-zoom-filter support, a graph-based overview, and direct links to datasets/models. Participants also liked configuring multiple search variables and text-based graph searches. However, they raised concerns about unclear parameters, lack of linking-and-brushing, and the small default window. Labels were seen as opaque, with a request for icons. For some tasks (i.e. Scenario 3.2), graph visualization was distracting, with a preference for list-based views. Dense graphs and zooming were unhelpful for focusing on nodes, and participants preferred single clicks. Memorizing names and interface complexity were issues. Two separate search views were confusing, and zooming needed improvement. Participants suggested improvements like a page with more details, enabling the AND operator in the browser, larger graph view, and enhanced zooming to improve node selection.

5 Discussion and Future Work

In this paper, we presented *Datalens*, a visualization approach to support dataset search leveraging multi-faceted search and network-based visualizations. We address dataset search by revealing hidden relationships through graph topologies, which expose patterns, structures, and interconnections. Beyond exact matches, our approach highlights links and overlaps between communities, helping users identify reusable and transferable datasets. We demonstrated the effectiveness of *Datalens* through use case scenarios and a formative evaluation with data practitioners using HuggingFace dataset catalog, as a case study.

While graph topologies reveal hidden links between datasets, they alone may not suffice for dataset search. As shown in our use case scenarios, multiple data dimensions and attributes require viewing the data from different perspectives. Thus, we provide multiple visualization techniques that present the relationships in complementary ways, while allowing users to refine their search, transitioning from one view to another. Our formative evaluation has confirmed the promising aspect of visualization to support dataset search, particularly, through the network visualization but also via an egocentric visualization showing detailed pairwise relationships between datasets. Participants appreciate the use of chained views to support the discovery process, but noticed that the tool (i.e. MG-Explorer) might be too complex to the intended goal. Furthermore, the tool suffers from scalability issues, as the network can become rapidly cluttered when there are too many nodes and links to be represented. Furthermore, while the visualizations are useful, starting with a network was not always seen as necessary during the evaluation. Thus, future work includes the investigation of aggregation and interaction methods to tackle scalability issues, as well as studying the relevance of visualization techniques to support different dataset search tasks.

We developed *Datalens* in a manner that it adapts itself based on available metadata. Thus, while we currently applied our approach to datasets published on Hugging Face, one can easily extend and apply the tool to any other dataset repository, such as *Kaggle*. In the spirit of open science, the source code is publicly accessible on GitHub[6]. As future work, to extend the coverage of *Datalens* and support dataset search in other domains, such as software engineering, we intend to dynamically integrate other dataset repositories.

Metadata quality issue remains a challenge in dataset discovery. For example, in the Hugging Face catalog, labels are only partially standardized, allowing users to add custom descriptions, which leads to heterogeneous dataset profiling. Furthermore, there are multiple attributes that do not have associated values, which hinders dataset discovery as we cannot classify it easily (e.g., over 57.7k datasets do not have an associated modality). To address this issue and enhance interoperability, future work includes integrating domain-specific metadata standards with controlled vocabularies, as recommended by the RDA Data Maturity Model[7]. We also intend to investigate the usage of semantic web technologies to model and represent dataset metadata in a inherently interoperable way. Additionally, we aim to enrich metadata by answering key questions (e.g., What is the dataset for? Who is it for? Why and when is it used?), focusing on data seekers' needs.

[6] https://github.com/amenin/datalens
[7] https://www.rd-alliance.org/

References

john Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press. Num Pages: 6.

Sonia Castelo, Rémi Rampin, Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proc. VLDB Endow.*, 14(12):2791–2794.

Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *VLDB J.*, 29(1):251–272.

Eleftherios Dimitrakis, Konstantinos Sgontzos, and Yannis Tzitzikas. 2020. A survey on question answering systems over linked data and documents. *J. Intell. Inf. Syst.*, 55(2):233–259.

Leila Feddoul, Sirko Schindler, and Frank Löffler. 2019. Automatic facet generation and selection over knowledge graphs. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 310–325, Cham. Springer International Publishing.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

Tianyang Liu, Fatma Bouali, and Gilles Venturini. 2014. EXOD: A tool for building and exploring a large graph of open datasets. *Comput. Graph.*, 39:117–130.

David Maier, V. M. Megler, and Kristin Tufte. 2014. Challenges for dataset search. In *Database Systems for Advanced Applications - 19th International Conference, DASFAA 2014, Bali, Indonesia, April 21-24, 2014. Proceedings, Part I*, volume 8421 of *Lecture Notes in Computer Science*, pages 1–15. Springer.

Aline Menin, Ricardo Cava, Carla Maria Dal Sasso Freitas, Olivier Corby, and Marco Winckler. 2021. Towards a visual approach for representing analytical provenance in exploration processes. In *25th International Conference Information Visualisation (IV)*.

Michalis Mountantonakis and Yannis Tzitzikas. 2020. Content-based union and complement metrics for dataset search over RDF knowledge graphs. *ACM J. Data Inf. Qual.*, 12(2):10:1–10:31.

Norman W. Paton, Jiaoyan Chen, and Zhenyu Wu. 2023. Dataset discovery and exploration: A survey. *ACM Comput. Surv.* Just Accepted.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Tyler J. Skluzacek, Matthew Chen, Erica Hsu, Kyle Chard, and Ian Foster. 2022. Models and metrics for mining meaningful metadata. In *Computational Science – ICCS 2022*, pages 417–430, Cham. Springer International Publishing.

A Questionnaire/Protocol used during the Formative Evaluation

A.1 Terms and Conditions Agreement

The goal of this user study is to assess the capabilities of DataLens, a web-based platform that integrates faceted search with advanced information visualization techniques, to facilitate the search and exploration of machine learning (ML) datasets. Specifically, we aim to explore the following:

Identifying user tasks and needs, with a focus on the types of queries and attributes essential for effectively searching and selecting ML datasets.

Evaluating the added value of leveraging both customizable network topologies and multi-perspective visualizations to effectively meet the needs of data practitioners.

Investigating the benefits of visual exploration of dataset networks through information visualization techniques. In this study, we use the visualization tool MGExplorer.

This study will last around 40 minutes and will be structured as follows. We will start with a few questions to learn about your profile, habits, and needs regarding the usage and exploration of ML datasets. Next, we will introduce you to the Datalens interface and its main features. Then, we will ask you to complete three exploratory tasks using either the Datalens interface or the HuggingFace search feature to explore datasets from the HuggingFace catalog. Lastly, we will collect your feedback and comments on the tool's capabilities.

Participation in this study is anonymous. The collected data will be used in this study only. You may withdraw at any time without providing a reason. If you choose to do so, your data will be immediately deleted and will not be used in this study.

Please, check the option below if you consent to participate in this study according to the conditions presented above.

◯ I consent to participate in this study.

A.2 Pre-questionnaire

1. What is your main research area?

2. How many years of experience do you have in your field?
   ◯ Less than 1 year ◯ Between 1 and 3 years ◯ Between 4 and 6 years ◯ Over 7 years

3. How often do you search for ML datasets? (5-point Likert scale)

4. Where to you usually search for datasets?
   ☐ Dataset Repositories (e.g. Kaggle, HuggingFace) ☐ Academic Papers ☐ Code repositories (e.g. Github) ☐ Personal or professional networks

5. How important are the following elements to you when searching for datasets? (5-point Likert scale)
   - The domain or application area of the dataset.
   - The size of the dataset (e.g., number of records, file size).
   - The language(s) represented in the dataset.
   - The modality of the dataset (e.g., text, audio, image).
   - Licensing information (e.g., open-source, commercial).
   - The time of data collection.
   - The source of dataset (e.g. institution, authors, etc.).
   - The popularity of the dataset (e.g. likes, downloads).
   - Tasks for which the dataset has been applied.
   - Scientific publications citing or utilizing the dataset.
   - Models trained using the dataset.
   - Temporal trends in dataset usage.

6. Are there any other factors that you consider important when searching for datasets? Please specify.

7. How often do you encounter problems when searching for datasets? (5-point Likert scale)

8. What king of problems do you encounter? If applicable.

9. How often do you use visualization tools to search datasets? (5-point Likert scale)

10. How important do you consider visualization techniques to support dataset search? (5-point Likert scale)

11. What features would you like to see in a dataset search and exploration tool?

A.3 Hands-on use of Datalens

Go to Datalens: http://dataviz.i3s.unice.fr/datalens/explorer.

Task 1: Identifying Suitable Datasets for a QA Chatbot

Your objective is to find datasets for developing a Question Answering (QA) model for a chatbot. Specifically, you want to explore available datasets and their modalities (e.g., text or audio) to define how users will interact with the chatbot. Follow the steps below:

1. Configure the Network Topology as follows:
    - **Source variable**: Dataset
    - **Target variable**: Modality
    - **Link variable**: Task Categories
    - **Thematic attribute**: default (not applicable)
2. To keep a set of interesting datasets to explore, filter the metadata to consider only QA tasks and relatively large datasets. For that, select the following filters:
    - **Task Categories**: question-answering
    - **Size Categories**: $1M < n < 10M$
3. Click on the **play** (blue button) to launch the visualization.

Question: Can you identify 2 datasets that you can use for both tabular and textual data?

Task 2: Identifying Related Tasks to Expand your Model's Applications

Still within the goal of developing a QA model for a chatbot application, your objective now is to investigate related or adjacent sub-tasks to expand your model's potential applications. Follow the steps below:

1. Configure the Network Topology as follows:
    - **Source variable**: Task Categories
    - **Target variable**: Task Categories
    - **Link variable**: Dataset
    - **Thematic attribute**: License
2. To keep a set of interesting datasets to explore, filter the metadata to consider only QA tasks and expand the search to include all possible sizes of datasets. For that, select the following filter:
    - **Task Categories**: question-answering
3. Click on the **play** (blue button) to launch the visualization.

Question: Can you find 2 datasets that support both Question Answering (QA) and Visual Question Answering (VQA) and that are available under the MIT License? Share below the links to their descriptions on HuggingFace.

Task 3: Identifying Relevant Datasets based on the Models that use them

In this task, your goal changes as you aim at improving your model selection process for audio-based applications. For that, you will try to identify the most popular datasets, based on their shared usage across models. Follow the steps below:

1. Configure the Network Topology as follows:
    - **Source variable**: Dataset
    - **Target variable**: Dataset
    - **Link variable**: Model
    - **Thematic attribute**: default (not applicable)
2. To keep a set of interesting datasets to explore, filter the metadata to consider only datasets within the audio modality. For that, select the following filter:
    - **Modality**: audio
3. Click on the **play** (blue button) to launch the visualization.

Question: Can you identify one widely used dataset for training models (i.e. one associated to many models) and three models that use it? Provide below the links to the models' descriptions on HuggingFace.

Post-task questions (after each task)

- How do you evaluate the difficulty of this task? (5-point Likert scale)
- How useful did you find the following elements to solve this task? Network topology, filters, visualizations (5-point Likert scale)

A.4 Post-questionnaire

1. SUS questionnaire (Brooke, 1996)
2. Please provide 3 things that you **liked** about the experience or the tool.
3. Please provide 3 things that you **disliked** about the experience or the tool.
4. What enhancements or new features would you recommend to improve the tool's usefulness and user experience?