# Auctus: A Search Engine for Data Discovery and Augmentation

Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire
New York University
{s.castelo,remi.rampin,aecio.santos,aline.bessa,fchirigati,juliana.freire}@nyu.edu

## ABSTRACT

The large volumes of structured data currently available open up new opportunities for progress in answering many important scientific, societal, and business questions. However, finding relevant data is difficult. While search engines have addressed this problem for Web documents, there are many new challenges involved in supporting the discovery of structured data for specific tasks. To tackle these challenges, we propose the dataset search engine Auctus. In this paper, we describe Auctus and present open questions and future work related to dataset discovery.

**Reference Format:**
Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. Auctus: A Search Engine for Data Discovery and Augmentation. In the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA Data 2021).

## 1 INTRODUCTION

While data are abundant, given the large number of datasets spread over a large number of sites and repositories, finding *relevant data* for a given task is difficult. Recognizing this challenge, a number of approaches have been proposed to *organize and index data collections* [3]. While these present a significant step towards simplifying data discovery, they have an important limitation: they only support keyword-based search queries over published dataset metadata. In addition, published metadata is often incomplete, and in many cases it is inconsistent with the actual data. Thus, relying solely on the metadata also hampers the discoverability of datasets. In this work we describe Auctus, a system we propose to tackle these limitations. We also introduce a number of open questions related to the problem of dataset discovery, as well as future work related to Auctus.

## 2 THE AUCTUS SYSTEM

Auctus [1, 2] is an *open-source dataset search engine designed to support data discovery and augmentation*. In addition to keyword-based search, the system *supports a rich set of queries* including spatial and temporal queries, as well as data integration and augmentation queries. These queries are enabled in part by a *data profiler that automatically extracts metadata* from the actual datasets. The profiler generates summaries (or sketches) of column contents and data types which are to construct indices that support efficient query evaluation. Users can explore large dataset collections through an *intuitive interface*. To help users identify relevant datasets, Auctus displays snippets that summarize the contents of datasets.


Published in the Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, co-located with VLDB 2021 (August 16-20, 2021, Copenhagen, Denmark) on CEUR-WS.org.

Auctus was *implemented with scalability in mind*: the system is containerized using Docker [5]. Each data discovery plugin corresponds to an independent container, allowing multiple plugins to be executed in parallel. Auctus can also spin up as many profiling and query containers as required in response to load. Users can access the system via a Web UI or programmatically via Python and REST APIs. Auctus has been successfully deployed and is currently used by different research groups within the DARPA D3M program [4, 7]. We refer the reader to [2] for additional details about the system and its architecture.

## 3 FUTURE WORK AND OPEN QUESTIONS

*Messy data.* In addition to the lack metadata, real data is messy and noisy. The Auctus data profiler represents a first step at identifying semantic types as well as summarizing datasets to support integration queries. However the recall and precision of discovery queries can be negatively affected by the presence of data quality issues. Robust and automated techniques are needed to automate data cleaning, discover semantic types, and support approximate join and union queries.

*Correlated data discovery.* One of the original motivations for Auctus was to support data augmentation to improve machine learning models. For this task, given a large collection of tabular datasets $C$ and a query table $Q$, we need to identify all datasets $d_i \in C$ that are both joinable with $Q$ and that contain an attribute that is correlated with the target variable in $Q$. However, computing these joins and correlations in real-time is not feasible for large tables and dataset collections. We have recently proposed a new sketching-based method that support the efficient evaluation of join-correlation queries [6]. To integrate this method with Auctus and further improve query efficiency, it would be interesting to explore techniques that have been successfully used to speed-up web search queries, as well as the use of locality-sensitive hashing (LSH).

*User interfaces for data discovery.* User interfaces for dataset search is a rather unexplored research area. We posit this is due the prior unavailability of efficient algorithms and systems for building dataset search engines. There are many open questions in how can we present dataset search results to users so that they can make sense of the data and efficiently and effectively perform relevance judgements about the suitability of the data for their task.

*Result ranking.* Different information needs and associated discovery tasks demand different ranking strategies – there is no one-fits-all strategy. Moreover, determining whether a dataset is better than another can be difficult even for a fixed task. Furthermore, datasets have other properties that contribute to their value, including the publisher (e.g., datasets published by reputable sources can be considered 'better' than datasets from unknown sources) or intrinsic quality measures (e.g., the number of NULL values). Research is needed both to better understand ranking in the context of structured datasets and to devise effective strategies.

## ACKNOWLEDGMENTS

This work was partially supported by the DARPA D3M program and NSF award OAC-1640864. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and DARPA.

## REFERENCES

[1] Auctus: Github Repository 2021. https://github.com/VIDA-NYU/auctus.

[2] Sonia Castelo, Remi Rampin, Aecio Santos, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A Dataset Search Engine for Data Discovery and Augmentation. In *Proceedings of the 47th International Conference on Very Large Data Bases (to appear)*.

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

[4] Data-Driven Discovery of Models (D3M). 2019. https://www.darpa.mil/program/data-driven-discovery-of-models.

[5] Docker. 2021. https://www.docker.com/.

[6] Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation Sketches for Approximate Join-Correlation Queries. In *International Conference on Management of Data (SIGMOD)*. 1531–1544.

[7] Aécio S. R. Santos, Sonia Castelo, Cristian Felix, Jorge Piazentin Ono, Bowen Yu, Sungsoo Ray Hong, Cláudio T. Silva, Enrico Bertini, and Juliana Freire. 2019. Visus: An Interactive System for Automatic Machine Learning Model Building and Curation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA@SIGMOD)*. 6:1–6:7.