DataGEMS

Data Discovery Platform with Generalized Exploratory, Management, and Search Capabilities



OTHER

D3.3
D3.3a Data Discovery Services

PU

28/11/2025

Disclaimer

This document contains description of the DataGEMS project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) protection rules so, prior to using its content please consult carefully the licensing information of this document and/or contact the Project and the Editor for any further information.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify the Project and the Editor immediately.

The editor and authors of this document have taken any available measure in order to ensure that its content is accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

This project has received funding from the European Union’s Horizon Europe framework programme under grant agreement No. 101188416.



Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency, neither of which can be held responsible for those views and opinions and are not liable for any use that may be made of the information contained therein.

Document Information

Type	OTHER
Identifier	D3.3
Title	D3.3a Data Discovery Services
Dissemination level	PU
Due Date	30/11/2025
Actual Delivery Date	28/11/2025
Lead Beneficiary	ARC
Other beneficiaries	ZHAW, CNRS
Editor	Anna Mitsopoulou (ARC)
Authors	George Katsogiannis (ARC), Antonis Mandamadiotis (ARC), Yi Zhang (ZHAW), Kurt Stockinger (ZHAW), Georgia Troullinou (CNRS)
Contributors	Mike Xydas (ARC), Elzbieta Szafarska (IDG)
Reviewer	George Kakaletris, Giorgos Papanikos (CITE)
Workpackage(s)	WP3 – Dataset Discovery
Keywords	Software, Documentation, In-Dataset, Cross-Dataset, Packaging, Analytics Patterns
Version	1.0
Version information	Number of pages: 15 Last file save: 28/11/2025 9:45 PM File revision: 9 File name: D3.3_M11_D3.3a Dataset Discovery Services.docx

Copyright Information

License: Creative Commons Attribution (CC-BY)
This document is released under the Creative Commons Attribution (CC-BY) license, which allows for the free use, distribution, and adaptation of the work, provided proper attribution is given to the original author(s). By accessing or using this report, you acknowledge and agree to comply with the terms and conditions of the CC-BY license. For the full text of the license, please visit: https://creativecommons.org/licenses/by/4.0/legalcode

Document Version

Version	Date	Change Description	Editor
1.0	28/11/2025	Submitted to EU	Anna Mitsopoulou

Document Tables

Table of Tables

Table 1 In-Dataset Discovery Information 9
Table 2 Cross-Dataset Discovery Information 10
Table 3 Dataset Packaging Information 12
Table 4 Data Analytics Patterns Information 12

Table of Figures

Figure 1 Dataset Discovery Components in the DataGEMS Architecture (highlighted) 8
Figure 2 An example response to a search request, from the Cross-Dataset Discovery API 11

Table of Images

No table of image entries found.

Table of Contents

Disclaimer 2
Document Information 3
Copyright Information 3
Document Version 4
Document Tables 5
Table of Contents 6
Abstract 7
1 Introduction 8
2 Software Components by WP3 9
2.1 In-Dataset Discovery Component 9
2.1.1 Component Overview and Location 9
2.1.2 In-Dataset Discovery in Action 9
2.2 Cross-Dataset Discovery Component 9
2.2.1 Component Overview and Location 9
2.2.2 Cross-Dataset Discovery in Action 10
2.3 Dataset Packaging Component 11
2.3.1 Component Overview and Location 11
2.3.2 Dataset Packaging in Action 12
2.4 Data Analytics Patterns 12
2.4.1 Component Overview and Location 12
3 Conclusions 13
Appendix I List of acronyms 14
Appendix II References 15

Abstract

The Data Discovery Services software deliverable provides information on the implementation of the dataset discovery components described in D3.1 Dataset Discovery Architecture Report. For each component, it includes a brief overview, the link of the corresponding source code and the documentation. Where applicable, the deliverable also presents examples of how the component is used within the DataGEMS platform.

1 Introduction

This document provides a brief description and the location of the implementation and the documentation of the Dataset Discovery components. All software implementations are made available in the general DataGEMS GitHub repository. The detailed descriptions of the Dataset Discovery components, highlighted in Figure 1, can be found in D3.1 Dataset Discovery Architecture Report.

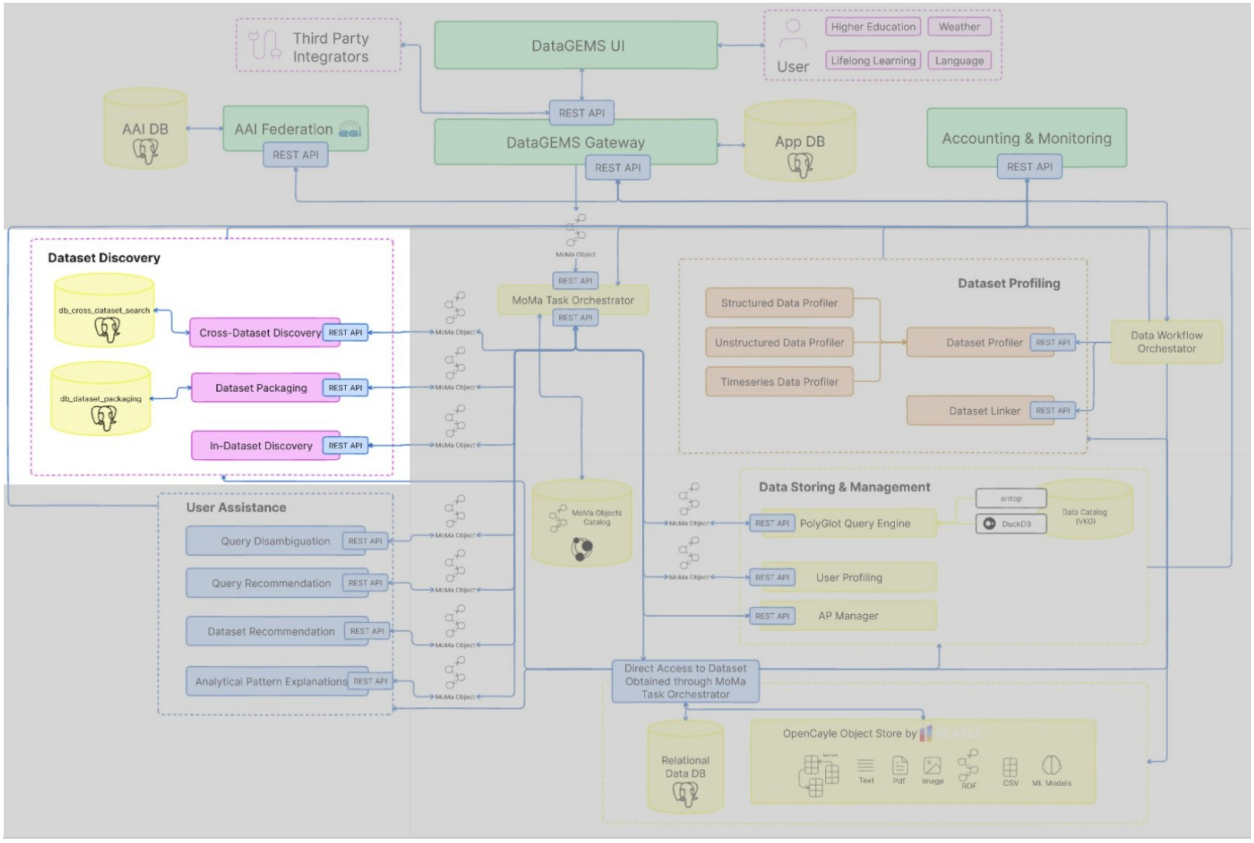


Figure 1 Dataset Discovery Components in the DataGEMS Architecture (highlighted)

The document is structured as follows:

- Section 2.1 covers the In-Dataset Discovery task, which enables content-level exploration of the datasets, by answering questions asked in natural language and in other forms, posed over datasets.
- Section 2.2 provides the information regarding the Cross-Dataset Discovery task, which identifies the datasets from the platform's catalogue that best match the user needs, provided in the form of a natural language question.
- Section 2.3 contains the information about the Dataset Packaging task, which groups datasets into packages that more effectively address the user's analytical intent.
- Finally, Section 2.4 covers the Data Analytics Pattern task, which unifies the DataGEMS components by structuring their processes into reusable, optimizable patterns.

2 Software Components by WP3

2.1 In-Dataset Discovery Component

2.1.1 Component Overview and Location

The **In-Dataset Discovery** component enables querying data within a specific dataset, e.g. a database or text documents, with query operators or in natural language without the need to know the schema of the database to be accessed, nor the database-specific query language or API. In case the data is stored in a relational database, the natural language question is automatically translated into SQL.

In the current implementation we have mainly addressed relational data, and we are covering the following functionality:

- text2geo: Translation of a natural language question to a geolocation object.
- text2SQL: Translation of a natural language question to a SQL query.

In the future, the system will be extended to cover question answering over text documents and images.

Table 1 In-Dataset Discovery Information

In-Dataset Discovery Component	
Service Name	In-dataset-discovery
Repository	https://github.com/datagems-eosc/In-dataset-discovery
Documentation	https://datagems-eosc.github.io/in-dataset-discovery-docs
Visibility (currently)	PRIVATE
Visibility (planned)	PUBLIC
LICENSE	MIT License

2.1.2 In-Dataset Discovery in Action

For details on how to use the services, see Deliverable D3.1 Dataset Discovery Architecture Report.

2.2 Cross-Dataset Discovery Component

2.2.1 Component Overview and Location

Cross-Dataset Discovery is responsible for identifying and retrieving useful datasets in response to Natural Language Questions (NLQs) submitted by a user. To achieve this, the component keeps track of all the datasets that are available in the platform by storing them in a dedicated index that is only used by this component. The use of a specialized index is necessary due to the large number of datasets that are expected to be available and the low latency requirements by our end users. When a question is given, the component uses the index to quickly search for relevant

datasets, which are then re-ranked and given to the user in descending order of relevance or usefulness in relation to their question.

Currently, the component is implemented using a BM25[1] index that stores the contents and metadata information of all datasets in the platform. When a question is posed, it is searched in the index and the final response uses the BM25 relevance scores calculated by the index (i.e., no re-ranking is currently performed). In the future, the component will be extended with a dense index (i.e., using dense vector representations that capture the semantics of data instead of just BM25) and a re-ranking step that will provide a better ordering of the datasets retrieved by the index. Additionally, the index will be extended with additional information that will be generated for all datasets (e.g., summaries, sample questions, named entities, etc.).

Table 2 Cross-Dataset Discovery Information

Cross-Dataset Discovery Component	
Service Name	cross-dataset-discovery
Repository	https://github.com/datagems-eosc/cross-dataset-discovery
Documentation	https://datagems-eosc.github.io/cross-dataset-discovery-docs
Visibility (currently)	PRIVATE
Visibility (planned)	PUBLIC
LICENSE	MIT License

2.2.2 Cross-Dataset Discovery in Action

Figure 2 shows the response of the Cross-Dataset Discovery Component to an API search request with the Natural Language Question “What is the fundamental theorem of calculus?” for the top 3 most relevant results. We observe that the service’s response contains the time the query took to complete, as well as a list of 3 results. Each result contains the piece of content that was found to be relevant to the question, the id of the dataset that it belongs to, the id of the dataset object it belons to (e.g., which file of the dataset), and a similarity score that allows us to rank the results from most to least relevant. This format allows us to present the relevant dataset to the user, along with the exact part of the dataset that is relevant to their question, if needed, facilitating data discovery. In fact, we observe that the response to the user’s question can be found in the highest-ranked result.

[1] https://en.wikipedia.org/wiki/Okapi_BM25 (BM25 = Best Match 25)

```
{
  "query_time": 126.59,
  "results": [
    {
      "content": "The fundamental theorem of calculus states that differentiation and integration are inverse operations. More precisely, it relates the values of antiderivatives to definite integrals. Because it is usually easier to compute an antiderivative than to apply the definition of a definite integral, the fundamental theorem of calculus provides a practical way of computing definite integrals. It can also be interpreted as a precise statement of the fact that differentiation is the inverse of integration.",
      "dataset_id": "1f6fba0c-9aea-4345-b5a3-457c924f9e0c",
      "object_id": "wiki_en_11340",
      "similarity": 14.658599853515625
    },
    {
      "content": "Differential calculus and integral calculus are connected by the fundamental theorem of calculus. This states that differentiation is the reverse process to integration. Differentiation has applications in nearly all quantitative disciplines.",
      "dataset_id": "1f6fba0c-9aea-4345-b5a3-457c924f9e0c",
      "object_id": "wiki_en_8497",
      "similarity": 14.205300331115723
    },
    {
      "content": "Calculus is the mathematical study of continuous change, in the same way that geometry is the study of shape, and algebra is the study of generalizations of arithmetic operations. Originally called infinitesimal calculus or \"the calculus of infinitesimals\", it has two major branches, differential calculus and integral calculus. The former concerns instantaneous rates of change, and the slopes of curves, while the latter concerns accumulation of quantities, and areas under or between curves. These two branches are related to each other by the fundamental theorem of calculus. They make use of the fundamental notions of convergence of infinite sequences and infinite series to a well-defined limit. It is the \"mathematical backbone\" for dealing with problems where variables change with time or another reference variable.",
      "dataset_id": "1f6fba0c-9aea-4345-b5a3-457c924f9e0c",
      "object_id": "wiki_en_11340",
      "similarity": 13.89739990234375
    }
  ]
}
```


Figure 2 An example response to a search request, from the Cross-Dataset Discovery API

2.3 Dataset Packaging Component

2.3.1 Component Overview and Location

Dataset packaging combines a set of given datasets, based on their metadata and inter-dataset links, to form useful packages that assist users in performing various data analysis tasks.

The process optimizes three main objectives: (a) coverage of key aspects/topics in the input set, (b) compatibility among datasets (e.g., joins/unions), and (c) redundancy reduction to avoid overlapping information.

The deployed dataset packaging service currently optimizes for coverage using a greedy search strategy to identify the optimal combinations. Its API provides the following operations: (a) retrieving package recommendations given a set of dataset IDs, (b) processing and adding a new dataset (by ID) to the internal database, (c) performing a health check to monitor service status, and (d) identifying missing datasets in the internal database relative to external data sources.

The service will be extended to incorporate all optimization objectives, leveraging more efficient search algorithms (such as a reinforcement learning approach). Next iterations of the service will include a better integration with the DataGEMS ecosystem, such as a user interface for dataset packaging recommendations, information retrieval about datasets (e.g., metadata and inter-dataset links), and event triggers to handle changes in the DataGEMS catalog when datasets are added or removed.

Table 3 Dataset Packaging Information

Dataset Packaging Component	
Service Name	dataset-packaging
Repository	https://github.com/datagems-eosc/dataset-packaging
Documentation	https://datagems-eosc.github.io/dataset-packaging-docs
Visibility (currently)	PRIVATE
Visibility (planned)	PUBLIC
LICENSE	MIT License

2.3.2 Dataset Packaging in Action

For details on how to use the component, see Deliverable D3.1 Dataset Discovery Architecture Report.

2.4 Data Analytics Patterns

2.4.1 Component Overview and Location

The Analytical Pattern (AP) Management component is designed to receive and manage APs. APs are represented using MoMa terminology and modeled as graphs composed of operators that encapsulate the functionalities of the platform's components on datasets and machine Learning models.

To support these capabilities, AP primitives are designed to generate APs for a task, store them and evaluate and compare them along three dimensions: system, data, and human. These primitives also support retrieving, explaining, executing APs, and displaying their structure and results. This set of primitives represents a preliminary step toward formalizing a full-fledged AP Management Algebra. They are intended to be extended in the future to incorporate additional functionalities and to support more complex AP management operations.

Table 4 Data Analytics Patterns Information

Analytical Pattern Management Component	
Service Name	ap-management
Repository	https://github.com/datagems-eosc/ap-management
Documentation	https://datagems-eosc.github.io/ap-management
Visibility (currently)	PUBLIC
Visibility (planned)	PUBLIC
LICENSE	MIT License

3 Conclusions

The current software deliverable provides the first release of Dataset Discovery components, In-Dataset Discovery, Cross-Dataset Discovery, Dataset Packaging, and Data Analytics Patterns. Future deliverables will contain the progress in the implementation towards the architecture and functionalities described in D3.1 Dataset Discovery Architecture Report.

Appendix I List of acronyms

Acronym	Description
API	Application Programming Interface
SQL	Structured Query Language
NLQ	Natural Language Question
AP	Analytical Pattern

Appendix II References