# DataGEMS

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




OTHER

# D2.3
# D2.3a Dataset Profiling 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** | D2.3 |
| **Title** | D2.3a Dataset Profiling Services |
| **Dissemination level** | PU |
| **Due Date** | 30/11/2025 |
| **Actual Delivery Date** | 28/11/2025 |
| **Lead Beneficiary** | ZHAW |
| **Other beneficiaries** | ARC, IDS, UVR, UPCite |
| **Editor** | Jonathan Fuerst, Yi Zhang (ZHAW) |
| **Authors** | Kurt Stockinger, Yi Zhang (ZHAW), Beatrice Amico (UNIVR), Yanlin Qi (UPCite) |
| **Contributors** | |
| **Reviewer** | Carlo Combi, Elisa Quintarelli (UNIVR) |
| **Workpackage(s)** | WP2 - Dataset Profiling |
| **Keywords** | Data profile, Multimodal, MoMa |
| **Version** | 1.0 |
| **Version information** | Number of pages: 19<br>Last file save: 28/11/2025 9:45 PM<br>File revision: 24<br>File name: D2.3_M11_D2.3a Dataset Profiling Services |


### 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 | Kurt Stockinger |
| | | | |
| | | | |
| | | | |

## Document Tables

### Table of Tables

Table 1 Dataset Profiling Component Location ..... 9
Table 2 Data Source: NOA DataGEMS Feed (Updates every 10 minutes) ..... 10
Table 3 Real-time Dataset Profiling Component Location ..... 11
Table 4 Dataset Linking Component Location ..... 15

### Table of Figures

Figure 1 Dataset Profiler Component Architecture in DataGEMS platform. ..... 8
Figure 2 Example time series anomaly report showing detected extreme weather events and sensor anomalies for a meteorological station. ..... 11
Figure 3 User interface of the Profile Similarity Report. ..... 14
Figure 4 The refinement report displays a deep-dive analysis of two selected profiles, where we obtain the keywords and assignments present in the link between these profiles. ..... 15

### 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 WP2 ..... 9
2.1 Dataset Profiling Component ..... 9
2.1.1 Component Overview and Location ..... 9
2.1.2 Dataset Profiling Component in Action ..... 9
2.1.3 Real-time Dataset Profiling Component in Action ..... 12
2.2 Dataset Linking Components ..... 13
2.2.1 Component Overview and Location ..... 13
2.2.2 Dataset Linking Component in Action ..... 16
3 Conclusions ..... 17
Appendix I List of acronyms ..... 18
Appendix II References ..... 19

## Abstract

This document describes the implementation of the dataset profiling services which extract important information from structured and unstructured data to enhance the discovery, management and exploration of dataset within the project ecosystem. The design and the architecture of the data profiling services are described in Deliverable D2.1, which should be read before this document.

# 1 Introduction

Figure 1 shows the architecture of the dataset profiling components. In this document, we describe the implementation of the following services:

- Structured dataset profiler
- Unstructured dataset profiler
- Dataset linker
- Timeseries dataset profiler

All software components are stored in GitHub along with the documentation of the respective services.

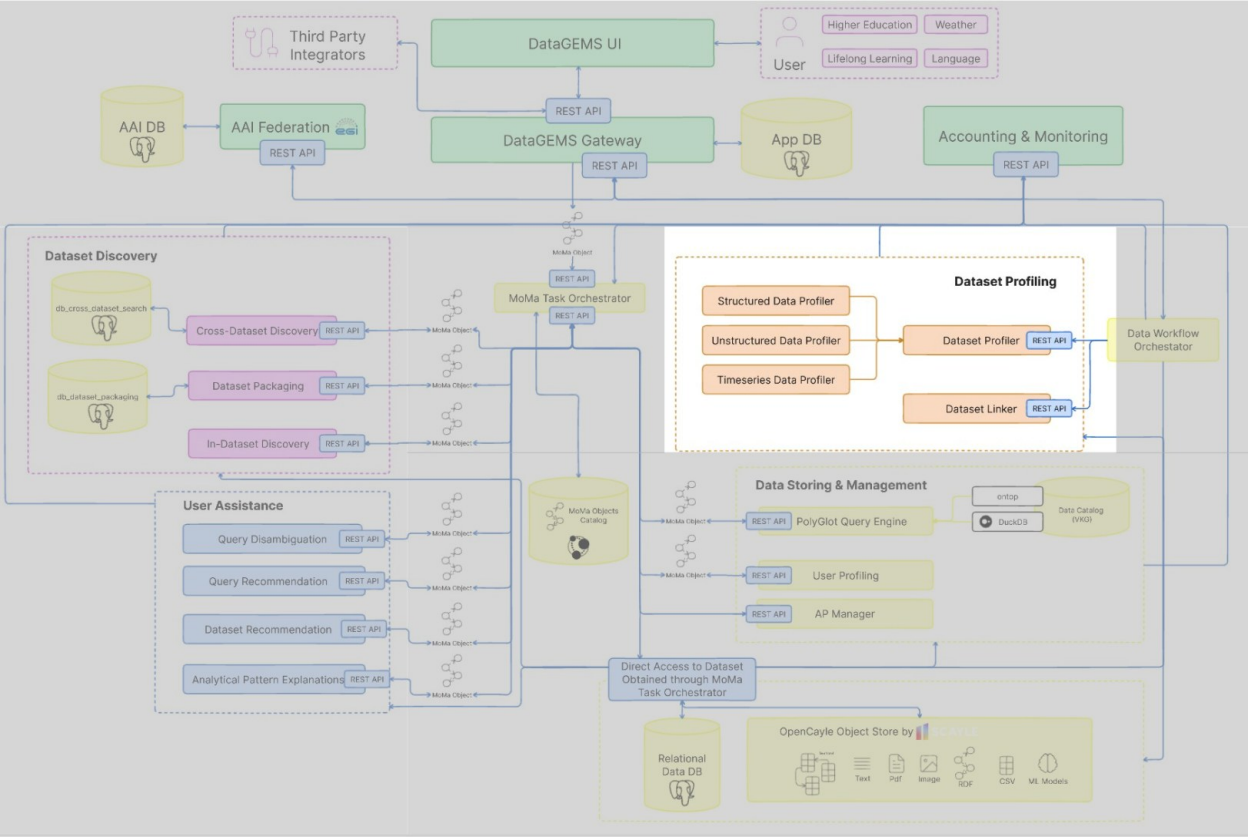



*Figure 1* Dataset Profiler Component Architecture in DataGEMS platform.

# 2 Software Components by WP2

## 2.1 Dataset Profiling Component

The **Dataset Profiler service** is a core component developed to support the project's data management objectives. Its primary function is the automated analysis and extraction of metadata from diverse dataset types. The service is designed to generate comprehensive profiles that describe the structure, content, and statistical characteristics of a dataset. This process is crucial for enhancing the **discoverability, usability, and interoperability** of data assets within the project ecosystem.

Initially, the service supports a range of common data formats, including CSV files, Excel spreadsheets, databases, text documents, and PDF files. The development roadmap includes the progressive addition of profiling capabilities for other data types and formats as the project evolves.

### 2.1.1 Component Overview and Location

The Dataset Profiler service is built with several key technical features to ensure scalability, interoperability, and ease of use:

- **Multi-format Support:** The service can ingest and process multiple data formats (e.g., CSV, Excel, databases, text, PDF).
- **Scalable Processing:** It leverages the **Ray framework for distributed processing**, enabling efficient analysis and profiling of large-scale datasets.
- **API-Driven Integration:** The service exposes a **RESTful API**, allowing for seamless integration with other project components and external services.
- **Automated Metadata Extraction:** It automatically inspects datasets to extract critical metadata, including column types, data distributions, and sample values.
- **Standardized Output:** A key feature is the generation of standardized metadata profiles in JSON-LD format. These profiles strictly adhere to the **Croissant Metadata Schema**, ensuring consistency and compatibility with other data-sharing initiatives.

*Table 1 Dataset Profiling Component Location*

| Cross-Dataset Discovery Component | |
|---|---|
| **Service Name** | dg-dataset-profiler |
| **Repository** | https://github.com/datagems-eosc/dg-dataset-profiler |
| **Documentation** | https://datagems-eosc.github.io/dg-dataset-profiler |
| **Visibility (currently)** | PUBLIC |
| **Visibility (planned)** | PUBLIC |
| **LICENSE** | MIT License |

### 2.1.2 Dataset Profiling Component in Action

For details on how to use the services, see Deliverable D2.1.

Real-time Dataset Profiling Components

The Real-time Dataset Profiling component monitors streaming time-series data – currently 10-minute meteorological sensor feeds – and uses a sliding-window spatial-temporal analysis to detect outliers and distinguish extreme weather events from potential device or sensor failures.

The anomaly detection focuses on a few key meteorological variables: **outdoor temperature (**temp_out**)** as the main indicator of weather events, **outdoor humidity (**out_hum**)** to characterize air masses, **wind speed (**wind_speed**)** which is highly variable and can reveal both gusts and stalled sensors, **barometric pressure (**bar**)** for storm and front detection, and **rainfall rate (**rain**)** to capture precipitation intensity. The details are shown in Table 2.

*Table 2 Data Source: NOA DataGEMS Feed (Updates every 10 minutes)*

| time | station_id | temp_out (°C) | out_hum (%) | wind_speed (km/h) | bar (hPa) | rain (mm) |
|---|---|---|---|---|---|---|
| 2025-11-22 17:00:00 | uth_volos | 10.1 | 85 | 12.5 | 1013.2 | 0.0 |
| 2025-11-22 17:00:00 | volos | 10.4 | 83 | 10.8 | 1013.5 | 0.0 |
| 2025-11-22 17:00:00 | zagora | 8.2 | 91 | 15.2 | 985.4 | 0.2 |
| 2025-11-22 16:50:00 | uth_volos | 10.3 | 84 | 11.9 | 1013.1 | 0.0 |
| 2025-11-22 16:50:00 | volos | 10.6 | 82 | 10.5 | 1013.4 | 0.0 |

Component Overview and Location

The service implements a two-step spatial-temporal anomaly detection workflow. First, each station is analyzed independently over a sliding temporal window (e.g., the last 6 hours). Statistical, Autoregressive and machine-learning models – including simple 3-Sigma rules, ARIMA-based time-series forecasting, and Isolation Forest – are applied to flag candidate outliers with respect to the station's own recent history. Second, for each temporal outlier, the service performs a spatial verification step by comparing the behavior of the target station with its geographical neighbors (defined by distance and elevation). This is the spatial relationships between stations **Interactive Map: View Station Neighbors**. Highly correlated behavior is interpreted as a regional environmental event (e.g., a weather front), whereas uncorrelated deviations are classified as likely device or sensor anomalies.

From an architectural point of view, the real-time dataset profiler is designed as a stateless processing engine over a persistent time-series store. New measurements are ingested from the data feed every 10 minutes and immediately written to a database backend (SQLite for local testing and PostgreSQL/Timescale DB for production deployments). The detection engine then queries a configurable sliding window (typically the last 6 hours) from the database when new data arrive, runs the temporal and spatial checks, and stores the resulting anomaly classifications back into the system. This design decouples storage from computation, avoids data loss, and allows the service to scale horizontally by adding more processing workers that share the same database.

In this report, the user provides the target stations, time window (e.g. last 6 hours) and detection method via command-line options, and the profiler queries the time-series database and returns a textual anomaly report listing, for each station and key variable, the detected anomalies, their classification (extreme weather vs device/sensor failure) and a short diagnosis table around each event.

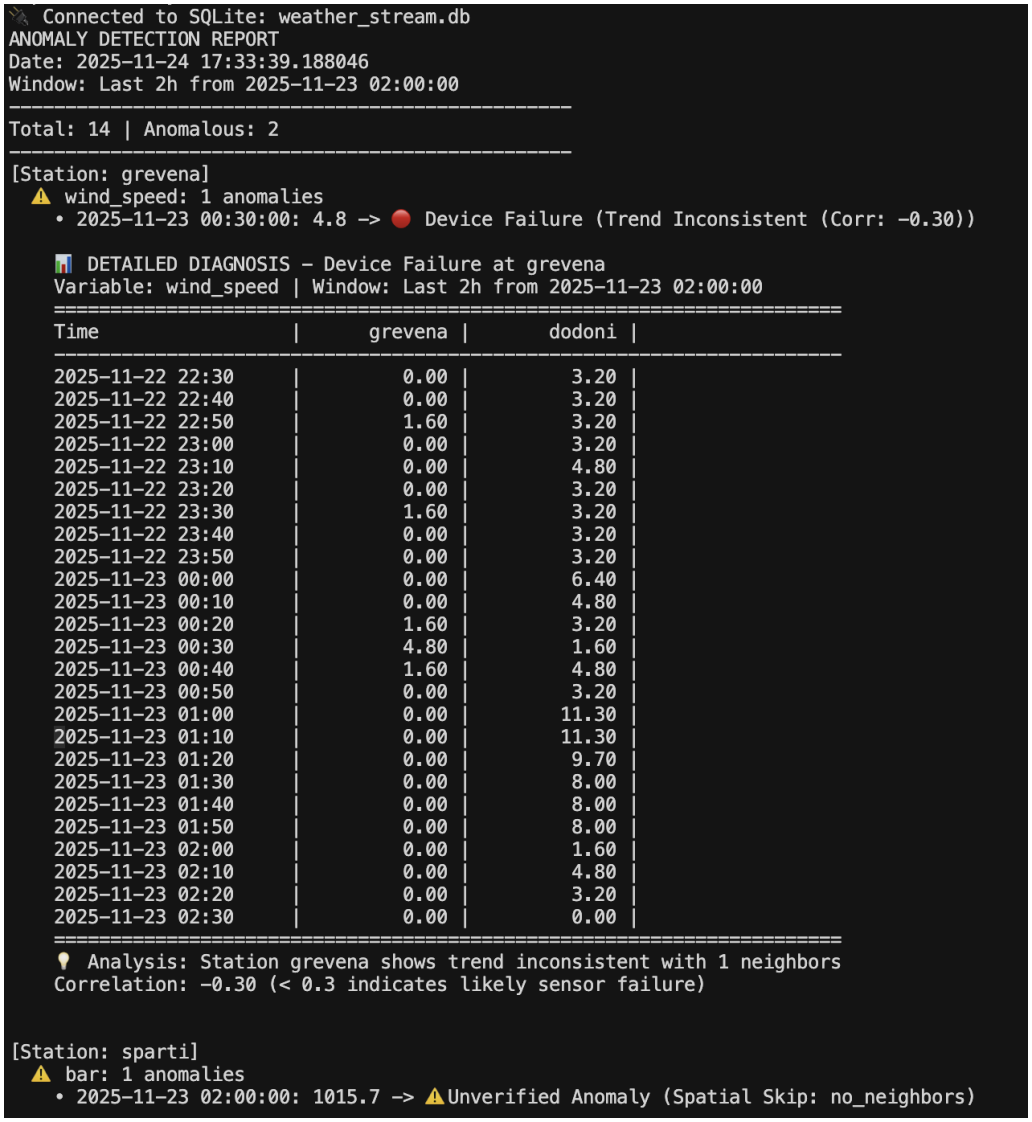



*Figure 2 Example time series anomaly report showing detected extreme weather events and sensor anomalies for a meteorological station*

*Table 3 Real-time Dataset Profiling Component Location*

| **Dataset Packaging Component** | |
|---|---|
| **Service Name** | Real-time Anomaly-detection |
| **Repository** | https://github.com/datagems-eosc/real-time-anomaly-detection |
| **Documentation** | https://datagems-eosc.github.io/real-time-anomaly-detection/ |
| **Visibility (currently)** | PUBLIC |

| | |
|---|---|
| **Visibility (planned)** | PUBLIC |
| **LICENSE** | MIT License |

### 2.1.3 Real-time Dataset Profiling Component in Action

In the current deployment, the real-time dataset profiling component is connected to a feed of meteorological observations from 14 stations, with new measurements arriving every 10 minutes. For each station, the service maintains a rolling 6-hour history window in the backing database. Whenever new data is ingested, the detection engine is triggered and retrieves the relevant window for the affected station.

Real-time Dataset Profiling Component in Action
In the current deployment, the real-time dataset profiling component is connected to a feed of meteorological observations from 14 stations, with new measurements arriving every 10 minutes. For each station, the service maintains a rolling 6-hour history window in the backing database. Whenever new data is ingested, the detection engine is triggered and retrieves the relevant window for the affected station.Step 1 – Temporal detection (time-series analysis).

For each station, the component analyses recent values to identify points that deviate from the expected short-term dynamics of the series. Several complementary detectors are applied:

- A simple 3-Sigma rule to capture sudden spikes or drops relative to the recent mean and variance.
- An ARIMA-based forecasting model to detect points that significantly diverge from the predicted value given the history.
- An Isolation Forest model that learns typical behavior from the sliding window and flags examples that are difficult to integrate into the learned pattern.

Any data point flagged by one or more of these models is treated as a temporal outlier and passed to the spatial verification stage.

In addition to short-term spatiotemporal detection, we are extending the component with a **long-term sensor health profiling** module. Instead of a few hours, this module analyses multi-day windows (e.g. 7–30 days) to detect chronic issues such as stalled wind-speed sensors, persistent missing data periods, or abnormally low variance in a variable. It computes simple statistics per station (zero-value ratio, missing-value ratio, variance, long-term correlation with neighboring stations) and flags sensors whose long-term behavior deviates from the expected operational pattern. The resulting health indicators will be integrated into the dataset profiles to complement the real-time anomaly reports.

Step 2 – Spatial verification (neighbor correlation).

For each temporal outlier, the service identifies neighboring stations based on geographical distance (e.g., within 100 km) and elevation difference. Using the same time window, it computes the correlation between the target station and each neighbor. The component then applies the following classification logic:

- If the target and its neighbors exhibit highly correlated trends around the outlier time, the system labels the event as extreme weather, i.e. a valid environmental phenomenon that consistently affects the region.
- If the target station deviates from the behavior of its neighbors (low correlation or opposite trend), the system labels it as a Device/Sensor anomaly, indicating a likely sensor malfunction or calibration issue.

For details on how to use the services, see Deliverable D2.1.

## 2.2 Dataset Linking Components

Dataset Linking allows users to identify potential connections between different datasets, expressed as dataset profiles in DataGEMS. The links between datasets are based on metadata dataset information such as keywords, descriptions, and column headers, while additional information is considered depending on each dataset's specific format.

In this initial phase, links are based on common information shared by all dataset types, such as keywords, descriptions, and headlines. The next steps will consider additional information specific for the data types and formats as the project evolves.

The development of this component started at the project month 6, adopting an agile development strategy. According to this strategy, early prototypes are built and then refined to reach stable versions of the overall component, periodically deployed in the DataGEMS platform. Deployment details will be added in the next iterations of software deliverables.

### 2.2.1 Component Overview and Location

The current prototype implementation provides a basic web application (see Figure 2) for analyzing and comparing JSON dataset profiles. It calculates the similarity between profiles based on keywords, descriptions, and column headers. On the main page, the user can select a folder containing the dataset profiles and assign weights to keywords (KW), descriptions (desc), and headers (HEAD). Such analysis allows the interested stakeholders to select suitable datasets for complex queries, possibly regarding the content of multiple datasets.

The preliminary results refer to an initial set of 12 profiles related to the DataGEMS use cases, provided by the dataset profile component, whose structures are not entirely defined. We have a first weight configuration at the top control bar, which indicates that the user has assigned a weight of 0.6 to keywords (KW), 0.3 to descriptions (DESC), and 0.1 to headers (HEAD). We also set a total similarity threshold (TH) of 30% to filter possible relevant links between two profiles. The resulting tables display a list of pairs of profiles (e.g., 19th_century_project.json vs. cedefop_skill_forecast_2025.json). For each pair, we have the individual similarity scores, respectively, for each metadata category, as well as a total similarity score.

Pairs exceeding the 30% threshold are marked with a green checkmark, and the option to produce the refinement report is available (Figure 3). Those below the thresholds are marked with a red cross.

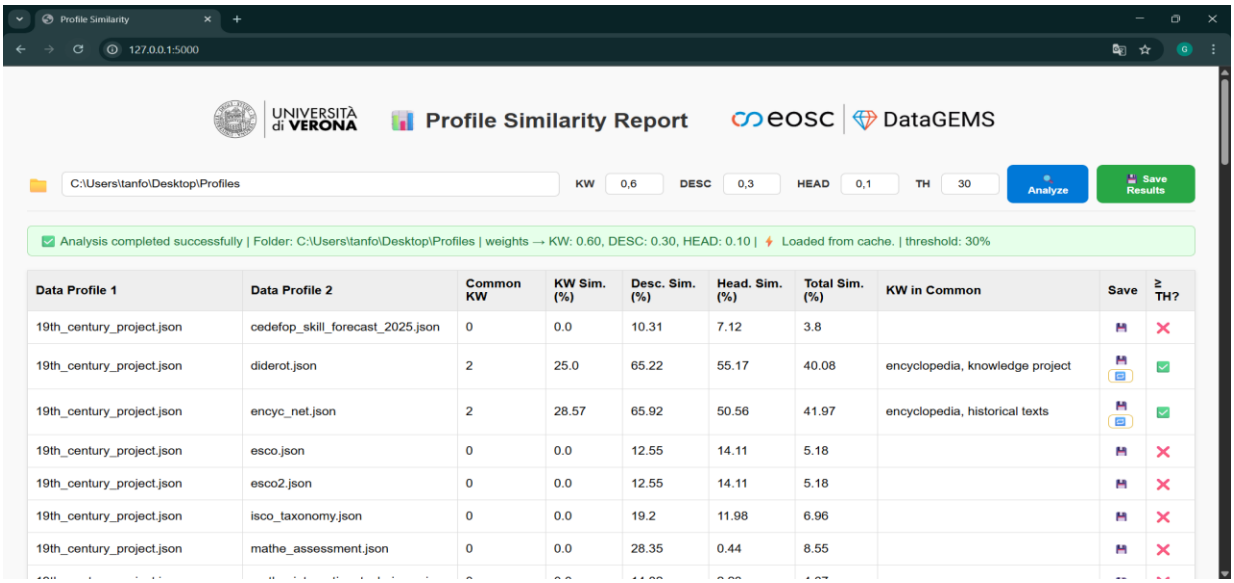



*Figure 3 User interface of the Profile Similarity Report.*

The next example is about the MathE use case. For datasets mathe_assessment and mathe_material, since their similarity exceeds the threshold, further comparison may be warranted, and additional correspondences may be verified.

This refinement report provides a detailed evaluation of the record set elements. The report details the JSON profiles *dataprofile1* and *dataprofile2* used for comparison, specifically mathe_assessment.json and mathe_material.json. It highlights the structural differences between the datasets; in this instance, *dataprofile1* is identified as having a "tabular" structure with one record set, while *dataprofile2* is "unstructured" with four record sets. Ultimately, the report explicitly lists the exercises (fields) for both profiles. This enables a granular verification of correspondence, comparing descriptive fields (such as "student id", "topic", and "keywords") against the specific data fields of the target dataset (e.g., numerical IDs or codes).




*Figure 4 The refinement report displays a deep-dive analysis of two selected profiles, where we obtain the keywords and assignments present in the link between these profiles.*

*Table 4 Dataset Linking Component Location*

| Dataset Packaging Component | |
|---|---|
| **Service Name** | Dataset-Linking |
| **Repository** | https://github.com/datagems-eosc/Dataset-Linking |
| **Documentation** | https://github.com/datagems-eosc/Dataset-Linking-docs |
| **Visibility (currently)** | PUBLIC |
| **Visibility (planned)** | PUBLIC |
| **LICENSE** | MIT License |

### 2.2.2 Dataset Linking Component in Action

For more details on how to use the services, see Deliverable D2.1.

# 3 Conclusions

The dataset profiling services have been fully implemented for structured and unstructured data with focus on CVS files, relational database tables, text documents and PDFs. The next step is to extend these services to images.

The dataset linking services are currently under development. The related prototype supports the main comparison of datasets' metadata (i.e., dataset profiles). Next steps will consider and complete i) a more detailed linking of dataset components, ii) the more complex linking of multiple dataset profiles, and iii) the integration of this service in the DataGEMS platform.

The real-time processing services provide anomaly detection for extreme weather events. As the project evolves, the real-time dataset profiling component will be generalized beyond the initial environmental use case. Planned extensions include support for additional types of time-series data (e.g., energy, traffic), more flexible configuration of detection algorithms and thresholds as part of the standard dataset profiles used in the DataGEMS platform.

# Appendix I List of acronyms

| Acronym | Description |
|---|---|
| **API** | Application Programming Interface |
| **SQL** | Structured Query Language |
| **AP** | Analytical Pattern |

# Appendix II References