Topio Marketplace: Search and Discovery of Geospatial Data

Ionescu, Andra; Alexandridou, Alexandra; Psarakis, Kyriakos; Patroumpas, Kostas; Chatzigeorgakidis, Georgios; Skoutas, Dimitrios; Athanasiou, Spiros; Hai, Rihan; Katsifodimos, Asterios

DOI
10.48786/EDBT.2023.73

Publication date
2023

Document Version
Final published version

Published in
26th International Conference on Extending Database Technology

Topio Marketplace: Search and Discovery of Geospatial Data

Andra Ionescu
a.ionescu-3@tudelft.nl
Delft University of Technology
Netherlands

Alexandra Alexandridou
Leonidas Ikonomou
alexandra@roleplay.gr
leonidas@roleplay.gr
Roleplay
Greece

Kyriakos Psarakis
k.psarakis@tudelft.nl
Delft University of Technology
Netherlands

Kostas Patroumpas
kpatro@athenarc.gr
Athena Research Center
Greece

Georgios Chatzigeorgakidis
gchatzi@athenarc.gr
Athena Research Center
Greece

Dimitrios Skoutas
dskoutas@athenarc.gr
Athena Research Center
Greece

Spiros Athanasiou
spathan@athenarc.gr
Athena Research Center
Greece

Rihan Hai
r.hai@tudelft.nl
Delft University of Technology
Netherlands

Asterios Katsifodimos
a.katsifodimos@tudelft.nl
Delft University of Technology
Netherlands

ABSTRACT

The increasing need for data trading has created a high demand for data marketplaces. These marketplaces require a set of value-added services, such as advanced search and discovery, that have been proposed in the database research community for years, but are yet to be put to practice. In this paper we propose to demonstrate the Topio Marketplace, an open-source data market platform that facilitates the search, exploration, discovery and augmentation of data assets. To support filtering, searching and discovery of data assets, we developed methods to extract and visualise a variety of metadata, as well as methods to discover related assets and mechanism to augment them. This paper aims at presenting these methods with a real deployment of the Topio marketplace, comprising hundreds of open and proprietary datasets.

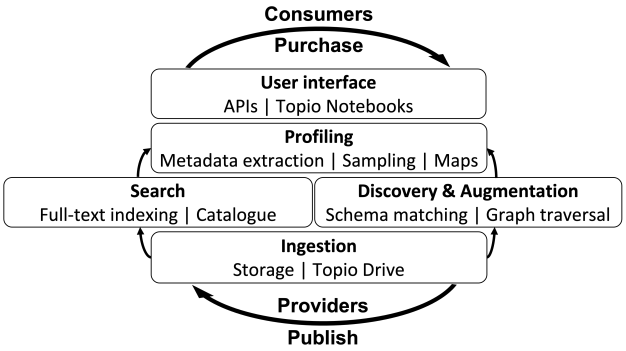


Figure 1: Topio's components for search and discovery.

1 INTRODUCTION

The growing interest in exchanging datasets and creating value from them has led to the development of data marketplaces (DMs). As such, DMs treat data as a commodity and aim at facilitating and streamlining data trading between data providers and data consumers. Data may be exchanged directly, by offering a dataset itself, or indirectly, by offering services on top of it [1]. DMs can be used to find and acquire specialized and high-quality data that are needed to train ML models, which are in turn crucial for many industrial or societal applications [8].

Many DMs have been developed over the last years with highly diverse characteristics. As a result, the landscape is quite fragmented, lacking any interoperability standards [1]. Moreover, research of DMs mostly focus on investigating pricing policies and models for data [4, 13]. However, DMs struggle with many traditional data management challenges, such as data profiling and integration, metadata curation and enrichment, dataset search and recommendation. Such problems have been studied in the context of data catalogs and data lakes [3, 9, 11, 12]. Data lakes, however, typically deal with open datasets or data exchanged among users of the same organization, whereas data in a marketplace is an asset to be traded. This makes even more imperative the need for mechanisms to facilitate buyers to quickly and easily discover relevant datasets and to be able to assess the suitability of a candidate dataset for a given task before proceeding to its purchase.

In this paper, we propose to demonstrate Topio[1], and instance of our open-source[2] marketplace platform for geospatial data, which facilitates data exploration, discovery, and augmentation. Topio facilitates the decision making process by offering a descriptive suite of metadata, mechanisms to discover related assets and to augment one or more assets from the purchased asset collection. In short, we make the following contributions:

- We offer a comprehensive and extensive suite of metadata profiling techniques which operate on multiple geodata formats.
- We facilitate the user in the exploration and discovery process through our adaptable and rich discovery service, which operates on multiple granularity levels.
- We provide the means to augment purchased assets and offer the users a step-by-step explanation on how the augmentation can be achieved.

[1] https://topio.market
[2] https://github.com/opertusmundi/

2 SEARCH & DISCOVERY COMPONENTS

Figure 1 illustrates Topio's components for searching and discovering of data assets. In short, once the assets are ingested into the platform (Section 2.1), Topio marketplace allows the user to search for assets using descriptive metadata (Section 2.2), discover related assets and ways of augmenting them (Section 2.3), explore an asset, such as visualising statistics about columns, and spatial attributes, examine samples, and finally experience and analyse an asset in Topio Notebooks (Section 2.4).

2.1 Data Asset Ingestion

The entry point of geospatial data assets is the storage, or Topio Drive. A data asset is uploaded, versioned, curated, and stored in the underlying storage, and from there delivered to consumers by first transforming it in the consumers' preferred format. The data suppliers can provide descriptive metadata about an uploaded asset (e.g. format, price, coverage, topic, etc.).

The ingestion service[3] encapsulates three main features: (*i*) reading, parsing, and extracting data-types, (*ii*) storing the asset into a PostgreSQL[4]/PostGIS[5] database, and (*iii*) registering the asset as a service (e.g. publishing a layer associated with a PostGIS datastore to GeoServer). After publication, the asset can be available as a Web Map or Feature Service (WMS/WFS) from a GeoServer instance.

2.2 Data Asset Search

Topio offers rich search capabilities with a wide range of optional filtering criteria so that prospective data consumers can quickly identify assets of their interest. All search operations are powered by indexing all assets and their metadata (provided by the supplier) and thus, supporting various search conditions (e.g., textual, numerical, spatial, temporal). Some of the filtering conditions may come from a set of pre-defined choices (e.g., asset types, file formats), while others can be user-specified (e.g., price range), enabling potential consumers to narrow down their selection to assets that mostly match their preferences based on multiple filtering criteria. The search options are illustrated in Figure 2. The platform uses tools such as PostgreSQL full text indexing as well as Elasticsearch[6].

2.2.1 Catalogue. The catalogue assumes the role of the geospatial-aware catalogue software, being responsible for managing and maintaining metadata integrity, as well as providing consistent, well-defined geospatial information to users and components. The catalogue services support the ability to publish and search collections of descriptive information (metadata) for data, services, and related information objects. The catalogue is based on a two-tier architecture. The first component is the *Metadata Store*, which comprises the RDBMS that stores and manages asset metadata. The metadata store is developed on top of the PostgreSQL with PostGIS (spatial) extension. The second component is the *Geospatial Catalogue API*, which publishes geospatial asset metadata to the other sub-systems of Topio architecture. The API is based on the OpenAPI 3.0 specification and is implemented in Python[7].

[3]https://github.com/OpertusMundi/ingest
[4]https://www.postgresql.org/
[5]https://postgis.net/
[6]https://www.elastic.co/
[7]https://github.com/OpertusMundi/catalogue-service

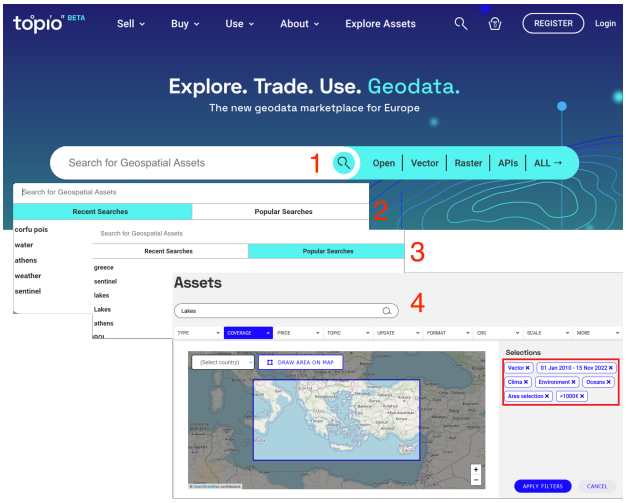


Figure 2: Multiple search criteria: 1. Keyword search, 2. Recent searches, 3. Popular searches, 4. Advanced search.

2.3 Data Asset Discovery & Augmentation

2.3.1 Discovery. Dataset discovery plays an integral part in Topio's asset discovery, enabling the platform to proactively or reactively recommend assets that are suited to their workflow context. The discovery service[8] allows end users to explore the collection of datasets by examining and understanding the relations between them and how they interconnect, which will be demonstrated together with searching in Scenario #1 (Section 3).

The core of discovery service is the *dataset relations graph* (DRG) [5, 6]. The role of DRG is to encode information from different sources in a simplified and principled manner. We define a DRG as a directed graph with nodes representing columns with properties derived from data profiles, and other automatically extracted metadata. The edges portray diverse relationships such as syntactic similarity, as well as subsumption relations and joinability conditions. The discovery service leverages Valentine [7], our open-source schema matching tool providing various methods for computing similarities. We store the DRG in Neo4J and we leverage graph traversal algorithms to create paths between the assets. The discovery service uses transitive relations and present the user multiple alternatives to link two assets.

2.3.2 Augmentation. The goal of augmentation is to recommend top-ranked assets which can be augmented to a given asset (named base asset). The approach consists of two steps. The first step is the enumeration of all the possible join paths to discover the assets that are not directly joinable with the base asset and that could connect to the base via a series of transitive joins. The second step is the ranking of join paths using a ranking function integrated with feature importance measures, in order to reduce the set of joined tables returned to the user. We use this feature to augment the purchased assets of an user and will further be demonstrated in Scenario #2 (Section 3).

2.4 Data Asset Profiling & Usage

Extracting metadata helps user in the decision making process. Metadata can be visualised (e.g. tables, charts, plots), and used for searching and filtering information, as well as data integration

[8]https://github.com/OpertusMundi/discovery-service

tasks. In data marketplaces, metadata helps the user understand an asset, determine its value and make informed purchases.

The users of Topio (providers and consumers) can access the platform through the user interface and have access to the services via the REST APIs. The profiling service is open-source[9] and served via APIs. The marketplace enables the user to invoke and use all available APIs either standalone or via Topio Notebooks, which will be demonstrated in Scenario #3 (Section 3).

2.4.1 Metadata. To compute the data profiles and metadata, we created *BigDataVoyant* [10], which repurposes and extends various existing open source software, bundled together in a streamlined and scalable manner. Data profiling for each type of supported data type (i.e., vector, tabular, raster, multidimensional) is handled by a separate software component in the profiler, and specifically: (*i*) GeoVaex[10] (an extension of Vaex [2]) developed for out-of-memory processing of vector assets, (*ii*) GDAL/OGR for raster assets, and (*iii*) the netCDF Python module for multi-dimensional assets.

2.4.2 Sampling. In Topio, the profiling information is used to describe a dataset by adding information that helps the users understand an asset better. Together with the profiling information, we show different data samples. For tabular data we use various techniques such as: random, stratified and cluster sampling. For geospatial data we created a sampling algorithm that selects random samples within a given bounded box, functionality implemented within *BigDataVoyant*.

2.4.3 Maps. Some metadata extracted from spatial attributes is represented using maps. We represent the *spatial extent* using the Minimum Bounding Rectangle (MBR), the *convex hull* of the geometries, and the *spatial data distribution* using heatmaps.

2.4.4 Topio Notebooks. Topio enables the consumers to directly use all geospatial assets purchased and uploaded, and perform operations such as data cleaning and enrichment, geocoding and trend detection, and analyzing satellite imagery in an online notebook. The notebook is backed by resources provided by Topio, which are charged to the data consumer in a separate agreement. This way, data analysis and transformation can be done without the need to download the assets, enabling the use of high-value/size and complex assets with minimal effort. While working in the notebooks environment, Topio can automatically recommend new data sources for enrichment and integration based on the data which is currently in use. This is possible with the integrated discovery service, enabling the consumer to discover (and purchase) relevant data for their data analysis workflows (Section 2.3).

3 DEMONSTRATION

Topio Marketplace[11] is currently in beta version. Topio provides open access for unregistered users, and provides a wide range of search possibilities to explore the asset catalogue. Unregistered users are able to search for assets using keywords, as well as access the most popular searches reuse them. The platform supports functionalities such as creating an account, logging in, visualising the dashboard and many more. The discovery service – the main subject of this demonstration – is still in alpha phase and it is not yet open to public exploration and usage in the live version of the platform.

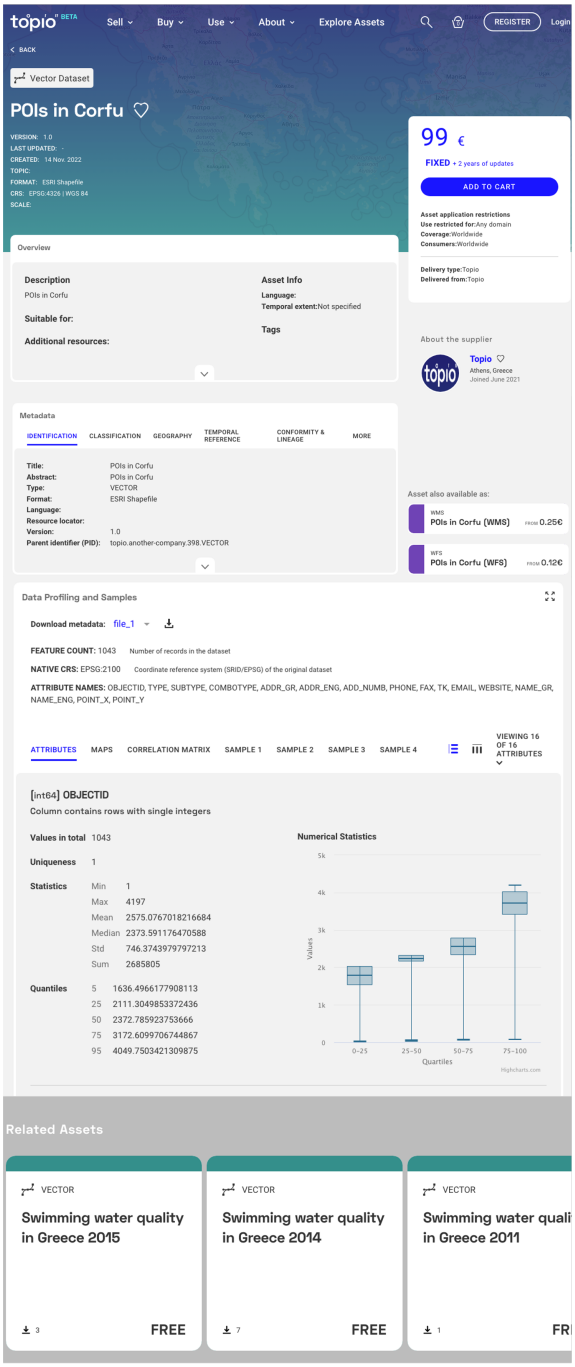
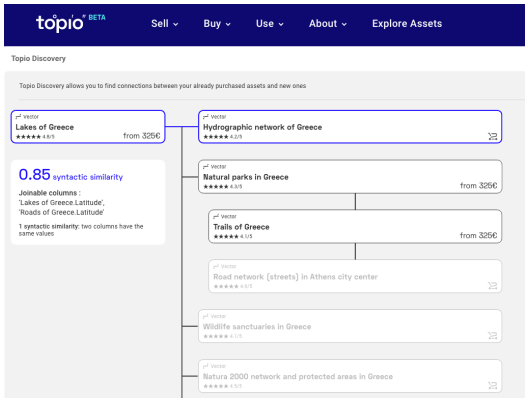


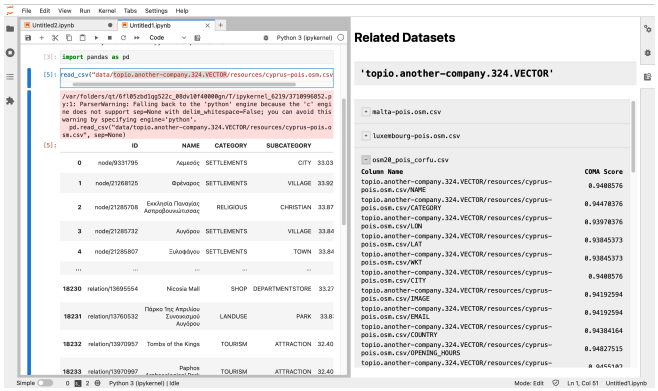
Figure 3: Asset view page: inspecting metadata, advanced profiling information and related assets.

With our demonstration, we will showcase Topio the platform in the following scenarios: (*i*) how to use Topio Marketplace to search for a wide variety of geospatial assets, and discover related and unexplored assets (Figure 3), (*ii*) how to link and augment an asset to the purchased asset collection (Figure 4a), and (*iii*) how to inspect an asset, understand its fit and purpose, and discover more related assets while working in Topio Notebook (Figure 4b).

[9] https://github.com/OpertusMundi/profile
[10] https://github.com/OpertusMundi/geovaex
[11] https://beta.topio.market/



(a) Full-screen view of augmenting purchased assets.



(b) View of the discovery service in the notebook environment.

Figure 4: Views of the augmentation service in the marketplace and discovery service in the notebooks.

Datasets. For our demonstration, we use hundreds of open data assets from the official portal for European data[12], as well as proprietary geospatial data assets.

3.1 Scenario #1: Asset search and discovery

Suppose that we want to explore the assets currently available in Topio Marketplace. As users, we can explore all the assets from the catalogue and perform an advanced search, based on the metadata provided during ingestion. With the advanced search, the users can filter the assets based on the type (e.g. vector, raster, tabular), coverage (which allows the user to draw an area of interest on a map), price range, topic (e.g. farming, health, oceans), last update date, format (e.g. CSV, KML, WMS), scale and more.

Once the users find an interesting asset during the search and exploration phase, they can inspect its own asset page as illustrated in Figure 3. Here, Topio helps the users discover even more assets related to the current one. The discovery service (Section 2.3) finds connections for every asset from the marketplace with the purpose of finding other connected and related assets. As such, on the asset page, the users have access to the related assets (if available). This process helps the user expand the search towards unknown areas.

3.2 Scenario #2: Augmenting purchased assets

Suppose that the user found an adequate asset, purchased it and then explored the asset catalogue once more. Topio supports the users in the new exploratory journey by showing how an asset from the catalogue can be connected to the purchased assets. Moreover, two presumably disjoint assets can be joined together through transitive joins. The users can visualise the join paths and how the disjoint assets can actually be connected and lastly augmented as illustrated in Figure 4a.

3.3 Scenario #3: Analyse and discover assets with Topio Notebooks

Suppose that the users found an interesting asset, but are still uncertain about the value of it. Once the users log in, Topio enables more advanced metadata computed by the profiling component described in Section 2.4. This information helps the users inspect statistics about the assets, look at different samples of the data and ultimately take an informed decision towards purchasing.

[12]https://data.europa.eu/en

To support the users even further towards assessing an asset value, Topio provides a notebook environment for analytics and discovery (Figure 4b). With Topio Notebooks, the users can assess the quality and fitness of an asset before purchasing. Inside this environment, the users benefit from the discovery service (Section 2.3) and can directly inspect the list of related assets. This functionality helps them discover more assets, without actually inspecting the asset page in the marketplace. Therefore, the users can only focus on the processing, analysing and understanding of an asset usefulness without interruptions.

ACKNOWLEDGMENTS

This work was partially funded by the the European Union's H2020 project OpertusMundi (870228).

REFERENCES

[1] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. A Survey of Data Marketplaces and Their Business Models. *SIGMOD Rec.* 51, 3 (2022), 18–29.
[2] Maarten A Breddels and Jovan Veljanoski. 2018. Vaex: big data exploration in the era of gaia. *Astronomy & Astrophysics* 618 (2018), A13.
[3] 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 (2020), 251–272.
[4] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data market platforms: trading data assets to solve data problems. *Proceedings of the VLDB Endowment* 13, 12 (2020), 1933–1947.
[5] Andra Ionescu, Rihan Hai, Marios Fragkoulis, and Asterios Katsifodimos. 2022. Join Path-Based Data Augmentation for Decision Trees. In *2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 84–88.
[6] Andra Ionescu, A Katsifodimos, and GJPM Houben. 2021. Interactive Data Discovery in Data Lakes. In *VLDB PhD Workshop*, Vol. 2971. CEUR-WS.
[7] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. 2021. Valentine: Evaluating matching techniques for dataset discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 468–479.
[8] Yifan Li, Xiaohui Yu, and Nick Koudas. 2021. Data Acquisition for Improving Machine Learning Models. *Proc. VLDB Endow.* 14, 10 (2021), 1832–1844.
[9] Renée J. Miller, Fatemeh Nargesian, Erkang Zhu, Christina Christodoulakis, Ken Q. Pu, and Periklis Andritsos. 2018. Making Open Data Transparent: Data Discovery on Open Data. *IEEE Data Eng. Bull.* 41, 2 (2018), 59–70.
[10] Pantelis Mitropoulos, Kostas Patroumpas, Dimitrios Skoutas, Thodoris Vakkas, and Spiros Athanasiou. 2021. BigDataVoyant: Automated Profiling of Large Geospatial Data.. In *EDBT/ICDT Workshops*.
[11] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data Lake Management: Challenges and Opportunities. *Proc. VLDB Endow.* 12, 12 (2019), 1986–1989.
[12] Paul Ouellette, Aidan Sciortino, Fatemeh Nargesian, Bahar Ghadiri Bashardoost, Erkang Zhu, Ken Pu, and Renée J. Miller. 2021. RONIN: Data Lake Exploration. *Proc. VLDB Endow.* 14, 12 (2021), 2863–2866.
[13] Jian Pei. 2022. A Survey on Data Pricing: From Economics to Data Science. *IEEE Trans. Knowl. Data Eng.* 34, 10 (2022), 4586–4608.