Topio: An Open-Source Web Platform for Trading Geospatial Data

Ionescu, Andra; Patroumpas, Kostas; Psarakis, Kyriakos; Chatzigeorgakidis, Georgios; Collarana, Diego; Barenscher, Kai; Skoutas, Dimitrios; Katsifodimos, Asterios ; Athanasiou, Spiros

DOI
10.1007/978-3-031-34444-2_25

Publication date
2023

Document Version
Final published version

Published in
Web Engineering

Citation (APA)
Ionescu, A., Patroumpas, K., Psarakis, K., Chatzigeorgakidis, G., Collarana, D., Barenscher, K., Skoutas, D., Katsifodimos, A., & Athanasiou, S. (2023). Topio: An Open-Source Web Platform for Trading Geospatial Data. In I. Garrigós, J. M. M. Rodríguez, & M. Wimmer (Eds.), *Web Engineering: 23rd International Conference, ICWE 2023* (pp. 336-351). (Lecture Notes in Computer Science; Vol. 13893). Springer. https://doi.org/10.1007/978-3-031-34444-2_25

Topio: An Open-Source Web Platform for Trading Geospatial Data

Andra Ionescu[1(✉)], Kostas Patroumpas[2], Kyriakos Psarakis[1], Georgios Chatzigeorgakidis[2], Diego Collarana[3], Kai Barenscher[4], Dimitrios Skoutas[2], Asterios Katsifodimos[1], and Spiros Athanasiou[2]

[1] Delft University of Technology, Delft, The Netherlands
{a.ionescu-3,k.psarakis,a.katsifodimos}@tudelft.nl
[2] Athena Research Center, Athens, Greece
{kpatro,gchatzi,dskoutas,spathan}@athenarc.gr
[3] Fraunhofer IAIS, Sankt Augustin, Germany
diego.collarana.vargas@iais.fraunhofer.de
[4] WIGeoGIS, Vienna, Austria
kb@wigeogis.com

Abstract. The increasing need for data trading across businesses nowadays has created a demand for data marketplaces. However, despite the intentions of both data providers and consumers, today's data marketplaces remain mere data catalogs. We believe that marketplaces of the future 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 not yet put to practice. With this paper, we report on the effort to engineer and develop an open-source modular data market platform to enable both entrepreneurs and researchers to setup and experiment with data marketplaces. To this end, we implemented and extended existing methods for data profiling, dataset search & discovery, and data recommendation. These methods are available as open-source libraries. In this paper we report on how those tools were assembled together to build `topio.market`, a real-world web platform for trading geospatial data, that is currently in a beta phase.

Keywords: Web platform · Data trading · Data marketplace · Open-source

1 Introduction

As the economic value of data becomes more prevalent, data marketplaces (DMs) have emerged, treating data as a commodity and aiming 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 [3]. 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

I. Garrigós et al. (Eds.): ICWE 2023, LNCS 13893, pp. 336–351, 2023.
https://doi.org/10.1007/978-3-031-34444-2_25

many industrial or societal applications [20]. They can be general-purpose, such as AWS Data Exchange[1] or Datarade,[2] or focused to a specific industry or type of data. For instance, big geospatial data providers (e.g., Carto[3], Here[4]) have recently integrated private marketplaces into their platforms. A DM is typically expected to deal with commercial data assets; nevertheless, as pointed out in [3], there also exist some DMs that generate revenue by monetizing the effort to collect and link open data, making them more easily and readily exploitable.

In this paper, we present Topio marketplace, alongside its main design decisions and the challenges that we had to overcome when developing it. Topio is designed with **openness and reusability** in mind: all of the components are packaged as reusable libraries[5] (e.g., for data discovery, data pipelines, data profiling, etc.). We believe that these reusable libraries can provide value to both researchers and practitioners alike. We also provide descriptions of the different libraries that we have developed, alongside links to their respective repositories. These libraries can be used together to form a platform on which different data marketplaces can be built.

The goal of Topio[6] is to develop a digital single market for proprietary geospatial data, addressing the heterogeneity, disparity, and fragmentation of geospatial data products in a cross-border and inclusive manner. Our goal is inspired by, and grounded on, the real-world landscape and industry-led challenges of the fragmented geospatial data value chain. The Topio marketplace is a central hub and a one-stop shop for the streamlined and trusted discovery, remuneration, sharing, trading, and use of proprietary and commercial geospatial assets [14]. Offering high-quality value-added services, it addresses the heterogeneity, disparity, and fragmentation of geospatial data products. The platform is simple, fast, cost-effective and safe for data providers and data consumers alike. In short, we make the following contributions:

- We provide insights into the needs of users, based on conducted surveys with 122 geospatial data asset providers and consumers (Sect. 3).
- We present the underpinnings of Topio - the first marketplace for geospatial data developed for publishing and purchasing assets which integrates data management tools for profiling and discovery (Sect. 4).
- We illustrate the asset lifecycle process throughout the platform and provide a pragmatic approach towards pricing (Sect. 5).
- We outline a suite of scalable, low-cost value-added services that we built on top of industrial geospatial assets published in the platform (Sect. 6).

[1] https://aws.amazon.com/data-exchange/.
[2] https://datarade.ai/.
[3] https://carto.com/spatial-data-catalog/.
[4] https://www.here.com/platform/marketplace.
[5] https://github.com/opertusmundi/.
[6] https://topio.market.

2 Related Work

Data Market Platforms. Although many DMs have emerged over the last few years, they are highly diverse with respect to their characteristics, and the landscape is quite fragmented, lacking any interoperability standards [3]. Moreover, DMs have recently become an active area of research, with many works focusing on investigating pricing policies and models for data [1,7,8,10,21]. Still, DMs deal with many traditional data management challenges, such as data profiling and integration, metadata curation and enrichment, dataset search and recommendation. Such problems have also been studied in the context of data catalogs and data lakes [6,22,25]. These, 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. Our assessment identified the lack of comprehensive and precise metadata as a significant deficiency of the current market landscape.

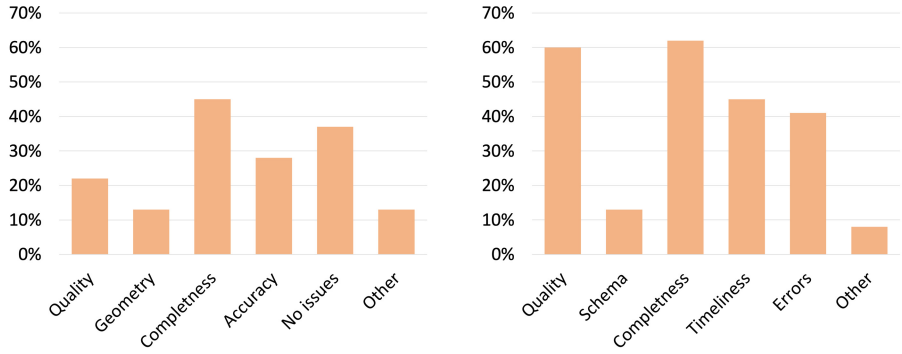
Open Data Platforms. Despite the extensive efforts of the research community towards data platforms openness, and their added benefits (e.g. developing data-driven insights and analytics modules) [24,27,28], to the best of our knowledge, there is no existing open-source platform that facilitates building and running data marketplaces. `Topio` is the first open-source set of tools that can be used to build a data marketplace. At the moment, `Topio` focuses on spatial data assets, but it can be easily extended to other data models and types.

3 User Surveys

Many data marketplaces or data sharing platforms focus on the data provider, and develop and support features tailored for the provider only [12]. Due to the difference between the viewpoints of the provider and consumer, matchmaking platforms have started to emerge [3,12]. With `Topio`, we want to develop and provide a platform which meets the requirements and preferences of both consumers and providers. Therefore, we conducted user surveys to discover and assess the qualities and features needed for a web data market platform from both perspectives: providers (27 responses) and consumers (95 responses).

3.1 Providers

The survey includes questions suitable for extracting user requirements from stakeholders with diverse backgrounds (e.g., geography, information technologies, marketing), roles (e.g., legal experts, analysts, managers, developers), and business fields (e.g., asset production, digitization, geo-marketing). The survey contains 44 questions categorized into five distinct groups: market activity, data assets, contractual life cycle, digital single market, `Topio` services.



(a) Typical issues raised by the consumers to the providers.

(b) Consumers' challenges when purchasing assets.

Fig. 1. Issues raised by consumers (a) to the providers, and (b) in the survey.

Market Activity. Most data providers currently offer less than ten geospatial data assets for sale, and typically sell two to ten geospatial data assets to the same customer. Moreover, most data providers did not adopt selling the assets via a digital marketplace, and almost half do not provide their assets as a service.

Data Assets. Most of the data providers are also the producers of the assets and do not offer their assets through a catalog or other asset management system. Most geospatial data providers do not offer access to their assets via web services. However, the providers that do, mostly prefer either OGC- (e.g., WMS, WFS), or RESTful API-based services. Finally, the providers reported that most consumers raised issues regarding the completeness of the data, and quite a few reporting complaints about the quality, the accuracy and the geometry of the assets, as illustrated in Fig. 1a.

Contractual Life Cycle. More than 60% of respondents provide their terms and restrictions as part of a contract (i.e., license embedding), while signature of a contract is needed only by 57% of the participating data owners and producers, and to a high extent, a *digital* signature is also accepted. Interestingly, a high number of providers do not need a signed contract. In terms of delivery of purchased data assets, data owners and producers usually deliver the assets through their websites, followed by email and delivery via physical media.

Digital Single Market. More than 95% of the questioned data owners and producers are interested in participating in the marketplace. However, the greatest challenges of joining a digital market platform are the standardization of pricing and contracts, and payment. To participate in a digital market platform, the providers would prefer a fixed commission on the price of each asset sale with no participation fee (42%), followed by zero fees (23%).

Topio Services. Finally, when asked about the willingness to use and adopt the services provided by a digital marketplace, more than 85% of data owners believe that the marketplace would increase their sales and revenue.

3.2 Consumers

The survey contains 25 questions categorized into three distinct groups, each one aiming to obtain insights on different aspects of geospatial asset searching and purchasing: market activity, data assets, digital single market.

Market Activity. Most geospatial data consumers mostly purchase geospatial data assets only once or once a year, and a vast majority of geospatial data consumers use *open* geospatial data assets.

Data Assets. Consumers typically use all census, place names and socio-demographic types of georeferenced data assets. Most consumers use services similar to Google Maps, and many also use OGC, RESTFul and Geospatial Analytics services. The major challenges of consumers are data availability (77%), followed by the lack of information on the quality offered (62%), and the license/contract terms (52%). The surveys also uncovered that the greatest *challenges when purchasing* data assets are their completeness (61%), quality (60%), timeliness (44%), as well as general errors (41%), also illustrated in Fig. 1b.

Digital Single Market. More than 95% of the questioned data consumers are interested in participating in the marketplace. As part of the marketplace, the consumers expect to easily find and purchase assets (85%), to have access to transparent terms and restrictions before purchasing assets (74%), high quality data (65%), transparent costs (63%) and uniform formats (50%).

3.3 Summary

Surveying both data providers and consumers, we observed the indication of a significant market interest and demand for the portfolio of services offered or envisaged by the `Topio` marketplace.

We identified the need of a digital marketplace for geospatial data assets provision, as most data owners did not embark in offering their assets via a platform. As such, with `Topio` we plan to offer multiple channels for delivering, visualising and using the data assets in support for the providers who deliver their assets via their website, or even email and physical media.

The survey also indicated that the consumers are in line with the producers in terms of assets format (SHP is the preferred format by both parties) and the usage of services such as OGC, REST APIs. Still, most geospatial data asset consumers also use services similar to Google Maps, which is expected, given the popularity of Google Maps and the bundled functionalities it provides.

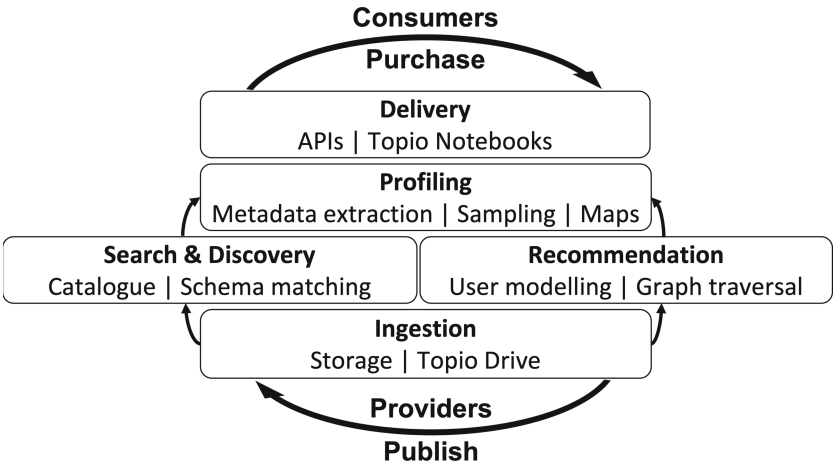


Fig. 2. Platform overview.

The major challenges indicated by the consumers perfectly frame and validate the issues addressed by Topio: make assets easier to publish and discover, and provide industry-focused and relevant metadata. Finally, these responses critically indicate that the actual quality and fit-for-use of a geospatial asset is largely an unknown entity before purchase, which deters both the use of the asset, as well as future purchases.

4 Platform Overview

The design of Topio marketplace is inspired from the insights gathered through surveys. Therefore, we focus on: (*i*) providing as much information about the assets as possible before acquisition; (*ii*) supporting multiple asset formats and delivering them via web services; and (*iii*) providing means to discover and integrate multiple assets with the aim to improve completeness, and quality. Through Topio design, we offer the absolute control of owners over their assets, and our flexible support for real-world value chain instances along the full lifecycle of geospatial data.

Figure 2 provides an overview of the components of Topio marketplace. First, the geospatial assets are ingested and stored in Topio Drive. A data asset is uploaded, versioned, curated, and stored in the underlying storage, and from there delivered to consumers directly transformed in their preferred format. The data asset lifecycle includes publishing, purchasing, delivery and also pricing based on the selected asset delivery option (Sect. 5).

We developed value-added services (VAS), including dataset discovery, recommender system, and profiler, to increase the benefits for the consumers. These benefits are twofold: (*i*) better understanding of the value of the assets based

on the metadata computed by the profiling service, and (*ii*) easier search and discovery, and personalised recommendations of related or complementary data assets (Sect. 6).

5 Data Asset Trading

In this section, we analyze existing works on data pricing (Sect. 5.1), and which of these existing ideas we have incorporated in the `Topio` platform. We then turn to the methods used to buy, sell and deliver data assets (Sect. 5.2).

5.1 Pricing Models

A lot of research has been done concerning pricing models for data [7,10,17,21]. Early works mostly focus on pricing views of data assets such that they are arbitrage- and discount-free [17]. These pricing schemes are useful for ensuring that: (*i*) a buyer will not buy "cheaper" views of a dataset whose union costs less than the original dataset, and (*ii*) the use of these concepts in practice requires both training of the data providers but also a complete pricing market architecture to support such pricing schemes.

During our research for pricing schemes, we investigated the possibility of deriving the prices from selling either subsets of the datasets, or views of those datasets, but this came to be a very challenging task. When talking to data providers during our surveys (Sect. 3), the most common request was that the providers set a price for their dataset and a separate price for each of their derivatives (e.g., a subset of the businesses in France) set by the suppliers.

At this stage, `Topio` prices datasets in two main ways: (*i*) pay per dataset; and (*ii*) pay per API call on a value-added service. The former is the simplest form of pricing: a provider offers a dataset to consumers for a fixed price and can provide discounts on bundles of datasets. For the latter, as described in Sect. 6, when consumers read data from value-added service APIs, providers can set a price per API call. API calls are logged and the consumers are charged on a per-call basis. We also offer consumers the possibility to buy API-call credits e.g., buy 1M calls for a fixed price.

5.2 Data Asset Lifecycle

Asset Provision. The provider of an asset has full and highly-granular control over the asset and can define if, when, and how an asset will be available at any point in time of the asset's lifecycle. An asset (e.g., file, database, service) is provided in a stand-alone manner, as a file with small or ad-hoc transformations, or derived/integrated with other assets. An asset is published in the platform along with its license, price policy, price and contract terms. Publishing can be limited to metadata publishing alone or the metadata and the data asset itself.

Asset Acquisition. Once an asset is uploaded in Topio Drive, the asset is immediately available throughout the application and all the services. The consumers can browse the asset catalog and discover the desired assets based on the available metadata (Sect. 6.1). The consumer retains the right to access and use the assets within the Topio platform through notebooks or maps.

Asset Delivery. Topio delivers the assets and services in three main ways: (*i*) via Jupyter notebooks after establishing an appropriate contractual agreement with the interested party (the platform or another asset owner) governing how joint value is created and shared, (*ii*) a service in one of the available APIs, and frameworks or (*iii*) integrated/derived and provided as a file. Following, we outline the asset delivery approaches.

- *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. The integrated discovery service (Sect. 6.2) enables the consumer to discover relevant data for their data analysis workflows while working in the notebooks environment. This way, we can automatically recommend new data sources for enrichment and integration based on the data currently in use.
- *Topio Maps.* Topio Maps is a comprehensive framework for creating, using, sharing, and integrating interactive maps in web and mobile applications. The consumer can create custom maps using not only the data and services provided by the platform, but also proprietary data.
- *Physical Delivery.* Finally, the purchase and delivery of the asset is performed within or outside the platform, according to owner preferences and asset type. When the files are very large or other constraints become an issue (e.g., company policies), data assets can be physically shipped to consumers.

6 Value-Added Services

Our surveys indicate that both providers and consumers face challenges coming from the assets themselves, such as quality, geometry, schema and most important: completeness and accuracy. The value-added services (VAS) provide a step forward towards facilitating asset completeness and accuracy, as they help discover new assets suitable for integration. Moreover, VAS help us to circumvent the deadlock where the consumer is unsure about the quality of the data, while the provider is not willing to reveal more information prior to payment.

As such, we have developed and integrated the following open-source value-added services: (*i*) data asset profiling (Sect. 6.1) to automatically extract various

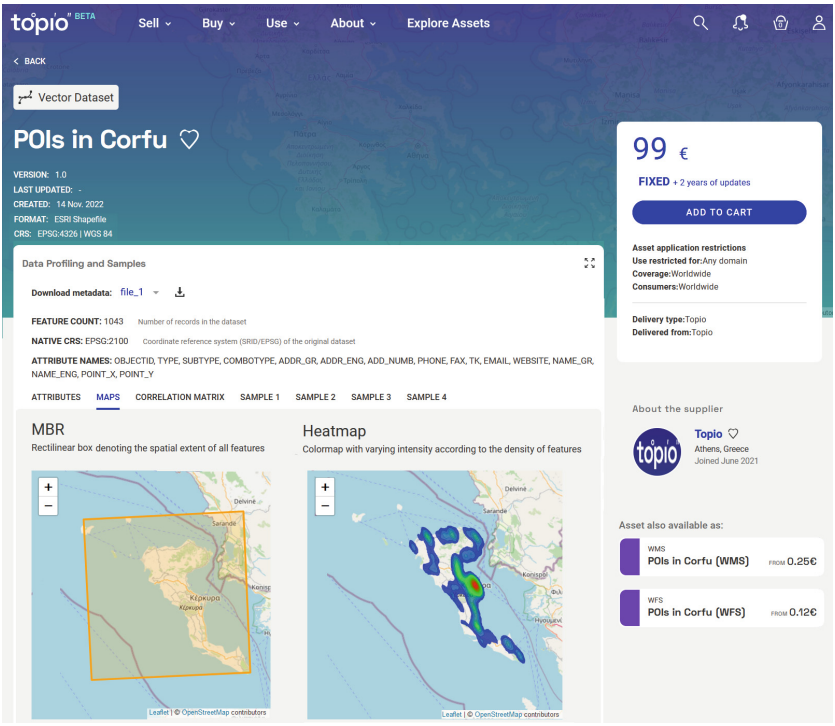


Fig. 3. View of asset details and metadata before purchase.

kinds of information from the content of a given asset and enrich its description, (*ii*) data asset search and discovery (Sect. 6.2) which offers metadata-, faceted-, keyword-based search functionalities throughout `Topio`'s catalog, and helps the user find related assets, and (*iii*) data asset recommendation (Sect. 6.3) to recommend to the consumer new assets based on already purchased, used, or visualised assets.

6.1 Data Asset Profiling

According to the user surveys (Sect. 3), providing comprehensive and precise metadata to prospective buyers for a given asset before a purchase increase transparency and trust. These observations led us towards prioritizing and strengthening the generation of automated metadata as a differentiator and unique selling point for the `Topio` marketplace.

Data profiling[7] comprises a collection of operations and processes for extracting metadata from a given dataset [26]. Such metadata may involve schema information, statistics, samples, or other informative summaries over the data, thus offering extensive and objective indicators for assessing datasets. This com-

[7] https://github.com/OpertusMundi/profile.

Table 1. Metadata computed based on the asset type.

Type	Level	Metadata
Vector & Tabular	Dataset	Feature count
	Thematic attributes	Names, data types, cardinality, distribution, N-tiles, unique values, frequency, value pattern type, special data types, keywords per column numerical value patterns, numerical statistics, correlation among numerical attributes, equi-width histogram, date/time value distribution, geometry type distribution, attribute uniqueness, compliance to well-known schema
	Geometry	Native CRS, Spatial extent, convex/concave hull, heatmap, clusters, thumbnail generation
Raster	Dataset	Native CRS; Spatial extent
	Raster specific	Resolution; Width, height; COG
	Band related	Number of bands; Band statistics; Value distribution; Pixel (bit) depth; NoData Value(s)
Multi-dimensional	Dataset	Native CRS; Dimension count/info; Variable count/info
	Variable related	Spatial extent; Temporal range; Value distribution; NoData Value(s)

ponent can be internally invoked as part of the data publishing workflow, or on demand when searching and browsing for datasets, as illustrated in Fig. 3.

The geospatial datasets can be organized in various types, commonly vector (and tabular), raster, and multi-dimensional. Although some of the profiling metadata are common among various data types (e.g., native CRS and spatial extent for spatial data), in principle a different set of metadata is used for each data type. Some of the metadata characterize the dataset as a whole (e.g., feature count for vector and tabular assets), while other metadata apply only on a specific feature of the data type. A summary of the metadata computed based on the asset type is listed in Table 1.

To compute the data profiles and metadata, we created BigDataVoyant [23], 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[8]

[8] https://github.com/OpertusMundi/geovaex.

(an extension of Vaex [5]) 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.

6.2 Data Asset Search and Discovery

Advanced 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 in the backend and thus supporting various search conditions (textual, numerical, spatial, temporal, etc.). 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 platform uses tools such as Postgres full text indexing as well as Elasticsearch.

Data Asset Discovery. Dataset discovery is the process of navigating numerous datasets in order to find relevant ones and the relationships between them [16]. The output of a discovery process represents the initial step in a data management pipeline and the input for schema matching, mapping and the subsequent processes [16]. In `Topio`, the discovery service[9] allows end users to explore the collection of datasets by examining and understanding the relations between them and how they interconnect. This process enables the users to make informed purchasing decisions, as they get more knowledgeable and understand the different layers of relatedness between the assets.

The data asset discovery process is primarily used with tabular data, such as CSV, web tables, and spreadsheets [16]. For geospatial data assets, the typical discovery process adopts methods from the Semantic Web, primarily using RDF and ontologies [4,19], while data mining and knowledge discovery approaches put more emphasis on searching for co-location patterns given location points [13]. In the context of data market platforms, where different types of datasets can be published and transformed for purchasing, we employ the methodologies existing in structured tabular data for geospatial data. As such, we reduce system complexity by utilizing the metadata extracted using the profiler component of the platform, previously described in Sect. 6.1. Such metadata is used as a filtering step to reduce the number of datasets to process for discovery. We use open-source software to transform geodata into CSV, such as mapshaper [11]. The tool addresses the challenges posed by geodata formats (e.g., Shapefiles, GeoJSON), which are non-topological data formats (i.e., do not store topological relationships between adjacent polygons). As such, the transformed files are compatible with existing open-source discovery services [15], which rely on schema matching algorithms for capturing semantic or syntactic relationships between datasets [16].

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

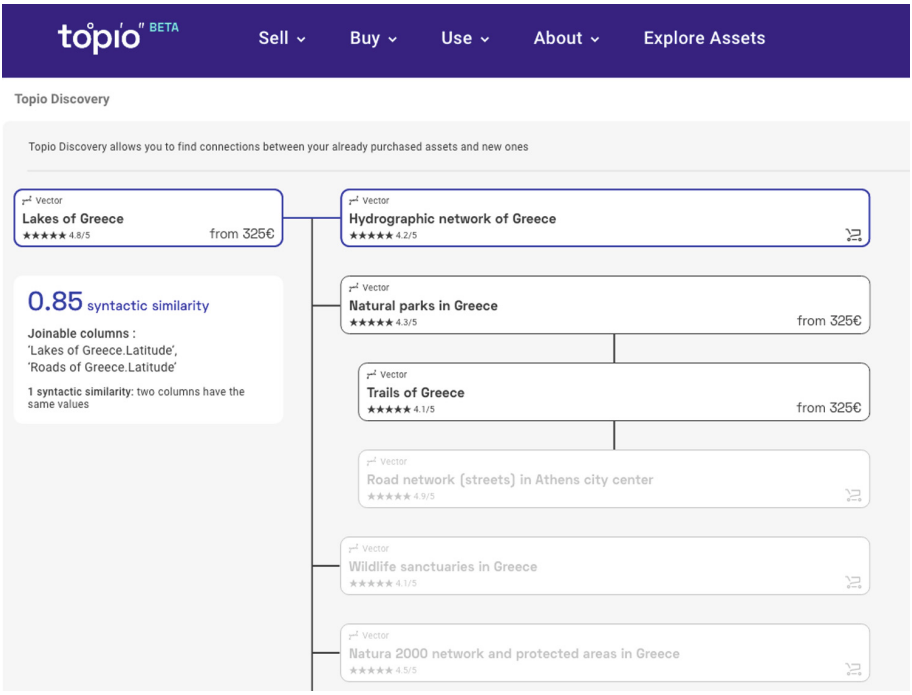


Fig. 4. View of the asset discovery and augmentation process.

Using the transformed assets, the discovery service is further used to discover assets which can be augmented. The approach leverages join paths traversal and ranking. We rank the join paths using a function integrated with feature importance measures, in order to reduce the set of joined tables returned to the user [15]. Then, using transitive joins, we determine the assets which are most appropriate for augmentation. Finally, we present all the assets used along the transitive join paths to improve the user experience through explanations. Figure 4 shows an example of the data asset discovery process used for augmentation.

6.3 Data Asset Recommendation

Topio provides contextualized asset recommendations to marketplace users, allowing the discovery of a wide range of related geospatial assets. Topio's recommender service combines several data sources from the marketplace into a consolidated knowledge graph following the DCAT[10] ontology. This knowledge graph serves as an expressive and powerful data structure that naturally models the user-item marketplace interactions. Then the recommender service applies knowledge graph embedding (KGE) models [9] to embed assets from the graph

[10] https://www.w3.org/TR/vocab-dcat-2/.

into a vector representation. Finally, the cosine function calculates the similarity among data assets in the graph.

The recommender service provides a REST API for its integration in the marketplace. The main service receives as input the asset identifier to produce the recommendations, an embedding model (currently, we support TransH [30], RotatE [29], and ComplExLiteral [18]. Many other models included in the PyKEEN framework [2] can be used), and the number of recommendations required (by default, three). With these parameters, the recommender executes its pipelines, giving a JSON response with the identifiers of the recommended assets.

When `Topio` collects more user feedback in the marketplace, such as search history, views, and buys, the recommender service will include and combine this information into the knowledge graph. More metadata will produce more robust embeddings of user interactions, making better recommendations. We plan to switch then to a collaborative filtering algorithm. The recommender service is open-source under the Apache 2.0 license[11].

7 Preliminary Usability Evaluation

We have successfully deployed the publicly available, beta version of the marketplace[12]. We used the beta version to assess the data lifecycle in the platform, measure time spent on the publishing and purchasing processes, and evaluate our design decisions. At the moment, we are evaluating the performance of each component (e.g., discovery service, recommender, profiler, etc.) using data gathered from the interactions of suppliers and consumers on our platform. However, this evaluation requires more users and specific experiments to be conducted. Until those experiments are completed, we have done preliminary investigation of how much time is required for publishing and purchasing assets. More specifically:

- Publishing an asset by an novice supplier (i.e., supplier with less than two assets published) takes on average three minutes from process start, to submission for review. We do not account for the time to upload an asset which is dependent on the size of the asset. Publishing an asset by an experienced supplier (i.e., supplier with more than five assets published) takes on average 25 s. Most suppliers opted to add optional metadata in the publishing wizard, which is a positive outcome as suppliers understand that the more metadata available, the easier for users to discover and purchase their assets.
- A supplier with an existing published asset spends, on average, five minutes to create an OGC service operationalized by `topio.market`. Most of this time is allocated to deciding the pricing of the created asset, rather than completing the wizard. This is an interesting insight as we did not observe it for data publishing; suppliers generally know well in advance the price they want to set. However, operationalizing their data represents a new market activity, and more consideration is needed to allocate the price point.

[11] https://github.com/OpertusMundi/recommender-system.
[12] https://beta.topio.market.

– The average time required for a prospective client to complete an asset purchase from visiting the cart, until asset delivery is 12 s. This is an expected duration as we based on the assumption that purchasing data assets does not differ from a standard e-shop.

8 Conclusion and Future Work

With Topio, we aim at laying the foundation for future open data marketplaces. The many components of the platform are openly available[13] and represent the starting point towards open web engineering and development. We have developed flexible and automated facilities for managing the entire lifecycle of geospatial asset trading, but these components can easily be extended to work beyond spatial data. By talking to data providers, we have come to the conclusion that commercial geodata products are updated and offered by data suppliers in regular intervals, which enables more research and development opportunities (e.g., (meta)data versioning, provenance, etc.). At the same time, providers find it very useful to use Topio to automatically offer and sell small regional data extracts/views (e.g. socio-demographics for three out of 11000 municipalities in Germany). Small regional views always require manual preparation, delivery and billing. For suppliers, there is always a lot of effort and little return, so Topio is of particular benefit to vendors in these cases. Consumers also benefit because the costs of the data extracts are reduced.

Future work for the market platform includes and is not limited to: (*i*) experimenting with different pricing algorithms and making them available for suppliers who are uncertain about pricing their assets, (*ii*) enhancing the user experience while working in Topio Notebooks by providing easy access to the data samples from the platform, (*iii*) giving the consumers the possibility to discover related assets and augmentation possibilities between existing assets from the marketplace and proprietary assets which they can upload on demand.

References

1. Agarwal, A., Dahleh, M., Sarkar, T.: A marketplace for data: an algorithmic solution. In: EC 2019, pp. 701–726 (2019)
2. Ali, M., et al.: PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings. J. Mach. Learn. Res. **22**, 82:1–82:6 (2021)
3. Azcoitia, S.A., Laoutaris, N.: A survey of data marketplaces and their business models. SIGMOD Rec. **51**(3), 18–29 (2022)
4. Batcheller, J.K., Reitsma, F.: Implementing feature level semantics for spatial data discovery: supporting the reuse of legacy data using open source components. Comput. Environ. Urban Syst. **34**(4), 333–344 (2010)
5. Breddels, M.A., Veljanoski, J.: Vaex: big data exploration in the era of Gaia. Astron. Astrophys. **618**, A13 (2018)
6. Chapman, A., et al.: Dataset search: a survey. VLDB J. **29**(1), 251–272 (2020)

[13] https://github.com/OpertusMundi.

7. Chawla, S., Deep, S., Koutrisw, P., Teng, Y.: Revenue maximization for query pricing. Proc. VLDB Endow. **13**(1), 1–14 (2019)
8. Chen, L., Koutris, P., Kumar, A.: Towards model-based pricing for machine learning in a data marketplace. In: SIGMOD, pp. 1535–1552 (2019)
9. Chu, Y., Yao, J., Zhou, C., Yang, H.: Graph Neural Networks in Modern Recommender Systems. Springer, Singapore (2022)
10. Fernandez, R.C., Subramaniam, P., Franklin, M.J.: Data market platforms: trading data assets to solve data problems. Proc. VLDB Endow. **13**(12), 1933–1947 (2020)
11. Harrower, M., Bloch, M.: Mapshaper.org: a map generalization web service. IEEE Comput. Graph. Appl. **26**(4), 22–27 (2006)
12. Hayashi, T., Ohsawa, Y.: TEEDA: an interactive platform for matching data providers and users in the data marketplace. Information **11**(4), 218 (2020)
13. Huang, Y., Shekhar, S., Xiong, H.: Discovering colocation patterns from spatial data sets: a general approach. IEEE TKDE **16**(12), 1472–1485 (2004)
14. Ionescu, A., et al.: Topio marketplace: search and discovery of geospatial data. In: EDBT (2023)
15. Ionescu, A., Hai, R., Fragkoulis, M., Katsifodimos, A.: Join path-based data augmentation for decision trees. In: IEEE ICDEW, pp. 84–88. IEEE (2022)
16. Koutras, C., et al.: Valentine: evaluating matching techniques for dataset discovery. In: IEEE ICDE, pp. 468–479. IEEE (2021)
17. Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., Suciu, D.: Query-based data pricing. J. ACM (JACM) **62**(5), 1–44 (2015)
18. Kristiadi, A., Khan, M.A., Lukovnikov, D., Lehmann, J., Fischer, A.: Incorporating literals into knowledge graph embeddings. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11778, pp. 347–363. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30793-6_20
19. Lacasta, J., Nogueras-Iso, J., Béjar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J.: A web ontology service to facilitate interoperability within a spatial data infrastructure: applicability to discovery. Data Knowl. Eng. **63**(3), 947–971 (2007)
20. Li, Y., Yu, X., Koudas, N.: Data acquisition for improving machine learning models. Proc. VLDB Endow. **14**(10), 1832–1844 (2021)
21. Liang, F., Yu, W., An, D., Yang, Q., Fu, X., Zhao, W.: A survey on big data market: pricing, trading and protection. IEEE Access **6**, 15132–15154 (2018)
22. Miller, R.J., Nargesian, F., Zhu, E., Christodoulakis, C., Pu, K.Q., Andritsos, P.: Making open data transparent: data discovery on open data. IEEE Data Eng. Bull. **41**(2), 59–70 (2018)
23. Mitropoulos, P., Patroumpas, K., Skoutas, D., Vakkas, T., Athanasiou, S.: BigDataVoyant: automated profiling of large geospatial data. In: EDBT/ICDT Workshops (2021)
24. Mucha, T., Seppala, T.: Artificial intelligence platforms–a new research agenda for digital platform economy (2020)
25. Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q., Arocena, P.C.: Data lake management: challenges and opportunities. Proc. VLDB Endow. **12**(12), 1986–1989 (2019)
26. Naumann, F.: Data profiling revisited. ACM SIGMOD Rec. **42**(4), 40–49 (2014)
27. Niculescu, M.F., Wu, D., Xu, L.: Strategic intellectual property sharing: competition on an open technology platform under network effects. Inf. Syst. Res. **29**(2), 498–519 (2018)
28. de Reuver, M., Ofe, H., Agahari, W., Abbas, A.E., Zuiderwijk, A.: The openness of data platforms: a research agenda. In: Proceedings of the 1st International Workshop on Data Economy, pp. 34–41 (2022)

29. Sun, Z., Deng, Z., Nie, J., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019 (2019)
30. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI, pp. 1112–1119. AAAI Press (2014)