SLIPO: Large-Scale Data Integration for Points of Interest

Spiros Athanasiou, Michalis Alexakis, Giorgos Giannopoulos, Nikos Karagiannakis,
Yannis Kouvaras, Pantelis Mitropoulos, Kostas Patroumpas, Dimitrios Skoutas
Information Management Systems Institute, Athena Research Center, Greece
{spathan,alexakis,giann,nkaragiannakis,jkouvar,pmitropoulos,kpatro,dskoutas}@imis.athena-innovation.gr

ABSTRACT

Points of Interest (POIs) are indispensable to many modern applications, services, and products. From navigation applications, to social networks, tourism, or logistics, we use POIs to search, communicate, decide, and plan our actions. In this demonstration, we showcase SLIPO, a system prototype that addresses the limitations, gaps and challenges in integrating, enriching, and sharing POI data. Leveraging the Linked Data paradigm to effectively extract the most out of open, crowdsourced or proprietary real-world data sources, SLIPO tackles their inherent spatial, temporal, or thematic ambiguities in POI data. Hiding all Linked Data complexities in the background, SLIPO orchestrates state-of-the-art software customized for POI data integration, enabling stakeholders to increase the value of their data and relieving them from labor-intensive, manual, error-prone, and costly updates.

1 INTRODUCTION

Points of Interest (POIs) refer to physical locations of some particular interest or utility, such as restaurants, shops, hotels, sport venues, etc. They are useful in our everyday lives (e.g., navigation, social networks, tourism) as well as in various commercial domains (such as logistics, advertising, or geomarketing). A POI is minimally characterized by its *name*, a *category*, and a *location*; however, POI profiles may often be quite complex, containing composite, multi-faceted and multi-modal information. This complexity may concern extra *thematic attributes* (address, contact details, opening hours, etc.) or their *relationship* to other entities (e.g., a shop within a mall).

Integrating POI data from multiple sources to create quality-assured, enriched, updated datasets is challenging. The advent of *open* data, *crowdsourcing*, and *social media* has provided new data sources of even greater volume, heterogeneity, diversity, veracity, and timeliness. Any consistent approach towards *POI data integration* also needs to rely on robust, flexible and semantically-rich modelling of POI profiles and handling of POI identifiers, especially when dealing with cross-sector, cross-border, and cross-lingual content. The greater the *size*, *timeliness*, *richness*, and *accuracy* of POI data, the better the end product's *value*.

Motivated by the highly fragmented landscape on POI data integration, curation and update of missing, out-of-date, or inaccurate information, we propose a pragmatic, yet highly effective approach. In the context of the SLIPO project[1], while maintaining interoperability with de facto POI standards, we opt to apply de jure *Linked Data* standards (RDF[2], OWL[3], GeoSPARQL[4]) for the inner workings of data integration, also offering capabilities to harness open POI data sources (e.g., OpenStreetMap). Linked Data technologies are ideal for handling the inherent geospatial, thematic, and semantic ambiguities of POIs. To this goal, we have built an open-source prototype system with a complete suite of software tools and services to orchestrate iterative POI *integration workflows* over multiple POI datasets, across all stages of the POI data lifecycle (transformation, linking, fusion, enrichment). Stakeholders should not adapt their current processes to collect, update, or roll-out POIs across services and products, since SLIPO hides all Linked Data complexities, and allows them to focus on their task: increase the value of their data.

The paper is structured as follows. In Section 2 we discuss the challenges in POI data integration. Section 3 overviews the data integration lifecycle as applied in SLIPO. Section 4 outlines the current status of our prototype. Finally, Section 5 showcases how a typical POI integration scenario can be handled in SLIPO.

2 ISSUES IN POI DATA INTEGRATION

POI data are by nature *semantically diverse* and *spatiotemporally evolving*, representing different entities depending on their geographical, temporal, and thematic context. Due to their use in various domains and contexts, POI-related information is typically found in diverse, heterogeneous sources. Assembling such pieces of information together is seriously hindered by the lack of common POI identifiers and data sharing formats. In addition, stakeholders also have to cope with volatile data in a POI profile, e.g., its facilities, opening hours, prices, events, etc.

Since integrating POI data with current approaches remains labor-intensive and does not scale, most stakeholders restrict their focus on domain-specific or small-sized datasets. But at a larger scale, all this complex process raises several cases of *ambiguity* that may severely hinder data integration of POIs. Addresses, coordinates, and place names are equally used throughout applications as pseudo-identifiers; but practice shows that they fail to effectively disambiguate POIs. Next, we outline these challenging issues in POI data integration.

i) Geospatial ambiguity:

- *Same POI, differing coordinates.* Locations of the same POI among different datasets almost never match exactly due to varying data collection procedures (e.g., field work, map digitization, GPS readings, crowdsourced markers).
- *Same POI, different shapes.* Although POIs are usually abstracted as point locations, they usually have a *shape* with a spatial extent (e.g., a building). But such detailed geometries are only an approximation, and their accuracy may vary significantly.
- *Different POIs, same location.* Multiple POIs may be co-located within a larger structure (e.g., multi-storey building) or a facility (e.g., shops in a mall). If abstracted as *points*, those distinct entities end up superimposed at the same location.
- *POI within another POI.* If POIs are represented by detailed *shapes* (e.g., polygons), they may exhibit topological (e.g., containment) relations. Sometimes, it is not clear whether a certain shape is a separate entity or merely a *part of* the larger one.

[1] Acronym for **S**calable **L**inking and **I**ntegration of big **PO**I data, http://slipo.eu/
[2] https://www.w3.org/RDF/
[3] https://www.w3.org/OWL/
[4] https://www.opengeospatial.org/standards/geosparql

ii) Temporal ambiguity:

- *Same POI, new location.* Location of a POI may change over time, e.g., a shop has moved to another (nearby?) place.
- *Defunct POI.* A POI may still be displayed on a map, a city guide, a navigation device, etc., but in the meantime may have stopped its operation or completely ceased to exist.
- *Same POI, change of type.* Often, the type of a POI or the operations, services and facilities it offers may change over time, e.g., a café turning to a restaurant or bar.

iii) Semantic ambiguity:

- ◇ *Same POI, different names.* POIs involving buildings, localities, etc. are often referred to by multiple names, in different contexts or time periods (e.g., "Acropolis/Parthenon", "Saint Petersburg/ Petrograd/ Leningrad"). The most typical case concerns *multi-lingual names* across datasets, possibly in different alphabets (e.g., "Acropolis" transcribed in Arabic, Cyrillic, or Chinese). Other textual characteristics can raise more concerns, especially *addresses* (e.g., renamed or renumbered streets).
- ◇ *Different POIs, same name.* It is relatively easy to disambiguate multiple locations or POIs with the same name when the spatial context is quite different (e.g., hotels with the same name in different cities). Instead, it can be quite challenging to infer what is the actual entity in the same spatial context (e.g., "Hyde Park" may refer to the park, to a nearby café, or to a hotel).
- ◇ *Same POI, different types.* Besides names, there is much heterogeneity in the use of *classification schemes*, category names and tags to semantically annotate and classify POIs. Each source typically employs its own vocabulary of categories and a hierarchy to classify POIs. Sometimes, user-defined *tags* may be assigned to POIs to describe them either instead of or in addition to predefined classification schemes.

All this makes it especially challenging and cumbersome to integrate and harmonize POI data from different sources. In SLIPO, our approach places particular emphasis to resolving ambiguity, as well as in coping with differing POI models, non-common identifiers, complex geometries, diverse attribute schemata, etc., by employing Linked Data technologies as explained next.

3 THE POI DATA INTEGRATION LIFECYCLE

In this Section, we provide an overview of the POI data integration *lifecycle*, as implemented in SLIPO. The underlying idea of our proposed system is to address the POI data integration challenges in the Linked Data domain, which is ideally suited to handle the inherent geospatial, thematic, and semantic ambiguities of POIs. Hence, POI data assets must first be transformed into RDF, so that POI profiles can be interlinked, fused, and enriched in successive steps. This is achieved through a virtuous cycle implementing iterative workflows (Figure 1) that progressively increase the size and/or the quality of the original POI data.. Next, we outline the processes and software tools involved in each step.

Transformation. In order to be handled in the Linked Data domain, POI assets from heterogeneous data sources must be transformed into RDF triples conforming to a common *OWL ontology*[5] for POI profiles. To provide a scalable and efficient transformation facility (shown as a thick red arrow in Figure 1), we extended our open-source software TripleGeo[6] to enable transformation of POI datasets from a variety of de facto geospatial formats into RDF triples with minimal overhead. Although TripleGeo is a general-purpose, spatially-aware ETL tool [3], we have included specific support for transforming POI data. This was possible through adaptable, configurable, and reusable mappings from existing attribute schemata into our POI ontology, and also supporting classification hierarchies in assigning categories to POIs. As TripleGeo inherently handles all geometry types and established coordinate reference systems, it can cope with the heterogeneity of POI formats and representations. Once data integration is complete, SLIPO introduces *reverse transformation* of the resulting integrated RDF data back to conventional POI formats (at the bottom in Figure 1), so that they can be exploited by existing products, systems, and services.

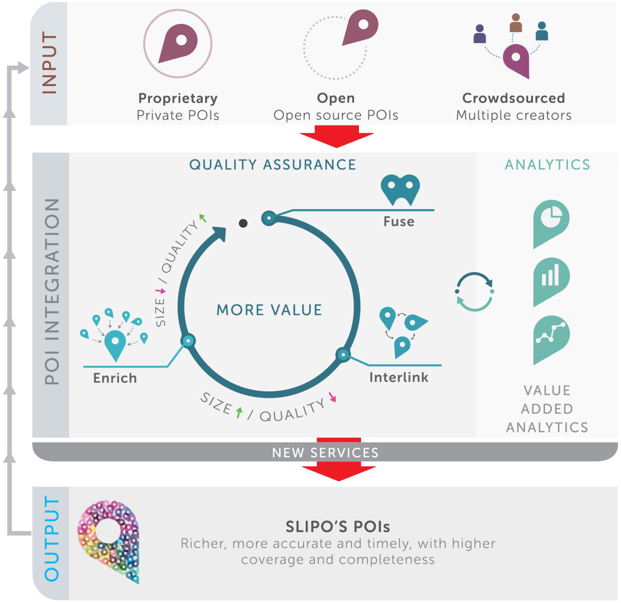


Figure 1: The POI data integration lifecycle.

All transformed RDF data are fed to a step-wise workflow abstracting a *virtuous circle*. This iterative cycle first increases the *size* (i.e., coverage, completeness, and richness) of POI data, and then refines them to increase *quality* of POIs by fusing intermediate results. For example, an expert user can repeatedly introduce additional data sources, apply different rules, etc. This iterative workflow involves the following stages:

Interlinking. This step is applied across the transformed RDF datasets coming from different sources in order to discover pairwise relations among real-world POI entities. We make use of Limes [5], a state-of-the-art interlinking software[7] that exploits the semantic structure of RDF data, textual similarities, proximity of geospatial representations, etc. In SLIPO, this actually concerns POI *deduplication*, as we wish to identify same real-world POIs based on user-specified metrics and thresholds. The output is `owl:sameAs` links between matching POI entities, which tackle the lack of common identifiers between POI entities across data sources, thus enabling their management at later stages of the integration process.

[5] Available at https://github.com/SLIPO-EU/poi-data-model

[6] Software available at https://github.com/SLIPO-EU/TripleGeo

[7] https://github.com/dice-group/LIMES

Enrichment. To produce enriched metadata and contextualize POI profiles based on information retrieved from external, third-party RDF data sources, we make use of Deer [4]. This software[8] identifies external (structured or unstructured) information related to POIs and creates extra properties. For instance, a POI profile can be enriched with opening hours, price ranges, event timelines, etc., available in SPARQL endpoints such as DBpedia[9]. It also discovers semantic interrelations between POI entities and other resources (e.g. areas, events, time), such as `partOf` relations (e.g., a shop is part of a shopping mall) or `occursAt` relations (e.g., events taking place at a certain venue).

Fusion. This stage consolidates linked POIs and their properties. From two linked POI entities it produces a unified representation, which is more complete, concise and accurate than the individual initial entities. In supporting scalable and quality-assured fusion of large POI datasets, we employ our fusion framework Fagi [1]. In SLIPO, we adapted and extended Fagi[10] with POI-specific similarity functions, learning mechanisms, and fusion actions. Such rules guide how POI properties will be fused (e.g., choose one, merge both) according to specific criteria (e.g. more complex, more timely) specified by stakeholders.

Value-Added Analytics. Having these integrated and enriched POI datasets it is then possible to provide added value services that involve *clustering* and *association discovery* among POIs. In SLIPO, we make use of Sansa [2], a software suite[11] that offers several large-scale aggregation strategies and predictive analytics, precious in geomarketing, tourism, logistics, etc.

As already mentioned, throughout this lifecycle we want to ensure that each phase produces correct and accurate results, taking into account dataset-specific and use-case-specific quality indicators and rules, including manual validation and authoring. Several indicators for such *quality assurance* can be used, most of them already adopted by industrial vendors that manage and exploit POIs: size, timeliness, coverage, accuracy, etc.

Last, but not least, we have implemented a service that allows users to track the integration and *evolution* of POI information across time and between different versions. This includes mechanisms for recording *provenance* by tracking the full history of changes per POI up to the current values of its various attributes. A graphical interface assists in visualizing and navigating through all available information, enabling users to intuitively explore where and how a POI actually changed across the various stages in the workflow.

4 THE SLIPO PROTOTYPE SYSTEM

We have been implementing a comprehensive, open-source software prototype that integrates tools for transforming, linking, fusing, enriching, and analyzing linked POI data aiming to support stakeholders in all stages of the POI data value chain. The SLIPO system[12] consists of the following main modules:

- *SLIPO Toolkit*: This is the collection of individual software components (Section 3) for transformation (TripleGeo), interlinking (Limes), fusion (Fagi), enrichment (Deer) and analytics (Sansa). Any tool can either be installed locally or invoked as part of the SLIPO workbench and APIs functionality.
- *SLIPO Workbench*: This web application allows users to orchestrate the Toolkit components and thus implement POI data integration workflows (like the one depicted in Figure 2) in a coherent, user-friendly, and flexible manner. It provides advanced capabilities for (a) uploading, searching and managing POI datasets in several formats, (b) designing, persisting and managing data integration workflows for POI datasets based on the features provided by the SLIPO Toolkit, (c) scheduling and monitoring the execution of data integration workflows, and (d) visualizing the results of such executions.
- *SLIPO APIs*: This is a collection of RESTful HTTP programming interfaces for invoking SLIPO Toolkit component functionality and integrating it into third-party systems. APIs only support the invocation of simple atomic functions (e.g., POI transformation). For composite operations, the Workbench web application must be used. Both SLIPO Workbench and APIs are exposed through the same web application server.

The SLIPO system is deployed within several virtual machines on top of the Synnefo cloud stack[13]. In the back-end, our prototype implements a workflow engine that executes data integration workflows and a scheduler for initializing workflow executions. The workflow engine and the SLIPO Toolkit components are deployed over a cloud infrastructure. Workbench and APIs exchange messages with the scheduler to execute workflows. A task is executed either in-process locally on the scheduler host, or remotely using Docker containers. Each component is deployed as a Docker Image and is responsible for providing a scalable instantiation for the requested operation (e.g., TripleGeo for transformation). A Toolkit component capable of partitioning its input and merging its output can also scale to multiple Docker containers. The scheduler only controls the total amount of resources allocated to a container, enforcing CPU/memory quotas derived from component-specific requirements and input size.

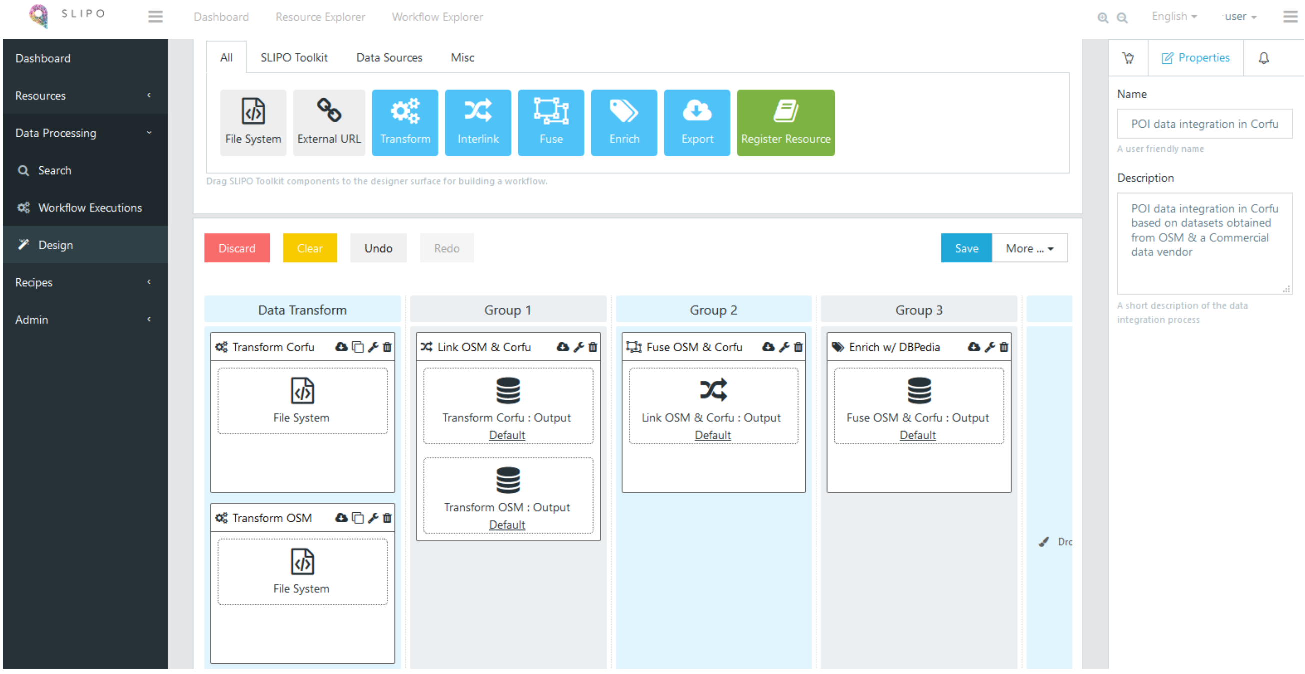
Thanks to its modular, service-oriented architecture, SLIPO offers stakeholders the option to directly use the provided functionalities following a Software-as-a-Service paradigm. Alternatively, they are able to select specific tools to customize, extend and incorporate in their own POI data management workflows using APIs according to their specific needs and requirements.

5 DEMONSTRATION

In this demonstration, we will showcase a data integration workflow using the SLIPO Workbench. This workflow accepts input POI datasets in a given geographical area (e.g., an island, a city, or a country). Data sources generally differ in schema, content, and quality; some concern *crowdsourced* information extracted from an open database (like OpenStreetMap[14] or GeoNames[15]), but others may be *proprietary* supplied by a commercial vendor.

Using the SLIPO Workbench, we will demonstrate how a user can define data integration workflows that deliver a single dataset in just a few minutes. Orchestrating the various tools into an executable workflow (like the one in Figure 2) can be carried out very quickly thanks to readily available profiles we have prepared for several common POI datasets. Such a workflow can be easily setup using drag and drop actions without the need to write any code, by only a basic parametrization per step (Section 3). This particular workflow first involves transformation of the input datasets. After discovering links between them, it fuses their

[8] https://github.com/dice-group/DEER
[9] https://wiki.dbpedia.org/
[10] https://github.com/SLIPO-EU/FAGI
[11] https://github.com/SANSA-Stack
[12] Current beta version is publicly available at https://github.com/SLIPO-EU
[13] https://www.synnefo.org/
[14] https://www.openstreetmap.org/
[15] https://www.geonames.org/



Name

POI data integration in Corfu

A user friendly name

Description

POI data integration in Corfu based on datasets obtained from OSM & a Commercial data vendor

A short description of the data integration process

Figure 2: Designing a POI data integration workflow in the SLIPO Workbench application.

properties according to user-specified rules and finally enriches the integrated result with external sources (e.g., DBpedia).

After executing such a workflow, the unified output dataset will be enhanced with information from all input datasets:

- The output dataset will contain *more POIs*. Starting with a base dataset (one of the input datasets), POIs missing from it will be complemented with information from the rest.
- Geometry representations can get a *more detailed shape*, e.g., polygons obtained from OpenStreetMap can replace (or complement) the original point (lat/lon) locations of certain POIs.
- *Extra thematic attributes* will be derived in the integrated dataset, by bringing together information (e.g., fax numbers, opening hours, links to photos, multi-lingual names) across the original data sources.
- Attribute values per POI will be *more accurate* and complete, e.g., missing telephone numbers can be filled or updated after checking against all available input.

We have prepared a short video[16] that demonstrates such a scenario on Corfu Island with a commercial POI dataset (*GET*)[17] enriched from OpenStreetMap (*OSM*). The map in Figure 3 shows how integration results (POIs depicted in blue circles or blue polygons) supersede by far and enhance the original information of the base dataset (*GET*) shown with red stars. Also, thematic properties per POI are substantially enriched with more attributes, while missing values in the base dataset are properly updated. Improvement in quality can be tracked graphically per individual POI by inspecting how it evolved along the integration progress, but also through statistics (attribute gain, confidence, etc.) estimated over the final dataset.

Overall, we believe that this demo will offer more insight not only about the challenges, but also regarding the benefits of POI data integration using SLIPO. As we continue our efforts to enhance and further develop our software, we expect rapid uptake of our innovations by stakeholders in a production setting without affecting any operations and processes already in place.

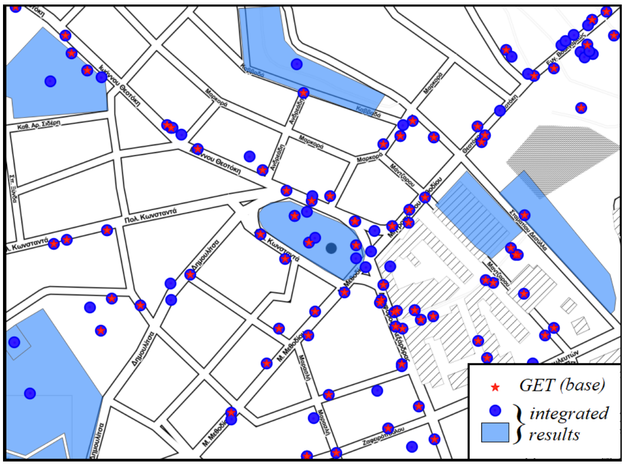


Figure 3: POIs before and after data integration in Corfu.

ACKNOWLEDGMENTS

This work was partially funded by the EU H2020 project SLIPO (#731581).

REFERENCES

[1] G. Giannopoulos, N. Vitsas, N. Karagiannakis, D. Skoutas, and S. Athanasiou. FAGI-gis: A Tool for Fusing Geospatial RDF Data. In *ESWC*, pages 51–57, 2015.

[2] J. Lehmann, G. Sejdiu, L. Bühmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A.-C. N. Ngonga, and H. Jabeen. Distributed Semantic Analytics using the SANSA Stack. In *ISWC*, 2017.

[3] K. Patroumpas, M. Alexakis, G. Giannopoulos, and S. Athanasiou. TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. In *EDBT/ICDT Workshops*, pages 275–278, 2014.

[4] M. Sherif, A.-C. Ngonga Ngomo, and J. Lehmann. Automating RDF dataset transformation and enrichment. In *ESWC*, 2015.

[5] M. A. Sherif and A.-C. N. Ngomo. A Systematic Survey of Point Set Distance Measures for Link Discovery. *Semantic Web Journal*, 2017.

[16] https://drive.google.com/file/d/1NPhl2mgbSdqH9A5KMZufQF3-7zihZ3zb/view

[17] Data sample courtesy of GET Ltd., http://www.getmap.eu/en/