DELIVERABLE D5.3

Final Evaluation Report

PROJECT NUMBER: 731581
START DATE OF PROJECT: 01/01/2017
DURATION: 36 months

Horizon 2020

SLIPO is a research project funded by European Commission's Horizon 2020 Programme.

The information in this document reflects the author's views and the European Community is not liable for any use that may be made of the information contained therein. 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.

Dissemination Level	Public
Due Date of Deliverable	Month 36, 31/12/2019
Actual Submission Date	20/12/2019
Work Package	WP5 – Pilots and Validation
Tasks	Tasks 5.2-6
Type	Report
Approval Status	Submitted for approval
Version	1.0
Number of Pages	36
Filename	D5.3_Final_Evaluation_Report.pdf

Abstract

This report provides an analysis of our pilot, during which all aspects of the SLIPO system were evaluated and validated on a real-world setting. We present key insights and observations, with emphasis on their business relevance and benefits for stakeholders of the broader geospatial value chain. *All* sensitive and business-critical information of our industrial partners (*e.g., data assets, clients, integration steps, market needs*) have been omitted. A complete and confidential version of our evaluation, including the above material, is available in the Confidential Deliverable D5.2 "Pilot Report". The reader is encouraged to contact the Consortium and request access to this report, including a hands-on demonstration of the SLIPO system.

History

version	date	reason	revised by
0.1	26/08/2019	First draft	Spiros Athanasiou
0.2	14/09/2019	Updated all sections; removed sensitive and business-critical information	Kostas Patroumpas, Dimitris Skoutas, Yannis Kouvaras, Michalis Alexakis, Pantelis Chronis, Pauline Baudens, Thodoris Vakkas, Pantelis Mitropoulos, Kai Barenscher
0.3	15/10/2019	Minor edits and revisions in all sections	Kostas Patroumpas, Dimitris Skoutas, Yannis Kouvaras, Michalis Alexakis, Pauline Baudens, Thodoris Vakkas, Pantelis Mitropoulos, Kai Barenscher
0.4	20/10/2019	Revisions in the evaluation section and executive summary	Spiros Athanasiou, Giorgos Chatzigeorgakidis
0.8	30/11/2019	First complete draft	Spiros Athanasiou
0.9	12/12/2019	Reviewed version	Georg Magenschab, Pauline Baudens, Thodoris Vakkas, Giorgos Chatzigeorgakidis
1.0	20/12/2019	Final version	Spiros Athanasiou

Author list

organization	name	contact information
ATHENA RC	Spiros Athanasiou	spathan@imis.athena-innovation.gr
ATHENA RC	Kostas Patroumpas	kpatro@imis.athena-innovation.gr
ATHENA RC	Dimitrios Skoutas	dskoutas@imis.athena-innovation.gr
ATHENA RC	Yannis Kouvaras	jkouvar@imis.athena-innovation.gr
ATHENA RC	Michalis Alexakis	alexakis@imis.athena-innovation.gr
ATHENA RC	Pantelis Chronis	pchronis@imis.athena-innovation.gr
ATHENA RC	Giorgos Chatzigeorgakidis	gchatzi@imis.athena-innovation.gr
GET	Thodoris Vakkas	tvakkas@getmap.gr
GET	Pantelis Mitropoulos	pmitropoulos@getmap.gr
TomTom	Pauline Baudens	Pauline.Baudens@tomtom.com
WIGeoGIS	Kai Barenscher	kb@wigeogis.com
WIGeoGIS	Georg Magenschab	gm@wigeogis.com

Executive Summary

This report presents an analysis of the insights generated from our 12-month pilot performed, during which the SLIPO system was evaluated and validated in a real-world setting. This report assumes the reader is familiar with the scope and services of the SLIPO system, an integrated cloud-based application for large-scale POI data integration and value-added service provision. A detailed presentation of the system, its components, and a walkthrough of its various capabilities is available in the public Deliverable D1.4 'Final SLIPO system'.

Our pilots were performed according to a formal evaluation protocol, in which five (5) data integration use cases (DINUC's) were established, addressing real-world data integration settings and a different focus for POI data integration in terms of coverage, completeness, accuracy, enrichment and efficiency. Each DINUC was independently executed several times from multiple domain-experts allowing us to objectively assess the performance and efficiency of the system over time and ensure the repeatability of our findings. In addition, our industrial partners applied the SLIPO system in everyday, ad hoc data integration jobs, evaluating it in an *open-ended* setting and in alignment with our exploitation actions. This enabled us to identify, implement, and perform several *additional new* data integration use cases of high commercial interest with a cross-sector and cross-border coverage.

Further, we designed and implemented an additional evaluation methodology to directly compare and assess the efficiency and accuracy of SLIPO vs. current POI data integration approaches. We selected three (3) DINUCs and performed data integration *manually*, measuring the effort involved. Next, we performed the data integration process via SLIPO using the optimal configurations for each use case (*i.e., configurations defined by expert users*), and *automatically* (*i.e., configurations selected automatically by the system*) and similarly measured the involved effort. We then directly compared the effort of these approaches and the integrated POIs obtained by each, allowing us to assess the *productivity* and *efficiency* gains, as well as directly compare the *accuracy* of POIs produced by the SLIPO system vs. those produced manually. These findings further confirmed the increases in productivity and cost-effectiveness offered by SLIPO, as well as the practically *identical* results achieved compared to expert user- and manual-driven POI integration.

We would like to emphasize that a discussion for the evaluation of the system is presented *both* in the current *Public* report, as well as in the preceding *Confidential* deliverable D5.2 "Pilot Report". In this report *all sensitive and business-critical information* of our industrial partners have been omitted. This approach is dictated from our piloting methodology aiming to ensure the evaluation of our work on a real-world setting, but without *compromising* in any way the business practices, relationships, or offerings of our industrial partners. The reader is encouraged to contact the Consortium and request access to the Confidential D5.2 deliverable, including a hands-on demonstration of the SLIPO system.

Based on our findings, the SLIPO Consortium has pursued an aggressive exploitation strategy, aiming to directly commercialize the output of the project and further exploit it. In summary:

- *Pre-commercialization*. We have approached the 12-month pilot of the system as an opportunity to apply and evaluate the SLIPO system in new POI data integration use cases of external commercial

stakeholders. This has afforded us the opportunity to further improve SLIPO's performance and services, but also to establish concrete collaboration ties with industrial/commercial entities outside the project Consortium.

- *Commercialization.* In alignment with our planned exploitation activities, all Consortium partners have been invited to participate in the formation of a startup company, exclusively focused on the commercialization of SLIPO's output (i.e., on-demand data integration, provision of SLIPO platform). The startup will commence its commercial activities right after the end of the project.
- *Exploitation.* Beyond the direct exploitation of SLIPO's output by our industrial (*i.e., application in data integration tasks*) and (*i.e., academic/research work*) academic partners, we will expand and integrate SLIPO in a new H2020 Innovation Action (IA), OpertusMundi, in which we will develop and operate an EU-wide industrial data marketplace for geospatial data assets, enabling geospatial data owners and producers across EU to discover, market, and purchase geospatial data and services. SLIPO will power three (3) types of value-added services: (a) data integration services for geospatial data assets (*including, but not limited to POIs*), (b) bundled geospatial analysis services based on authoritative geospatial data assets (*e.g., geomarketing, site-planning*), and (c) hosted industrial geospatial data science offerings (*i.e., Jupyter notebooks, embedded with access to proprietary data assets and SLIPO's python-based geospatial analysis API*).

Overall, we consider that SLIPO has fully achieved its objectives. SLIPO provides POI data integration output with a **similar accuracy** when compared against manual data integration, at a **fraction** of the comparative **effort and cost**, and is inherently **scalable** in terms of the size and frequency of data integration workflows at a world-scale. As such, SLIPO **empowers** industrial stakeholders with the means to **cost-effectively scale** POI data integration well beyond their current means, enabling them to harness **value** from their assets and knowhow, and expand the **reach** of their product and service offerings.

The layout of the document is as follows.

In Section 1, we frame the scope of data integration in general, and POI data integration in particular.

In Section 2, we first present a summary of the evaluated data integration use cases (DINUCs) applied to evaluate the SLIPO system, and next, the performance, scalability and efficiency of the system as measured by the corresponding KPIs.

In Section 3 we present the evaluation methodology we introduced, and the results obtained for assessing the productivity gains and accuracy achieved by the SLIPO system compared to current manual and expert-driven data integration approaches.

Finally, in Section 4 we present our key findings from the evaluation of the SLIPO system.

Abbreviations and Acronyms

AOI	Area of Interest
DINUC	Data INtegration Use Case
EV	Electronic Vehicle
GIS	Geographic Information System
KPI	Key Performance Indicator
POI	Point of Interest
QA	Quality Assurance

Table of Contents

1. POI Data Integration 10

1.1. Data integration 10

1.2. POI Data integration 11

2. DINUC Evaluation 13

2.1. DINUC A.1 13

2.2. DINUC A.2 15

2.3. DINUC A.3 17

2.4. DINUC B.1 18

2.5. DINUC C.1 20

3. Comparative Evaluation 23

3.1. Evaluation Methodology 23

3.1.1. Manual POI integration 23

3.1.2. SLIPO POI integration 25

3.1.3. Automated SLIPO POI integration 25

3.1.4. Accuracy 26

3.1.5. Productivity 29

4. Evaluation Findings 32

4.1. KPI Performance 32

4.2. Comparative Evaluation 33

4.2.1. Accuracy 34
4.2.2. Productivity 34
4.2.3. Cost-benefit 35
4.3. Summary 35

1. POI Data Integration

1.1. Data integration

In general, the purpose of data integration is to collate and combine information from *multiple* data assets to produce a *single* data asset removing *duplicate* entities (*aka deduplication, duplicate elimination, linking*), *fusing* available information for each entity (*i.e., keep, remove, decompose or combine information related to the same feature*), and optionally *enriching* with information from other sources (*i.e., add information for a given feature from a trusted source*). Data integration is a complex, open-ended, and frequent process in a business setting, since the need to combine information from independent data sources (*i.e., data producers, systems*) to produce a single data asset is a pre-requisite when developing new products and services, collaborating with new stakeholders (*e.g., a new client with its own CMS*), or even attempting to improve the efficiency of an organization (*e.g., optimize supply chains*). In this setting, data integration is challenging because:

- There is no *single nor correct* representation of real-world entities. Each data asset is produced and maintained by a different system, entity and process, with no one being an entirely accurate nor a complete representation of the world.
- Data assets are created *independently* and for *different purposes*. Each asset maintains a different identification scheme for its entities (*typically local and non-interoperable*), conventions and assumptions regarding *what* entities are, what are their *features* and how often they *change*. Further, assets may be created intentionally or as a by-product (*exhaust data*) of an activity.
- Creating the data asset is *infeasible*. The reason data integration is needed is because it is *impossible* on a practical setting (*cost-benefit, scale*) to create the required data asset from scratch.
- Information that *does not exist* cannot be created. The output of data integration cannot contain any information beyond what is *already* available in its input. Therefore, the success of data integration is not measured against the real-world, but rather on *baseline* integration processes.
- Integration is *expert-, process-, and purpose*-biased. Each data integration job, even if it applies the same input, is *unique*, with its output depending on *who* performs integration, using *what* tools and processes, as well as with *what* goal. Therefore, evaluating the success of data integration processes and tools with a set of generic indicators, *is not advisable*.

Software and services in support of data integration are positioned accordingly, providing a suite of services for the core steps of data integration (*i.e., linking, fusion, enrichment*), as well as several optional services for pre-processing, schema mapping, visualization, editing, and asset export. The definition of data integration software is *broad* given the diversity of data, sectors, and processes applied, comprising standard relational databases (*e.g., perform linking using SQL*), spreadsheets (*e.g., fuse features using VBA or MS Excel Data Tools*), scripting and data processing frameworks (*e.g., Python, R Studio, Jupyter notebooks*), visualization tools (*e.g., spreadsheet charts, web-based visualizations, interactive maps*), as well as independent or

bundled data integration-specific software (*e.g., SAS, CA, Oracle, Pentaho*). Depending on the specific organization, its workflows, input assets, integration requirements, and available resources, such tools are applied in a different manner to perform a data integration job. At all cases, their goal is to augment and facilitate the data integration process, ensuring *cost-savings* compared to what an entirely manual- and expert-driven integration would achieve. As such, their **success** is judged based on what degree they can **replicate** of the output of an expert user, for a **specific** data integration use case, as well as the amount of **effort** required.

Regardless the level of automation of a specific instance of a data integration job (*i.e., input data, tool configuration*), its **accuracy** refers to the comparison of its output vs. expert- and manual-performed data integration. This is a frequent point of miscomprehension, since the individual methods and tools applied in a data integration setting may also be characterized by their accuracy, which however refers to an *entirely different metric*. For example, a deduplication process may apply a *similarity* function with a specific accuracy guarantee, which is however measured *independently* for all supported similarity functions available, and of course is different than the accuracy of an *instance* of a data integration process. In another example, the notion of accuracy for the tools comprising a data integration process may be entirely *absent*, as is the case for simple data transformation or enrichment when entities follow the same identifier scheme.

In summary, the rationale of data integration is "*do the best with what you have*". Each data integration job, even if it refers to the same input data, is a *unique* instance, as its output is determined by its specific *goal*. The output of a data integration *instance* is evaluated on a cost-benefit approach, using an ad hoc weighted measure of its accuracy and effort compared to an expert- and manual-driven integration.

1.2. POI Data integration

In the context of SLIPO, data integration refers to *POI data integration*, i.e., producing a *single* dataset of POIs collating multiple other POI datasets or sources. The same concepts, terms, and conventions applied to generic-purpose data integration, also hold in our case. Specifically:

- Each data integration use case is *unique*. The required input assets and purpose of a data integration establish a unique data integration scenario, whose success is measured via independent and use case-specific KPIs. Besides *accuracy* and *effort*, which are relevant for all data integration jobs (*and as elaborated are unique for the specific integration use case*), a collection of other indicators can be defined, which represent the specific purpose of a given data integration use case. For example, one such indicator could be coverage (*i.e., increase the geographic coverage of the output asset*) or density (*i.e., increase the number of POIs for a given geographical extent*).
- A generic-purpose performance and effectiveness *score* for the entire system is not possible to produce, similarly to any data integration software. With each POI data integration use case applied on different data assets, under different purposes, and with its accuracy and effort dependent on the comparison with a manual and expert-driven process, it is not possible to produce generic-purpose scores that *objectively* measure the performance and effectiveness of our system. This is the reason we have defined concrete data integration use cases and provide *reproducible* KPIs for each one, objectively evaluating the merits of our work.

- Expert- and manual-driven data integration is *user-dependent*. Having an established purpose for a data integration use case (*e.g., increase coverage*), as well as a methodology for performing it in a reproducible manner, the results of expert- and manual-driven data integration still differ. Each expert may arrive to different decisions regarding the similarity of POIs and their content, reaching different decisions when combining them in a single POI. This is both understandable and extremely frequent due to the differences in the quality of the input POI assets. For example, an expert may link two POIs based on their spatial proximity, even if their name slightly differs (*e.g., 'St. Paul' vs. 'Cathedral of St. Paul'*), while another may reach a different conclusion after observing that one POI refers to the location of the church and three other POIs refer to the entrances of the church. In another example, an expert may keep the polyline geometry of a park, while another may only keep the point geometry after observing that the polyline includes a nearby business park.
- *Accuracy* for a data integration use case is *different* from the notion of accuracy when referring to the specific function or operation of a SLIPO toolkit method, and different from the term accuracy (*aka precision*) when referring to the geometries of input POI assets. As elaborated previously, the accuracy of the output for a data integration use case is compared against the output of an expert- and manual- driven process (*i.e., to what degree the system can replicate a human data integrator*). Accuracy for the geographical features in POIs refers to the (*typically unknown*) method and equipment used for establishing the location (*geometries in general*) of POIs (*e.g., digitization over base maps, ortho-interpretation, field surveys*). In contrast, accuracy for SLIPO Toolkit components refers to methods provided by each component for automatically performing certain operations (*when relevant*). For example, this includes the schema mapping facilities of TripleGeo, the similarity functions of LIMES, or the fusion rules selection of FAGI. The accuracy of each such component or function is presented in its corresponding deliverables and documentation, while it is obviously dependent on the specific experimental context (*i.e., feature space, training set, validation set*).
- POI integration is not an entirely automated process. Similarly to all data integration software, SLIPO provides a suite of tools to perform the core steps of data integration (*Toolkit components*), an interface for creating and invoking data integration workflows, and exposes the configuration parameters of individual functions/operations for each data integration tool. As such, we allow users to have full control over every aspect of data integration, iteratively select optimal configuration parameters, and optimize the output of the data integration process. This is the exact industry paradigm reflected and implemented in the context of our five (5) DINUCs. In addition, and to further simplify the use of the system, we have implemented automated operation modes for the individual Toolkit components, enabling the execution of entire data integration workflows (*or parts of them*), without the user's interaction. We do not however consider this mode of operation as a realistic scenario due to the expansive scope of POI integration in terms of the characteristics of input data and the unique purposes of each data integration use case.

2. DINUC Evaluation

We have defined five (5) data integration use cases in the context of our pilot (DINUC A.1, A.2, A.3, B.1, C.1). Each use case applied real-world data in a current industrial business case, providing a different focus for POI data integration in terms of coverage, completeness, accuracy, enrichment and efficiency.

For each DINUC, we elaborated on the application field it corresponds to, discussing: the importance of the described use case in a commercial setting; potential end products and services that are related/depend on the use case; the current state regarding software and processes that are applied (if any) and their major shortcomings. Then, we detailed the datasets to be utilized in the data integration task and elaborated the steps that comprise the data integration workflow of the DINUC, as well as the KPIs to be applied. The purpose of analyzing, framing, and presenting the DINUC's in such detail, was to ensure the *repeatability* of their execution, and thus the assembly of *objective* assessment metrics and evaluation results. Each DINUC was independently executed several times from multiple domain-experts and from all industrial partners (*>2K executions in total*), allowing us to objectively assess the performance and efficiency of the system over time and ensure the repeatability of our findings.

In the following, we present an overview of the results obtained for each DINUC, providing for each one an intentionally concise description of its scope and the values for its defined KPIs.

2.1. DINUC A.1

Sector: Geomarketing – B2B sales

This data integration use case describes a situation that is common in geomarketing: A company in the Business-to-Business (B2B) sector needs to serve its own *customers*, while constantly striving to acquire new customers (called *potentials*). It is essential for the performance of such company to be able to constantly identify and try to optimally exploit the potentials. Hence, the goal of this data integration task is to obtain market information that is as comprehensive and correct as possible about the existing customers and the possible potentials.

In this use case, SLIPO aims to improve the efficiency of the processes (*measured through time and human effort spent on the task*), but also the coverage and completeness, along with several other KPIs, as listed below:

- **Enrichment**. This KPI measures *the number of new properties* added.

  The execution of the DINUC resulted into **8 new attributes** in addition to the **original 8 attributes**. This incurs a **100%** increase on the number of available properties per POI.

- **Completeness**. The KPI measures the *average per POI number of non-empty POI property values* in the integrated dataset and is compared to the respective value in the *initial* customers dataset.

In the integrated dataset, there are **11.59** property values per POI, whereas in the initial data there are only **8.84** property values per POI. This means that after integration, on average, there is an *increase by* **31%** in the available properties per POI.

- **Efficiency**. The KPI measures *the effort required by the user* of the SLIPO Workbench, starting from data loading, including multiple cycles of re-configuration to optimize data integration results, and ending with the provision of the integration results. The result is compared against the time currently required to perform a similar workflow manually.

  The time required for an expert user of the SLIPO Workbench to perform this DINUC is **0.89 person-days**, while the estimated effort to perform data integration manually is **7.4 person-days**. As such, integration via SLIPO requires just **12%** of the effort compared against manual integration.

- **Scalability**. The KPI evaluates whether the SLIPO Workbench can support the data sizes of the described scenario. Since there is no baseline procedure to compare with, the KPI comprises a Boolean value that assesses whether the SLIPO Workbench was able to integrate the specific data in an acceptable time.

  As detailed in Section 3, the value of this indicator is 'Yes', since SLIPO successfully executed this DINUC in **3.5 hours** (Table 2).

- **User friendliness**. The KPI evaluates how complex is the usage of the system to a user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **4.1/5** and has been calculated by averaging the scores provided by the users of the SLIPO system using the system with the roles of 'Data Integrator' and 'QA User'.

- **Advanced facilities**. The KPI evaluates if the functionality of the system covers the needs of an expert user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **4.2/5** and has been calculated by averaging the scores provided by the users of the SLIPO system with the roles 'Organization Administrator' and 'Workflow Editor'.

These results have the following direct and indirect benefits on a business setting:

- *Obtain comprehensive and correct market information.* SLIPO succeeds in increasing the available properties per POI by **100%**, while also increasing the completeness of the available information by **31%**. This consists a significant increase over the original data assets, directly increasing the value of the geomarketing analysis, and thus resulting in one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over other companies offering geo-marketing services, (b) increased consumer satisfaction, (c) increase in the cost (and profits) for the provision of the service to clients, (d) open up new possibilities for geomarketing: from strategic market analysis to data-driven operational market development, and (e) increased flexibility in terms of the data sources for potential analysis.
- *Repeatable, low-cost execution of potential analysis.* SLIPO succeeds in this objective by requiring only **12%** of the effort compared against manual data integration (i.e., data loading, multiple cycles of re-configuration to optimize data integration results, provision of the output), with subsequent executions requiring only a few minutes (*i.e., negligible effort*). In addition, the potentials analysis can be performed at a much larger scale (*e.g., EU-wide*) without any increase in required effort.

These result into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over other companies offering geo-marketing services by enabling the low-cost periodic and/or ad-hoc execution of the analysis with negligible effort (*the provision of frequent updates based on pre-defined recipes especially, is very important from end-customer point of view*), (b) competitive advantage over other companies offering geo-marketing services by supporting the provision of analyses at a much larger geographical scale, and with negligible impact on the required effort, (c) increased consumer satisfaction by enabling a more timely and frequent execution of the analyses with minimal extra cost/effort and by independence of data-sources (e.g., *different provider of business-addresses*), (d) increase in the cost (and profits) for the provision of the service to clients.

2.2. DINUC A.2

Sector: Geomarketing – Location planning

This data integration use case concerns a location planning scenario. A company is examining their entry into an Eastern European country. Among other data (catchment areas, demographic data of households, population, etc.) necessary as input to a customized model for evaluating candidate locations to open new pharmacies, this model also requires POI data. In particular, for each catchment area, such a model requires:

i. **Competitors**: POIs belonging to the respective area, and specifically, companies in a specific subset of market categories (*depends on the particular market area/focus of the client, e.g., restaurants*).

ii. **Magnets**: POIs belonging to the respective area, and specifically, companies in a specific subset of market categories (*depends on the particular market area/focus of the client, typically clothing, cafes, restaurants*).

This data integration use case is currently performed in an exclusively manual way. In this use case, SLIPO aims to improve the efficiency of the processes (measured mainly through time and human effort spent on the task), but also the coverage and completeness, as listed in these KPIs:

- **Enrichment**. The KPI measures *the number of new properties* added.

  The execution of the DINUC resulted into **18** extra attributes included in the integrated dataset. This incurs a **138%** increase on the number of available properties per POI.

- **Size**. The KPI measures *the number of POIs* that are included in the final, integrated POI dataset.

  The integrated dataset contains **17,414** POIs against only 5,899 POIs in the original dataset for Bucharest. This indicates an *increase by* **195%** in the number of POIs, i.e., that the dataset is practically tripled in size.

- **Completeness**. The KPI measures *the average per POI number of non-empty POI property values* in the final dataset.

  In the integrated dataset, there are **10.32** property values per POI, whereas in the initial data there are only **9.73** property values per POI. This means that after integration, on average, there is an *increase by* **6%** in the available properties per POI.

- **Efficiency**. The KPI measures *the effort required by the user* of the SLIPO Workbench, starting from data loading, including multiple cycles of re-configuration to optimize data integration results, and ending with the provision of the integration results. The result is compared against the time currently required to perform a similar workflow manually.

  The time required for an expert user of the SLIPO Workbench to perform this DINUC is **1.33 person-days**, while the estimated effort to perform the data integration manually is **19.01 person-days**. As such, integration via SLIPO requires **6.99%** of the effort compared against manual integration.
- **Scalability**. The KPI evaluates whether the SLIPO Workbench can support the data sizes of the described scenario. Since there is no baseline procedure to compare with, the KPI comprises a Boolean value that assesses whether the SLIPO Workbench was able to integrate the specific data in an acceptable time.

  The value of this indicator is 'Yes', since SLIPO successfully executed this DINUC in **3 minutes**.
- **User friendliness**. The KPI evaluates how complex is the usage of the system to a user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **4.2/5** and has been calculated by averaging the scores provided by the users of the SLIPO system using the system with the roles of 'Data Integrator' and 'QA User'.

These results have the following direct and indirect benefits on a business setting:

- *Increase coverage and completeness*. SLIPO succeeds in this objective, by delivering a dataset with increased size, richness, *and* completeness. Specifically, the produced dataset has simultaneously greater number of POIs by **195%** (*i.e., more competitors/magnets to apply in the analysis*), extra **138%** attributes (*i.e., more information to select/filter/assess competitors/magnets*) and **6%** more complete properties per POI. This consists a significant improvement compared to the original dataset, resulting into one or more of the following (*depending on the specific client and business case*): (a) increased value for the site-planning analysis output (*business decisions can be made safer because of the better data basis*), (b) competitive advantage over other companies offering similar services, (c) increased consumer satisfaction, (d) increase in the cost (and profits) for the provision of the service to clients.
- *Improve efficiency*. SLIPO succeeds in this objective by requiring only **6.99%** of the effort compared against manual integration (i.e., data loading, multiple cycles of re-configuration to optimize data integration results, provision of the output), with subsequent executions requiring only a few minutes (*i.e., negligible effort*). In addition, the location planning analysis can be performed at a much larger scale (*e.g., EU-wide*) without any increase in required effort. These result into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over other companies by offering similar services by enabling the low-cost periodic and/or ad-hoc execution with negligible effort, (b) competitive advantage over other companies offering location planning services by supporting the provision of analyses at a much larger geographical scale, and with negligible impact on the required effort, (c) increased consumer satisfaction by enabling a more timely and frequent execution of the analyses with minimal extra cost/effort, (d) increase in the cost (and profits) for the provision of the service to clients.

2.3. DINUC A.3

Sector: Geomarketing – Dealer network analysis

This data integration use case refers to a specific type of geomarketing analysis that helps to optimize the dealer network of manufacturing companies. Dealer networks, or specialist traders, may offer products of several manufactures (e.g., cars), but each manufacturer may have its own network of dealers. To assess the regional market coverage, companies visualize the locations of their network of dealers (i.e., POIs) as well as their competitors. However, the same dealer POI may be included in more than one dealer datasets, each dataset belonging to competing manufacturers. Information regarding the same dealer may differ among such datasets in terms of names, location, address, contact details, etc. In this use case, we consider seven (7) datasets with dealer POIs, each one corresponding to the dealer network of a different manufacturer. Our goal is to integrate those seven individual datasets into one containing all dealer POIs, and also listing the manufacturers offering their products to each dealer. Currently, such integration of dealer POIs is done *manually* by map-based visual exploration and considers spatial proximities of POI locations to find possible matches. In SLIPO, we make use of transformation, interlinking, and fusion to efficiently yield a unified, quality-assured dataset.

This data integration use case is currently performed in an exclusively manual way. Furthermore, this integration processes potentially needs to run frequently and/or in short notices. In this use case, SLIPO improves the efficiency of the processes (measured mainly through time spent on the task), but also the coverage and completeness, as listed in these KPIs:

- **Size**. This KPI measures *the number of POIs* that are included in the final, integrated POI dataset, against the number of POIs in the initial dataset.

  The final dataset contains **3,927** POIs, while the original dataset has **1,077** POIs. Since there is an *increase by* **365%** in its contents, the final integrated dataset practically almost quadruples the amount of available POIs.
- **Deduplication effect**. This KPI is measured by assessing *the number of same dealer POIs that are identified* in each consecutive interlinking iteration. In this setting, this expresses the amount of *duplicate POIs* detected along the integration process, and hence is equivalent to the *redundancy reduction* that has been achieved in the final, integrated dataset.

  In total, **663** POIs are identified as duplicates in the six integration cycles. Compared to the total **4,590** POIs that could be obtained from the union of the original datasets (without deduplication), this value signifies that **15%** of the original POIs are duplicates and have been eliminated from the final, integrated dataset.
- **Efficiency**. The KPI measures *the effort required by the user* of the SLIPO Workbench, starting from data loading, including multiple cycles of re-configuration to optimize data integration results, and ending with the provision of the integration results. The result is compared against the time currently required to perform a similar workflow manually.

The time required for an expert user of the SLIPO Workbench to perform this DINUC is **1.48 person-days**, while the effort to perform the data integration manually is **7.40 person-days**. As such, integration via SLIPO requires **20.2%** of the effort compared against manual integration.

- **Scalability**. The KPI evaluates whether the SLIPO Workbench can support the data sizes of the described scenario. Since there is no baseline procedure to compare with, the KPI comprises a Boolean value that assesses whether the SLIPO Workbench was able to integrate the specific data in an acceptable time.

  The value of this indicator is 'Yes', since SLIPO successfully executed this DINUC in **2 minutes.**

- **User friendliness**. The KPI evaluates how complex is the usage of the system to a user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **3.4/5** and has been calculated by averaging the scores provided by the users of the SLIPO system using the system with the roles of 'Data Integrator' and 'QA User'.

These results have the following direct and indirect benefits on a business setting:

- *Increase coverage and completeness*. SLIPO succeeds in this objective, by delivering a dataset with increased size. Specifically, the produced dataset has greater number of POIs by **365%** compared t to the original dataset (*i.e., more dealer network POIs*), while also identifying 15% of the POIs in the as duplicates. This consists a significant improvement compared to the original dataset, resulting into one or more of the following (*depending on the specific client and business case*): (a) increased value for the dealer-network analysis output, (b) competitive advantage over other companies offering similar services, (c) increased consumer satisfaction, (d) increase in the cost (and profits) for the provision of the service to clients.
- *Increase efficiency; frequent and timely execution*. SLIPO succeeds in this objective by requiring only **20.2%** of the effort compared against manual integration (i.e., data loading, multiple cycles of re-configuration to optimize data integration results, provision of the output), with subsequent executions requiring only a few minutes (*i.e., negligible effort*). In addition, the location planning analysis can be performed at a much larger scale (*e.g., EU-wide*) without any increase in required effort, as well as more frequently and at a short notice, with minimal effort. These result into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over other companies offering similar services by enabling the low-cost periodic and/or ad-hoc execution with negligible effort, (b) competitive advantage over other companies offering dealer network analysis services by supporting the provision of analyses at a larger geographical scale, and with negligible impact on the required effort, (c) increased consumer satisfaction by enabling a more timely and frequent execution of the analyses with minimal extra cost/effort, (d) increase in the cost (and profits) for the provision of the service to clients.

2.4. DINUC B.1

Sector: Tourism and Leisure – Food tourism

This use case concerns data integration of POIs related to *food (gastronomic) tourism*, which is considered as a vital component of the tourism experience worldwide, with over a third of tourist spending devoted to food. In this setting, we consider an initial, proprietary POI dataset which contains general-purpose POIs, i.e., not only food-related POIs, but also public buildings, bus stops, etc., updated on an *ad-hoc basis, for specific areas/on client demand/on error detection*, leading to a *non-homogeneous product*. Our primary goal is to improve it in terms of size, coverage, richness, and completeness, with information extracted from third-party data sources. Further, we want to enrich this data with additional properties specifically regarding food-related POIs (e.g., producers, gastronomic festivals, food fairs, events, farmers' markets) from third-party data sources.

This data integration use case describes a real-world setting where the integration process is currently performed in an exclusively manual way. This results to a throughput of approximately 100 POIs/per man-day, where validation may also regard POI addition, updating, interlinking, fusion or enrichment. Given that, in this use case, SLIPO aims to improve the efficiency of the processes, as well as the size and the coverage of the produced POI datasets, along with several other KPIs, as listed below:

- **Efficiency**. The KPI measures *the effort required by the user* of the SLIPO Workbench, starting from data loading, including multiple cycles of re-configuration to optimize data integration results, and ending with the provision of the integration results. The result is compared against the time currently required to perform a similar workflow manually.

  The time required for an expert user of the SLIPO Workbench to perform this DINUC is **1.04 person-days**, while the estimated effort to perform the data integration manually is **178.62 person-days**. As such, integration via SLIPO requires just **0.58%** of the effort compared against manual integration.

- **Enrichment**. The KPI measures *the number of new properties* that were added to the GET dataset.

  The execution of the DINUC resulted into **13** extra attributes included in the integrated dataset, which indicates an increase of **81%** with respect to the original **16** attributes.

- **Coverage**. This KPI measures the number of *food-related POIs* that were added to the final integrated dataset compared to those that existed in the original dataset.

  The original data contains **15,649** (restaurants, taverns, fast-food, but also café and bars). In the integrated data there are **13,713** extra *food-related POIs*. Hence the final dataset contains **29,362** such POIs in total, which indicates an *increase by* **88%** compared to the original dataset.

- **Size**. The KPI measures *the number of POIs* that are included in the final, integrated POI dataset, against the number of POIs in the initial dataset.

  The integrated dataset contains **160,306** POIs against only **72,373** POIs in the original dataset. This indicates an *increase by* **121%** in the number of POIs, i.e., practically more than doubled in size.

- **Completeness**. The KPI measures *the average per POI number of non-empty POI property values* (e.g., for address, website, phone properties) in the final dataset. This number is compared to the respective value in the initial dataset.

  In the integrated dataset, there are **8.2** *property values per POI*, whereas in the initial data there are **8.39** *property values per POI*. This means that after integration, on average, there is a slight *decrease by* **2%** in the available properties per POI. This result may seem rather poor, but it should be noted

that the third-party dataset applied is not particularly rich in properties, as most POIs include only basic tags (like name, categorization, and sometimes a contact).

- **Scalability**. The KPI evaluates whether the SLIPO Workbench can support the data sizes of the described scenario. Since there is no baseline procedure to compare with, the KPI comprises a Boolean value that assesses whether the SLIPO Workbench was able to integrate the specific data in an acceptable time.

  As detailed in Section 3, the value of this indicator is 'Yes', since SLIPO successfully executed this DINUC in **22 minutes**.

- **User friendliness**. The KPI evaluates how complex is the usage of the system to a user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **4.2/5** and has been calculated by averaging the scores provided by the users of the SLIPO system using the system with the roles of 'Data Integrator' and 'QA User'.

These results have the following direct and indirect benefits on a business setting:

- *Increase efficiency*. SLIPO succeeds in this objective by requiring only **0.58%** of the effort compared against manual integration (i.e., data loading, multiple cycles of re-configuration to optimize data integration results, provision of the output), with subsequent executions requiring only a few minutes (*i.e., negligible effort*). In addition, integration can be performed at a much larger scale (*e.g., EU-wide*) without any increase in required effort. These result into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over companies offering similar thematic POI data assets by offering a dataset at a fraction of the effort, (b) potential expansion of the asset's geographical coverage (*e.g., country/EU-wide*) at a fraction of the effort, (c) increased consumer satisfaction by enabling a more timely data asset, (d) increase in the cost (and profits) for the provision of the data asset to clients.
- *Increase coverage and completeness*. SLIPO succeeds in this objective, by delivering a dataset with increased size and richness. Specifically, the produced dataset has greater number of food-related POIs by **88%** compared to the original dataset (*and **121%** more POIs in total*) and with **81%** more attributes, while simultaneously increasing its completeness by **2%**. This is a significant improvement compared to the original dataset, resulting into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over companies offering similar thematic POI data assets by offering a dataset of increase size and richness, (b) increased consumer satisfaction by enabling a more complete data asset, (c) increase in the cost (and profits) for the provision of the data asset to clients.

2.5. DINUC C.1

Sector: Navigation – EV Charging station search

This data integration use case presents a significant challenge for Electric Vehicle (EV) management in navigation services. It refers to integration of EV POIs, mainly EV charging stations, from different providers,

given the large diversity in models, standards and protocols for representing EV POIs. This makes the integration of EV POI content from several suppliers a major challenge.

This use case mainly assesses the efficiency of the integration process, as well as the quality of the results produced. Considering the above, there are several KPIs that can meaningfully measure the added value of the SLIPO system:

- **Efficiency**. The KPI measures *the effort required by the user* of the SLIPO Workbench, starting from data loading, including multiple cycles of re-configuration to optimize data integration results, and ending with the provision of the integration results. The result is compared against the time currently required to perform a similar workflow manually.

  The time required for an expert user of the SLIPO Workbench to perform this DINUC is **1.48 person-days**, while the effort to perform data integration manually is 0.32 person-months or **6.83 person-days**. As such, integration via SLIPO requires **21.6%** of the effort compared against manual integration.
- **Coverage**. This KPI measures the number of *EV-related POIs* that were added to the final dataset compared to those *EV-related POIs* that existed in the original dataset.

  The integrated dataset contains **11,955** EV POIs against only **1,845** EV POIs in the original dataset. This indicates an *increase by* **548%** in the number of *EV-related* POIs, i.e., practically more than quintupled in size.
- **Completeness**. The KPI measures *the average per POI number of non-empty POI property values* (e.g., for address, website, phone properties) in the final dataset concerning *EV-related POIs*. This number is compared to the respective value of *EV-related POIs* in the initial dataset.

  In the integrated dataset, there are **21.51** *property values per EV-related POI*, whereas in the initial data there are **12.45** *property values per EV-related POI*. This means that after integration, on average, there is a significant *increase by* **73%** in the available properties per EV-related POI.
- **Scalability**. The KPI evaluates whether the SLIPO Workbench can support the data sizes of the described scenario. Since there is no baseline procedure to compare with, the KPI comprises a Boolean value that assesses whether the SLIPO Workbench was able to integrate the specific data in an acceptable time.

  The value of this indicator is 'Yes', since SLIPO successfully executed this DINUC in **1.6 hours**.
- **User friendliness**. The KPI evaluates how complex is the usage of the system to a user. It comprises a user rating in a scale of [1-5].

  The value of the KPI is **4.2/5** and has been calculated by averaging the scores provided by the users of the SLIPO system using the system with the roles of 'Data Integrator' and 'QA User'.

These results have the following direct and indirect benefits on a business setting:

- *Increase efficiency*. SLIPO succeeds in this objective by requiring **21.56%** of the effort compared against manual integration (i.e., data loading, multiple cycles of re-configuration to optimize data integration results, provision of the output), with subsequent executions requiring only a few minutes (*i.e., negligible effort*). In addition, integration can be performed at a much larger scale (*e.g.,*

EU-wide) without any increase in required effort. These result into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over companies offering similar thematic POI data assets by offering a dataset at a fraction of the effort, (b) potential expansion of the asset's geographical coverage (*e.g., country/EU-wide*) at a fraction of the effort, (c) increased consumer satisfaction by enabling a more timely data asset, (d) increase in the cost (and profits) for the provision of the data asset to clients.

- *Increase quality*. SLIPO succeeds in this objective, by delivering a dataset with increased size and richness. Specifically, the dataset has greater number of EV-related POIs by **548%** compared to the original dataset and with **73%** more property values per POIs. This results into one or more of the following (*depending on the specific client and business case*): (a) competitive advantage over companies offering similar thematic POI data assets by offering a dataset of increase size and richness, (b) increased consumer satisfaction by enabling a more complete data asset, (c) increase in the cost (and profits) for the provision of the data asset to clients.

3. Comparative Evaluation

In the context of the system evaluation and to further highlight the scope and goals of POI data integration in a real-world setting, we designed and implemented an additional evaluation methodology to directly compare and assess the efficiency and accuracy of SLIPO vs. current POI data integration approaches.

Specifically, we selected three (3) DINUCs, and for each one five (5) geographic areas comprising *at least* 200 POIs to be integrated in total. For each area, five (5) data integrators *independently* performed the data integration *manually*, using only the standard software tools available to them and a *common* methodology, measuring the *effort* involved and extrapolating for the *entire* geographical area of each DINUC. Next, we performed the data integration process via SLIPO using the optimal configurations for each DINUC (*i.e., configurations optimized for each use case by our experts*), and *automatically* (*i.e., configurations selected automatically by each component*) and similarly measured the involved effort. Next, we directly compared both the effort of these approaches and the integrated POIs obtained by each, allowing us to assess the *productivity* and *efficiency* gains, as well as directly compare the *accuracy* of POIs produced by the SLIPO system vs. those produce manually.

These findings further confirm the *orders of magnitude* increases in productivity and cost-effectiveness offered by SLIPO, as well as the practically *identical* results achieved compared to expert user- and manual-driven POI integration. Finally, these results serve as a clear and necessary reminder when framing and assessing the scope, ambition, success, and impact of our work.

3.1. Evaluation Methodology

In this section we establish and elaborate on the evaluation methodology we applied to assess the accuracy and efficiency of the SLIPO system compared to baseline manual- and expert-driven data integration.

3.1.1. Manual POI integration

We assembled a panel of five (5) experts[1] to carry out the manual data integration process, all of which were versed in manual data integration, but with different backgrounds and responsibilities representing the real-world mix of individuals involved in a typical data integration process.

For each the three (3) DINUCs to be examined (A.1, A.2, B.1), we requested from the corresponding industrial partner to prepare a concise set of guidelines describing how the manual data integration process should be performed, replicating to the fullest degree the manual process already performed by their organizations. In addition, each industrial partner was requested to select five (5) areas of interest for each DINUC (*i.e., bounding boxes*) with different characteristics (e.g., urban, rural, islands) for which the data integration should take place. This set of guidelines contains the following for each DINUC:

[1] Two (2) additional experts were recruited and performed the manual integration tasks to mitigate the risk of potential delays in the delivery of the panel's output; the output of these two (2) experts was not ultimately used, as the panel performed their assigned tasks on time.

- *Bounding box*. The five (5) bounding boxes containing the input assets to be used for the data integration job. The panel was provided with access to the full input datasets and were instructed to keep only the subset corresponding to the prescribed five (5) areas of interest.
- *Data integration scope*: A link was provided to the elaborated description for the specific DINUC,.
- *Tools*. The collection of software tools to be used when performing the data integration job. These comprised the software currently used by the responsible industrial partner to actually perform data integration in a real-world setting.
- *Integration steps*. A detailed list of steps to be followed for performing data integration, including directions to homogenously handle linking, fusion, and any other use case- and tool-specific issues (*e.g., base maps, visual inspection, name abbreviation schemes*). The provided steps resemble the instructions given to data integrators when performing the integration in a real-world setting.

Next, the panel was provided with these guidelines and received 2h remote hands-on training presenting in detail the process to be applied. Panel participants were reminded to follow the guidelines as provided to them and were requested to perform the manual data integration at their own accord, documenting (a) the exact *time* it took them to perform integration for each area of each DINUC (*not including any breaks, pauses, etc.*), and (b) prepare and send the output dataset to the corresponding industrial partner.

The following table contains the total time required by each panel member, area, and use case.

Table 1: Time required to perform data integration per DINUC, area, and panel member

	Input Data (#POIs)		Panel Member (time in minutes)				
DINUC A.1							
Area	Dataset A	Dataset B	A	B	C	D	E
A.1.1	34	27766	156	48	90	108	110
A.1.2	37	24008	62	35	70	100	95
A.1.3	40	8287	61	55	45	100	120
A.1.4	39	3615	50	52	50	60	100
A.1.5	34	5846	38	51	65	80	90
DINUC A.2							
Area	Dataset A	Dataset B	A	B	C	D	E
A.2.1	110	299	144	115	100	80	110
A.2.2	97	248	78	108	75	80	90
A.2.3	94	178	81	78	50	80	100
A.2.4	95	154	68	81	50	71	110
A.2.5	104	333	73	109	65	68	120
DINUC B.1							
Area	Dataset A	Dataset B	A	B	C	D	E
B.1.1	110	80	87	71	55	105	100
B.1.2	100	201	60	64	30	97	90
B.1.3	108	89	27	60	30	56	60

B.1.4	103	198	27	67	25	105	105
B.1.5	102	535	57	61	40	105	155

3.1.2. SLIPO POI integration

We performed the data integration workflow for each DINUC via the SLIPO system (*only one execution*):

- Data integration was performed for the entire data space of the input data, and not only for the selected five (5) areas per DINUC. The reason is that due to the small number of POIs within the five (5) areas, system runtime would be very small (*thus lowering resolution*) and comprise mostly from initialization/overhead, rather than actual data integration time.
- The configuration parameters for each individual SLIPO Toolkit component were those created by the data integrator experts. As such, they are considered as the *optimal* that can be achieved.

The following table contains the total runtime required for each DINUC.

Table 2: Total runtime required per each DINUC and area (*normal mode*)

	Runtime (min)
DINUC A.1	210
DINUC A.2	3
DINUC B.1	22

3.1.3. Automated SLIPO POI integration

We performed the data integration workflow for each DINUC via the SLIPO system (*only one execution*), using *all automated facilities* of the system. This is what we consider as a *hands-off* method, which despite its unrealistic nature, was worth to examine. To ensure clarity:

- Data integration was performed for the entire data space of the input data, and not only for the selected five (5) areas per DINUC. The reason is that due to the small number of POIs within the five (5) areas, system runtime would be very small (*thus lowering resolution*) and comprise mostly from initialization/overhead, rather than actual data integration time.
- The configuration parameters for each individual SLIPO Toolkit component were automatically selected by each component. As such, they are not considered as the *optimal* that can be achieved.

The following table contains the total runtime required for each DINUC.

Table 3: Total runtime required per each DINUC and area (*automated mode*)

	Runtime (min)
DINUC A.1	136
DINUC A.2	3
DINUC B.1	22

3.1.4. Accuracy

To determine the accuracy of SLIPO (*normal and automated modes*) vs. the expert- and manual-driven integration, we defined the *similarity score SSC*. A perfect score is '100', in which case all POIs (geometries and remaining features) match exactly.

We have calculated the similarity score SSC to compare the normal and automated output of SLIPO vs. the output of the manual integration, per DINUC and panel member output[2]. Specifically:

- Table 4 presents the similarity score SSC of the normal integration output vs. the output of manual integration for each panel member and area of interest. Further, the *average* similarity scores per area of interest are also calculated.
- Table 5 presents the similarity score SSC of the automated integration output vs. the output of manual integration for each panel member and area of interest. Further, the *average* similarity scores per area of interest are also calculated.
- Table 6 presents the *average similarity* scores SSC for the normal and automated modes per area of interest across all panel members (last column of Table 4 and Table 5), as well as the average similarity score for each DINUC (*rows 'Average'*). Further, Table 6 also presents a similarity score calculated by *LIMES* when comparing the normal and automated output of SLIPO vs. the output of the manual integration process, per DINUC and panel member. In this setting, LIMES is used as a generic-purpose interlinking software, discovering links between the two input datasets, with a similarity score calculated as follows:
  - Link the dataset produced manually with the dataset produced by SLIPO:
    - Keep only *one* link per POI of the manual dataset (i.e., *the one with the highest score*);
    - If no links are discovered, set the score for the particular POI to 0;
  - Average the scores calculated by LIMES for each POI of the manually produced dataset.

  Given the sensitivity of LIMES to configuration changes and the input datasets, the score produced can only serve as an *indication* of the similarity of the two datasets, which we only report for completeness.

Table 4: Similarity score SSC per DINUC, area of interest and panel member of the normal data integration output vs. the manual data integration output

DINUC A.1	Panel Member					
Area	A	B	C	D	E	Average
A.1.1	82,1	90,4	78,1	86,8	81,3	**83,74**
A.1.2	83,9	83,8	81,6	76,5	81,9	**81,54**
A.1.3	80,5	76,4	83,4	83,4	88,7	**82,48**
A.1.4	79,2	59,7	85,1	91,2	64	**75,84**
A.1.5	90,7	95,3	93,5	75,9	65	**84,08**
DINUC A.2	Panel Member					

[2] Please be reminded that the manual data integration output of each expert may differ; hence the similarity score per DINUC and panel member may also differ.

Area	A	B	C	D	E	Average
A.2.1	70,7	64,1	70,9	83,2	71,5	**72,08**
A.2.2	67,5	67,7	67,2	78,6	67,7	**69,74**
A.2.3	80	71	71	81	71	**74,8**
A.2.4	62,7	62,8	62,9	69,3	64,6	**64,46**
A.2.5	72,3	72,2	72,5	77,5	75	**73,9**
DINUC B.1	Panel Member					
Area	A	B	C	D	E	Average
B.1.1	70,4	87,1	87,4	83	81,9	**81,96**
B.1.2	88,6	78,8	96,2	70,8	64,1	**79,7**
B.1.3	100	86,7	100	85,4	86,8	**91,78**
B.1.4	99,8	87,7	99,8	53	51,5	**78,36**
B.1.5	94,3	58,6	88,9	81,6	83	**81,28**

Table 5: Similarity score SSC per DINUC, area of interest and panel member of the automated data integration output vs. the manual data integration output

DINUC A.1	Panel Member					
Area	A	B	C	D	E	Average
A.1.1	82,9	91,3	79,1	87,6	81,3	**84,44**
A.1.2	87,8	88,4	86,1	80,5	85	**85,56**
A.1.3	81,3	77,8	83,9	85,4	90,2	**83,72**
A.1.4	82,8	62,7	88,7	96,3	65,1	**79,12**
A.1.5	94	95,3	96,9	78,5	65,9	**86,12**
DINUC A.2	Panel Member					
Area	A	B	C	D	E	Average
A.2.1	76,3	69,2	76,7	90	75,2	**77,48**
A.2.2	76,4	76,7	76,1	87,2	75,7	**78,42**
A.2.3	77,3	77,1	77,6	87,9	77,2	**79,42**
A.2.4	78,6	79	79,7	86,2	76,5	**80**
A.2.5	83,1	82,3	84,1	88,3	82,2	**84**
DINUC B.1	Panel Member					
Area	A	B	C	D	E	Average
B.1.1	70,4	84,9	85,1	83	82	**81,08**
B.1.2	87,5	78,9	95,1	75,9	65,2	**80,52**
B.1.3	100	86,7	100	85,4	86,8	**91,78**
B.1.4	97,4	85,3	97,3	55,4	54	**77,88**
B.1.5	91,9	58,9	87	79,9	83,7	**80,28**

Table 6: Average similarity scores SSC per DINUC and area of interest for normal and automated modes; average similarity calculated from LIMES per DINUC and area for normal and automated modes

	Normal mode (SSC score)	Automated Mode (SSC score)	Normal mode (LIMES)	Automated Mode (LIMES)
DINUC A.1				
Area				
A.1.1	83,7	84,4	79,8	78,0
A.1.2	81,5	85,6	76,1	78,2
A.1.3	82,5	83,7	75,9	77,9
A.1.4	75,8	79,1	75,7	80,7
A.1.5	84,1	86,1	72,4	73,9
Average	**81,5**	**83,8**	**76,0**	**77,7**
DINUC A.2				
Area				
A.2.1	72,1	77,5	89,3	89,7
A.2.2	69,7	78,4	88,7	87,6
A.2.3	74,8	79,4	90,3	91,5
A.2.4	64,5	80,0	89,8	91,0
A.2.5	73,9	84,0	91,7	94,1
Average	**71,0**	**79,9**	**90,0**	**90,8**
DINUC B.1				
Area				
B.1.1	82,0	81,1	94,0	94,1
B.1.2	79,7	80,5	96,5	94,0
B.1.3	91,8	91,8	96,6	92,1
B.1.4	78,4	77,9	92,1	83,6
B.1.5	81,3	80,3	97,1	95,8
Average	**82,6**	**82,3**	**95,2**	**91,9**

Finally, we introduced a third metric calculated by the posterior **blind manual** validation of the manual and SLIPO (*normal and automated mode*) integrated datasets. Specifically, we recruited **two (2) experts** *not involved in the manual integration efforts (i.e., they have not been subjected previously to our experimental methodology)*, to manually examine the input datasets for DINUCs A.1 and B.1 with the three (3) respective outputs for each case (*i.e., manual, SLIPO normal, SLIPO automated*). The experts were not provided with identifying information regarding the provenance of the output datasets, ensuring a blind examination. As such, they were provided with a list of the **30 combinations** of input and output datasets (2 DINUCs * 5 areas of interest * 3 types of integration output) and were instructed to calculate a score establishing the accuracy of the output when compared against the input datasets.

The resulting scores are presented in the following table.

Table 7: Blind manual validation scores of integration output for DINUCs A.1 and B.1 (manual, SLIPO-normal, SLIPO-automated)

	Manual integration		SLIPO integration (normal)		SLIPO integration (automated)	
	Expert A	Expert B	Expert A	Expert B	Expert A	Expert B
DINUC A.1						
Average of Expert	92,0	77,3	89,5	76,8	77,3	75,5
Average	**84,6**		**83,1**		**76,4**	
DINUC B.1						
Average of Expert	93,0	83,0	88,0	88,0	88,0	82,0
Average	**88,0**		**88,0**		**85,0**	

3.1.5. Productivity

To assess the differences in *efficiency* between the expert- and manual-driven data integration vs. the SLIPO approach (*both for normal and automated modes*), we first estimated the average person-months required to complete the *entire* DINUC (*i.e., the entire data space*) from a single expert. Hence, for each DINUC:

- We calculated the percentage of the number of POIs in the left input dataset vs. the number of POIs in the examined area of interest for the same dataset used by the expert (C);
- We calculated the person-days (PC) required by the expert to perform manual data integration for the area of interest it was provided with, after considering that during a person-day (8 hours) an expert can work only for 45 consecutive minutes (*followed by compulsory 15-minute breaks*) and at most for 6 45-minute sessions per day (*typical practice to ensure quality and avoid fatigue*);
- We estimated the person-days (PT) required by the expert to perform manual data integration for the *entire* DINUC by dividing PC with C.
- We averaged all estimates of person-days (PT) with the number of experts. Therefore, the rows '*Est. effort (entire DINUC)*' in Table 9 provide:
    - The average person-days required to perform manual data integration for the entire DINUC by each panel member, and
    - The final cell of each row provides the average person-days required to perform manual data integration for the entire DINUC across all panel members. As such, the average estimated effort for the manual data integration of **DINUC A.1 is 7.40 person-days, for DINUC A.2 19.01 person-days, and for DINUC B.1 178.62 person-days**.

The results are presented in the following Table.

Table 8: Estimated effort in person-days per DINUC for manual data integration

	Panel Expert (effort in person-days)					Average
DINUC A.1						
Area	A	B	C	D	E	
A.1.1	0,58	0,18	0,33	0,40	0,41	0,38
A.1.2	0,23	0,13	0,26	0,37	0,35	0,27

A.1.3	0,23	0,20	0,17	0,37	0,44	0,28
A.1.4	0,19	0,19	0,19	0,22	0,37	0,23
A.1.5	0,14	0,19	0,24	0,30	0,33	0,24
Est. effort (entire DINUC)	7,26	4,68	6,31	8,74	9,99	**7,40**
DINUC A.2						
Area	A	B	C	D	E	
A.2.1	0,53	0,43	0,37	0,30	0,41	0,41
A.2.2	0,29	0,40	0,28	0,30	0,33	0,32
A.2.3	0,30	0,29	0,19	0,30	0,37	0,29
A.2.4	0,25	0,30	0,19	0,26	0,41	0,28
A.2.5	0,27	0,40	0,24	0,25	0,44	0,32
Est. effort (entire DINUC)	19,19	21,36	14,71	16,62	23,17	**19,01**
DINUC B.1						
Area	A	B	C	D	E	
B.1.1	0,32	0,26	0,20	0,39	0,37	0,31
B.1.2	0,22	0,24	0,11	0,36	0,33	0,25
B.1.3	0,10	0,22	0,11	0,21	0,22	0,17
B.1.4	0,10	0,25	0,09	0,39	0,39	0,24
B.1.5	0,21	0,23	0,15	0,39	0,57	0,31
Est. effort (entire DINUC)	131,98	165,63	91,81	240,81	262,88	**178,62**

In the following, we established the total effort in person-days to perform data integration using SLIPO, both for the normal and automated modes. Specifically:

- *Normal mode*. We calculated the time required by an expert (*extracted by the system's logs*) to enter the input data in the system, create the data integration workflow, setup the required configuration parameters, perform any ad hoc optimizations, and produce the data integration output. Next, we calculated the corresponding person-days using the same considerations as in the manual mode (*i.e., actual working hours per person-day*).
- *Automatic mode*. We calculated the time required by an expert (*extracted by the system's logs*) to enter the input data in the system, create the data integration workflow using the automatically provided configurations, and produce the data integration output. Next, we calculated the corresponding person-days using the same considerations as in the manual mode (*i.e., actual working hours per person-day*).

The results are presented in the following Table.

Table 9: Effort in person-days for SLIPO integration (normal and automated modes)

	Normal Mode	Automated Mode
DINUC A.1		
Effort	**0,89**	**0,11**
DINUC A.2		
Effort	**1,33**	**0,11**
DINUC B.1		

Effort	1,04	0,15

Finally, we estimated the *total cost* required to perform data integration for each DINUC. Specifically:

- *Manual data integration*. We assumed *no cost* for the software used (*since open source software can be easily applied*) and using an average person-month cost of 4.5K Euros (*22 working days per month*), we estimated the total cost for the manual data integration of the entire data space and for each DINUC.
- *SLIPO (normal mode)*. We assumed a cost of 1K Euros[3] for using the system and using an average person-month cost of 9K Euros[4] (*22 working days per month*), we estimated the total cost for the normal data integration of the entire data space and for each DINUC.
- *SLIPO (automated mode)*. We assumed a cost of 1K Euros for using the system and using an average person-month cost of 4.5K Euros (*21 working days per month*), we estimated the total cost for the automated data integration of the entire data space and for each DINUC.

The resulting cost estimations and corresponding savings are presented in the following Table.

Table 10: Estimated total cost in Euros for manual and SLIPO data integration (normal and automated modes)

	Manual	SLIPO (normal) – (savings)	SLIPO (automated) – (savings)
DINUC A.1			
Cost	1.512,88 €	1.363,64 € (9.9%)	1.023,81 € (35.4%)
DINUC A.2			
Cost	3.888,91 €	1.545,45 € (60.3%)	1.001,08 € (75.4%)
DINUC B.1			
Cost	36.536,57 €	1.424,24 € (96.1%)	1.001,44 € (97.4%)

[3] This is an exaggerated cost assuming a one-time cost for using SLIPO and does not reflect a typical annual and resource-aware pricing policy.

[4] Involved personnel have a higher average PM rate due to the increased technical skills required.

4. Evaluation Findings

In this section, we collate and discuss our findings from the extensive evaluation and validation of the entire SLIPO system. Overall, we consider that SLIPO has fully achieved its innovation objectives. SLIPO provides data integration output with a **similar accuracy** when compared against manual data integration, at a **fraction** of the comparative **effort and cost**, and is inherently **scalable** in terms of the size and frequency of data integration workflows. As such, SLIPO **empowers** industrial stakeholders with the means to **cost-effectively scale** POI data integration well beyond their current means, enabling them to harness **value** from their assets and knowhow, and expand the **reach** of their product and service offerings.

4.1. KPI Performance

The performance of the SLIPO system across all established KPIs and DINUCs has been extremely successful, with key highlights the increased efficiency, coverage, and size of its output.

- **Enrichment**. As expected, enrichment posed no challenge for the system, as all additional attributes identified in the input datasets to be integrated were successfully handled. The values of this indicator serve to portray the increase in the attributes of the original dataset used by each corresponding DINUC, which in turn can potentially increase the scope and relevance of the integrated output, as well as its applicability into new service and application scenarios.
- **Completeness**. The values of this KPI is directly dependent on the characteristics of the input datasets. As such, in cases where (a) the input data contained a low number of non-empty property values (e.g., DINUC A.1), or where (b) the input data with high number of non-empty property values was of comparable size with the other input data used (e.g., DINUC A.2), completeness of the integrated output increased. In contrast, in cases where the comparative size of the input dataset with a high number of non-empty values was dominant (e.g., DINUC B.1), completeness slightly decreased. This is to be expected, but it should be noted that in these cases the output dataset also more than *doubled* in size, with a marginal (<3%) decrease in completeness.
- **Size**. The size of the final integrated dataset at all relevant DINUCs more than doubled, thus resulting into a tangible impact in the perceived value of the output dataset compared to the original one. As identified early in the project's duration, the absolute size of POI data assets has *both* a direct value for stakeholders, as well as a direct impact in marketing activities, with the advertised size of POI assets positioned as a competitive advantage of different POI products, driving client demand.
- **Coverage**. The thematic coverage of the final integrated output significantly increased compared to the original POI asset used by the corresponding industrial partners, thus directly contributing to POI data assets with an increased thematic relevance and usefulness for end-users.
- **Efficiency**. The efficiency of the SLIPO-powered data integration compared to standard manual integration effort, was as anticipated, significant, with the benefits increasing as the size of the input data assets expanded. Further, when our expert users assessed the quality of the SLIPO output, they empirically evaluated it as especially high. This empirical finding was extremely important to further

explore and document in detail, as it would enable us to assess SLIPO from a cost-benefit perspective. Towards this, we pursued an expended experimental evaluation protocol (Section 3), in which we compared the output of the SLIPO system against the output of a manual integrator. These findings are discussed next and signal the success of SLIPO in essentially *replicating* the output of the resource and time-intensive manual integration.

- **Scalability**. The SLIPO system was arguably not challenged by the size and complexity of the data assets and workloads involved across all DINUCs, completing its execution in a few minutes and up to a few hours. This was anticipated, as our focus in the pilot was on evaluating SLIPO's operation in real-world industrial use cases, which are understandably grounded in coverage by the practical constraints of current data integration practices. This directly relates to our ambition for the exploitation of SLIPO; first validate its operation and effectiveness for *current* use cases, and then prove it can support the current data integration workflows at a much larger scale. As such, SLIPO's adoption would allow the transfer of *existing* data integration workflows and thus services/products and knowhow into new markets, at a much larger scale, and with minimal effort. Towards this, we have **successfully benchmarked the SLIPO system at a world-scale *(>90MPOIs)***, an activity presented in Deliverable D1.4 "Final SLIPO system".
- **User friendliness**. The end-users of the SLIPO system were satisfied with the usability and overall friendliness of the system at the end of the pilot, a finding also confirmed by anecdotal feedback. It is worth highlighting that the achieved score was much lower in the early stages of the pilot, while it also differs according to the perceived complexity of each DINUC. The observed increase in the user friendliness score as the pilot progressed is understandable, due to the improvements introduced in the system to streamline and simplify user experience, as well due to the increased familiarity of the users resulting from their prolonged exposure to the system.
- **Advanced facilities**. The power users of the SLIPO system (i.e., workflow editors, admins) rate the system in terms of advanced feature quite high, with a remarkable improvement compared against the start of the pilot. This was a direct result of the improvements and new features we introduced according to their feedback and observations.

4.2. Comparative Evaluation

Our extensive efforts to objectively assess the accuracy of SLIPO's output against the output of manual POI integration in a transparent and disciplined manner (see Section 3), has paid dividends in documenting the exceptional value of the system, as well as in validating our original motivation behind the project. The obtained results *exceeded* our expectations, leading us to additional scrutiny, by introducing alternate scores and evaluation methodologies, as well as examining and validating the operation of the system in detail. These actions have only confirmed our findings.

4.2.1. Accuracy

The output of data integration workflows produced by **the SLIPO system essentially *replicates* the output of data integration performed manually by human experts**. We can observe from the scores reported in Section 3 that the output of SLIPO (*both normal and automated modes*) and for all comparative scores examined, are very *high* and practically *identical* to the scores achieved when comparing the output produced manually by experts. To frame our finding differently: the deviation of the output produced by manual experts is practically the same with the one when comparing the output of SLIPO with that of manual users. Hence, **SLIPO's accuracy can be considered as practically identical to the much more resource-intensive output of manual integration**.

In addition, the **automated operation of SLIPO** (i.e., SLIPO automated mode) **produces results of *at least similar accuracy*, when compared against the output of data integration workflows prepared and optimized by experts** (i.e., SLIPO normal mode). This finding may appear surprising at first, since it implies that an expert data integrator manually and iteratively defining and testing different configurations for transforming, linking, and fusing data in SLIPO, will *at best match* the accuracy score for an entirely automated process. However, upon closer inspection this observation is understandable. First, SLIPO focuses on the *highly-specific* data integration challenge of *POI* data assets, allowing us to train and use accurate ML models that can be transferred across most POI data integration scenarios. This validates our motivation to pursue ML-based methods to automate data integration, given the specialized domain of our Big Data assets. Second, we need to consider the typical workflows of expert data integrators when using the system, in which the output of interim steps is typically visually examined to inform subsequent configuration changes in the SLIPO toolkit components. This methodology is inherently biased, as it focuses on arbitrary and non-random areas of interest (e.g., a city center, business district), which cannot scale for Big POI data assets.

However, we must also raise an additional point regarding the real-world relevance of our finding. The training space used for the various ML models deployed is objectively small and similar to the evaluation space. For example, in POI transformation we observed that most POI schemas share very common approaches for the core attributes of POI assets (i.e., name, geometry, address) which is of course not surprising; POIs are specialized entities modeled throughout the world in a similar manner, while many POI assets are essentially derivatives of authoritative or de facto data assets. Similar observations hold for interlinking and fusion operations. The former can automatically derive optimized linking configurations across the POI space, avoiding the bias of human experts in specific areas of interest. The latter can very accurately replicate the relatively small number of distinct fusion operations of human experts. As such, we need to emphasize that we consider the automated mode as the *best-fit for most POI data integration scenarios addressed by most stakeholders of the value chain*. We anticipate that the automated operation will not produce similarly good results when applied for complex unknown schemas or non-homogeneous POI assets (*e.g., systematical errors in POIs name for large geographical areas or categories*).

4.2.2. Productivity

Our findings regarding the comparison of effort required to perform manual vs. the normal and automated SLIPO data integration workflows were as anticipated. First, manual data integration effort increases with the number POIs and the complexity of the data integration effort, ranging from a few person-days to tens

of person-months, which makes (as expected) manual data integration an extremely resource-intensive and unrealistic endeavor for big data assets and complex data integration flows. This burden multiplies considerably when data integration workflows are required to be performed *periodically*, making even relatively small and simple cases resource-intensive. Second, the effort required to author and perform data integration workflows in SLIPO is **practically the same** regardless of the type of problem and size of assets. Third, the automated mode requires only **10%-15%** of the effort when compared against the normal mode, since it entirely abolishes the effort spent for defining, testing, and optimizing configurations for the various SLIPO toolkit components.

4.2.3. Cost-benefit

Our attempt to highlight our findings from a cost-benefit perspective using very modest estimates regarding the cost of the system, portray a very interesting picture. At all cases, SLIPO provides **savings of at least 10%** against manual data integration, which rise to **~97%** for larger and more complex POI data integration workflows. This means that even for relatively simple data integration workflows, and despite the need for an expert data integrator, *SLIPO remains the optimal approach*. Further, should we consider the case of data integration workflows being required to be periodically performed, the benefits are even greater; SLIPO flows can executed *as-is*, or with minimal configurations changes, while repurposing effort in manual data integration is infeasible. When examining the efficiency of SLIPO's automated mode, the benefits are even greater, with **savings at least ~33%** compared against manual integration, and certainly higher than SLIPO's normal mode since the system does not require an expert user for the definition and optimization of data integration workflows.

4.3. Summary

In the following we summarize (a) the key findings established from the real-world evaluation of SLIPO, and its (b) competitive advantages and USPs.

- SLIPO's fully **automated** operation delivers data integration output which is practically **identical** with that of data integration performed **manually** by human experts.
- SLIPO's fully automated mode delivers data integration output of **at least similar accuracy**, when compared against the output of data integration workflows prepared and optimized by **experts**.
- The effort required to edit and optimize data integration workflows in SLIPO is a **fraction** of the effort of manual data integration, even for small data integration jobs, and is practically the **same** regardless of the data integration's geographic **coverage** and **size** of data assets.
- SLIPO provides **savings of at least 10%** against manual data integration, which rise to **~97%** for larger and more complex POI data integration workflows.
- SLIPO's data integration workflows can be applied and be executed *as-is* at a **world-scale**, without any changes required by the user, or to the underlying system.

- SLIPO requires **no specialized expertise** from its end-users to perform data integration jobs; any user with a **basic** understanding of GIS and integration concepts can fully exploit the system with minimal training and effort.
- SLIPO requires no local installation or configuration; it is a trusted and secure **cloud-based** system (SaaS) providing its users with private data-spaces and user management.
- SLIPO can fully support and be integrated with any existing **Quality Assurance** workflows to ensure the quality of geospatial data assets.
- SLIPO maintains and presents full **provenance** information for the data integration jobs at the POI and attribute level, enabling users to know exactly the lineage of the information in their assets.
- SLIPO's services are fully exposed via its open **API** to any **third-party** workflow/application and can be integrated with them as required by the end-users.
- SLIPO provides **value-added** services and support for industrial **data science** pipelines via its integrated suite of **geospatial analysis** algorithms and Python-based **APIs**.