DELIVERABLE D3.5

Initial report on Quality Assurance Metrics, Indicators and Processes

SLIPO

PROJECT NUMBER: 731581

START DATE OF PROJECT: 01/01/2017

DURATION: 36 months

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

Horizon 2020

Dissemination Level	Public
Due Date of Deliverable	Month 24, 31/12/2018
Actual Submission Date	20/12/2018
Work Package	WP3 - Data Integration and Quality Assurance
Tasks	3.5 Quality Assurance
Type	Report
Approval Status	Submitted for approval
Version	1.0
Number of Pages	42
Filename	D3.5_Initial_Report_on_Quality_Assurance_Metrics.pdf

Abstract

This report presents the initial Linked-Data-specific measures supported by the SLIPO Workbench and its components for assessing the quality of integrated POI datasets. We briefly describe the current state of the art in geospatial linked data quality assurance. Then, we present the quality metrics at each step of the POI integration lifecycle: transformation (TripleGeo), POI interlinking (LIMES), POI fusion (FAGI) and POI enrichment (DEER). Finally, we present an exemplary assessment of the proposed quality assurance metrics on a real-world example.

History

version	date	reason	revised by
0.1	11/11/2018	Deliverable structure and initial content	Mohamed Ahmed Sherif
0.2	13/11/2018	Quality Measures for transformation	Spiros Athanasiou, Kostas Patroumpas
0.3	22/11/2018	Introduction and state of the art	Mohamed Ahmed Sherif, Abdullah Fathi Ahmed
0.4	2/12/2018	Quality metrics for interlinking and enrichment	Mohamed Ahmed Sherif, Kevin Dreßler
0.5	8/12/2018	Evaluation scenario & quality of transformation results	Kostas Patroumpas, Nikos Karagiannakis
0.6	14/12/2018	Quality results of interlinking and enrichment	Kevin Dreßler
0.7	16/12/2018	First draft formatting	Mohamed Ahmed Sherif, Abdullah Fathi Ahmed
0.8	19/12/2018	Proof-reading	Spiros Athanasiou
1.0	20/12/2018	Final Version	Mohamed Ahmed Sherif

Author list

organization	name	contact information
INFAI	Mohamed Ahmed Sherif	sherif@informatik.uni-leipzig.de
INFAI	Kevin Dreßler	dressler@informatik.uni-leipzig.de
INFAI	Abdullah Fathi Ahmed	ahmed@informatik.uni-leipzig.de
TomTom	Pauline Baudens	Pauline.Baudens@tomtom.com
WIGeoGIS	Kai barenscher	kb@wigeogis.com
GET	Thodoris Vakkas	tvakkas@getmap.gr
ATHENA RC	Spiros Athanasiou	spathan@imis.athena-innovation.gr
ATHENA RC	Nikos Karagiannakis	nkaragiannakis@imis.athena-innovation.gr
ATHENA RC	Kostas Patroumpas	kpatro@imis.athena-innovation.gr

Executive Summary

Point of Interest (POI) data constitute the cornerstone of many current applications and services. Such applications and services vary from navigation to social networks, tourism and logistics. We use POI data to search, communicate, decide and plan our actions. POIs are semantically diverse and spatio-temporally evolving entities, having geographical, temporal and thematic relations. Currently, integrating POI data to increase their coverage, timeliness, accuracy and value is a resource-intensive and mostly manual process, with no specialized software available to address the specific challenges of this task. In SLIPO, we implement an integration toolkit capable of transforming, linking, fusing and enriching POI data. In particular, we demonstrate how Linked Data technologies can address the limitations, gaps and challenges of the current landscape in Big POI data integration. We have built a prototype application that enables users to define, manage and execute scalable POI data integration workflows built on top of state-of-the-art software for geospatial Linked Data.

Integrating various POI features from different datasets is not straightforward. Common problems in this respect include distorted geometries and divergent meaning of meta-information. In SLIPO, we aim to address an extensive range of users including customers in an industrial environment. Therefore, the quality of our POI datasets resulting from integrating other datasets becomes a crucial factor for the acceptance and distribution of the project's results. We assure the automation of the quality during all phases of POI integration. Therefore, we build a set of quality metrics at *each stage* of POI integration lifecycle used to assess the quality of the produced results in order to manually or automatically improve the efficiency of the specific data integration step for a particular use case/data asset, feed the following tools along the workflow, and inform the user about the quality of the data integration process. As with any data integration software and workflow, this aims to support an exploratory and iterative process towards gradually improving the data integration output according to the user's own goals (*e.g., increase number of POIs, enrich with information from another source*).

In this report we present our first implementation of quality assurance metrics, indicators and process for POI data integration. We begin by briefly presenting the state of the art in the field of linked data quality assurance and also present a survey of linked data quality features. Thereafter, we present our developed quality metrics for each of the SLIPO integration tools. i.e., *TripleGeo* for POI transformation, *LIMES* for POI interlinking, *FAGI* for POI fusion and *DEER* for POI enrichment. Finally, we carry out an integration process with real datasets, where we present and discuss the resulted quality metrics by each of the SLIPO integration tools.

The structure of this document is as follows: After introducing our work in Section 1, we provide a short overview of the state-of-the art and present a list of the metrics used for POI-related data quality metrics in Section 2. In Section 3, we introduce our quality metrics implemented in each of the SLIPO integration components. Finally, in Section 4 we present an evaluation of our quality metrics using real datasets.

Abbreviations and Acronyms

CEN/TC 287	European Committee for Standardization Technical Committee 287
CGDI	Canadian Geospatial Data Infrastructure
CRS	Coordinate Reference System
DBMS	Data Base Management System
DIGEST	Digital Geographic Information Exchange Standards
ETL	Extract-Transform-Load
GIS	Geographic Information System
JSON	JavaScript Object Notation
MBR	Minimum Bounding Rectangle
MUM	Multidimensional User Manual
NSD	National Spatial Data Infrastructure
OSM	Open Street Map
OWL	Web Ontology Language
POI	Point Of Interest
QIMM	Quality Information Management Mode
QGC	Open GIS Consortium
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SOLAP	Spatial On-Line Analytical Processing
SAIF	Spatial Archiving and Interchange Format
SDTS	Spatial Data Transfer Standard
WGS84	World Geodetic System 1984 (EPSG:4326)

Table of Contents

1. Introduction

Data Quality is the degree of excellence exhibited by the data towards the actual scenario in-use. It is generally thought of as a multi-dimensional concept and is most commonly referred to as "*Fit-for-use*", i.e., some applications are more critical towards high data quality and others may only require data of adequate quality [15]. For example, an application providing soccer players information using DBpedia may not require very high quality of data. On the other hand, prescribing a treatment to a cancer patient (*such as drugs, people ethnicity and countries etc.*) is simply not sufficient.

Point Of Interest (POI) are especially prone to errors because they contain data from multiple providers and use different assumptions about structure and semantics of data [13]. In many of the use cases (e.g., transportation, navigation and tourism), a very high quality of POI data is required. Low POI data quality can result in severe accidents such as: 1998 ski-lift accident in Italy and 1999 accidental bombing of the Chinese Embassy in Belgrade [13]. In recent years, the concern for POI data quality has increased due to a number of factors including [10]:

- Increased data production by the private sector and non-government agencies, which are not governed by uniform quality standards (production of data by national agencies has long been required to conform to national accuracy standards), and
- Increased reliance on secondary data sources, due to the growth of the Internet, data translators, and data transfer standards, making poor quality data ever easier to get.

In this report we provide a survey state-of-the-art metrics used for assessing data quality. Thereafter we introduce the initial version of the quality metrics we implemented in SLIPO to assure the quality of the POI dataset integration process. Finally, we present an evaluation of our quality metrics in a real dataset use case.

In the next section, we provide a short overview of the state-of-the art and present a list of the metrics used for POI-related data quality. In Section 3, we introduce our quality metrics implemented in each of the SLIPO integration components. Finally, we then show an evaluation of our quality metrics in a real dataset in Section 4.

2. State of The Art

In this section, we survey the state of the art pertaining to POI data quality. First, we introduce industrial data quality *standards* related to geospatial assets. Then, we discuss various POI-quality-related scientific research contributions. Finally, we present the current state of the art of data quality *metrics* for POI geospatial data in *general* and geospatial *linked* data in particular.

2.1. POI-related Data Quality Standards

ISO/TC 211[1] provides a series of standards that deal with various aspects of POI-related geospatial data quality. In particular, **ISO 19115-1:2014**[2], **ISO 19113:2002**[3], **ISO 19114:2003**[4] and the technical specification **ISO/TS 19138**[5] Data quality measures are important to be considered.

ISO 19115-1:2014 defines the schema required for describing POI geographic information and services by means of metadata. It provides information about the identification, extent, quality, spatial and temporal aspects, the content, spatial reference, portrayal, distribution, and other properties of digital geographic data and services.

ISO 19113:2002 establishes the principles for describing the quality of POI geographic data and specifies components for reporting quality information. It also provides an approach to organizing information about data quality.

ISO 19114:2003 provides a framework of procedures for determining and evaluating quality that is applicable to any digital geographic dataset such as the POI ones. This ISO consistent with the data quality principles defined in **ISO 19113**. It also establishes a framework for evaluating and reporting data quality results, either as part of data quality metadata only or also as a quality evaluation report.

ISO/TS 19138:2006 defines a set of data quality measures. These can be used when reporting data quality for the data quality sub-elements identified in **ISO 19113**. Multiple measures are defined for each data quality sub-element and the choice of which to use will depend on the type of data and its intended purpose.

---

[1] http://www.isotc211.org/

[2] https://www.iso.org/standard/53798.html

[3] https://www.iso.org/standard/26018.html

[4] https://www.iso.org/standard/26019.html

[5] https://www.iso.org/standard/32556.html

The Spatial Data Transfer Standard (SDTS6), is a robust way of transferring earth-referenced spatial data between dissimilar computer systems with the potential for no information loss. It is a transfer standard that embraces the philosophy of self-contained transfers, i.e. spatial data, attribute, georeferencing, data quality report, data dictionary, and other supporting metadata all included in the transfer.

Other organizations that provide POI-related data quality standards includes: The Open GIS Consortium (**OGC**[7]), the European Committee for Standardization Technical Committee 287 (**CEN/TC 287**[8]), Canadian Geospatial Data Infrastructure (**CGDI**[9]), National Spatial Data Infrastructure (**NSDI**[10]), Digital Geographic Information Exchange Standards **(DIGEST**[11]), and Spatial Archiving and Interchange Format (**SAIF**[12]).

2.2. Geospatial Data Quality Scientific Contributions

The work in [23] explores measurement standards on the quality of open Geospatial data. The main purpose of [23] was the data curation optimization and scientific research support enhancement. A set of dimensions for data quality measurement is proposed in order to develop appropriate metrics. Pipino et al. [19] describes the principles that can help organizations develop usable spatial data quality metrics. The authors of [1] addressed some of the issues in spatial data quality, especially the need to incorporate visualisation of data quality into graphics and maps. [22] provides a detailed discussion of various spatial data quality components. Boin et al. [4] question whether or not the quality information that is typically provided in such spatial metadata is actually effective. This research employs qualitative research approaches to explore how users of spatial data determine the quality of a dataset. Consumer feedback emails and semi-structured interviews have been analyzed to discover the perceptions, actions and goals of individual data consumers from a range of professional backgrounds.

Multidimensional User Manual (MUM) [9] allows the management of geospatial data quality and the communication of the quality information using indicators that can be analyzed at different levels of

---

6 http://www.fgdc.gov/metadata/csdgm/02.html

7 http://www.opengeospatial.org/

8 http://www.centc287.eu/index.php/education-a-training/education-a-training-centc-287

9 https://www.nrcan.gc.ca/earth-sciences/geomatics/canadas-spatial-data-infrastructure/8902

10 http://www.fgdc.gov/nsdi/nsdi.html

11 https://www.dgiwg.org/digest/

12 http://archive.ilmb.gov.bc.ca/crgb/pba/saif/

detail. [2] presents GIS related issues and a set of spatial data quality metrics. Caprioli et al. [5, 6] examined the quality approaches in GIS contexts and some of most meaningful spatial Data Quality standards. Quality Information Management Model (QIMM) [8] allows a user to easily and rapidly navigate into the quality information using a Spatial On-Line Analytical Processing (SOLAP) client-tied to its GIS application. [11] discusses the implementation of various ISO spatial Data Quality standards. Mooney et al. [18] provide measures of quality for Open Street Map (OSM) which operate in an unsupervised manner without reference to a "trusted" source of ground-truth data. Other recent contributions [3, 16, 17] discusses the various spatial Data Quality metrics and their implementations.

In the next Section, based on above scientific contributions and standards, we present the set of spatial Data Quality metrics and rank them according to the number of citations.

2.3. State-Of-The-Art Geospatial Data Quality Metrics

Here we introduce a ranked list of data POI quality metrics:

- **Accuracy**: Accuracy is critical quality metric for location information services. Georeferencing helps align POI entities to the underlying base map, which requires accurate transformation of data. It includes: positional accuracy, attribute accuracy, and temporal accuracy. Positional accuracy refers to the accuracy of the spatial component (*e.g., point, line*), attribute accuracy refers to the accuracy of thematic component (*e.g., type*), and temporal accuracy refers to the agreement between encoded and actual temporal information (*if available*).
- **Consistency**: It refers to the extent to which data is consistent and presented in same format. Data values from various sources referring to the same geospatial feature need to be consistent.
- **Completeness**: The extent to which data is not missing and is of sufficient breadth and depth for the task at hand. Key data fields and the other types of data supporting spatial analysis and presentation should be associated with each POI object to ensure usability and appropriateness of data values.
- **Reputation**: It refers to the extent to which data is highly regarded in terms of its source or content. Sources of POI data may indirectly indicate quality. Authoritative sources can come from the government. Examples include population counts, census tracts, and satellite imagery provided by the public organizations.

- **Currency**: It refers to the extent to which POI data is sufficiently up-to-date for the task at hand. We need to make sure that changes to POI data are updated, both on maps and in text (e.g., opening hours, change of name).
- **Objectivity**: It refers to the extent to which the data is unbiased, unprejudiced, and impartial.
- **Relevancy**: It refers to the extent to which the data is applicable and helpful for the task in hand.
- **Accessibility**: It refers to the extent to which the data is available or easily and quickly retrievable. When open access is taken into consideration, data accessibility and understandability have a new meaning. Moreover, it also means an evaluation on the sources of availability, e.g., satellite imagery from government agencies like NASA vs. imagery from commercial companies.
- **Sufficiency**: It refers to the extent to which the volume of the data is sufficient for the task at hand.
- **Compatibility**: It refers to the extent to which data are compatible between different technological systems.
- **Discoverability**: It refers to the extent to which the data is in appropriate language, symbols, units, and the definitions are clear. It is also known as interoperability, to support metadata harvesting.
- **Integrity**: It refers to the extent to which the data is regarded as true and credible. It is also known as believability, i.e., the user should believe that the data comes from credible source.
- **Repurposing**: This is more an open access issue. However, reusability helps verify the quality of the data.
- **Transparency**: Transparency in the process of the data creation and acquisition can help confirm high standards in the output data.
- **Validity**: It refers to the extent to which the data is reasonable and in correct format. For example, account numbers usually fall within a specific values range, numeric data are all digits, dates always have a valid day, month, and year format.
- **Verifiability**: It refers to the extent to which the data is verifiable. Usually, POI data are verifiable after a manual on-the-field validation process.
- **Visualization**: It refers to the quality of the visual presentation (e.g., color, visibility etc.) of the data in terms of geospatial maps.
- **Value-Added**: It refers to the extent to which the data is beneficial and provides advantages from its use.
- **Resolution**: It refers to minimum size of features that are discernible from a dataset.

- **Lineage**: It refers to proper documentation of the source materials, explaining method of derivation and transformations applied to the initial data.

Overall, accuracy, completeness, and consistency of the data are the *key metrics* for any POI application to produce credible results. Based on the literature, Table 1 shows a ranked list of data quality metrics, which are applicable for POI datasets.

Rank	Metric	Citation
1	Accuracy	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2]
1	Consistency	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2]
1	Completeness	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11 ,2]
2	Lineage	ISO/TC 211, SDTS, [13, 4, 5, 8, 11]
3	Resolution	ISO/TC 211, [19,2,5]
4	Currency	[13, 4, 2]
5	Accessibility	[19, 23]
5	Discoverability	[19, 23]
5	Integrity/Believability	[19, 23]
6	Validity	[19, 23]
7	Objectivity	[19]
7	Relevancy	[19]
7	Security	[19]
7	Sufficiency	[19]
7	Compatibility	[19]
7	Legibility	[19]
7	Repurposing	[19]
7	Transparency	[19]
7	Visualization	[19]

Rank	Metric	Citation
7	Value-Added	[19]
7	Reputation	[19]
7	Verifiability	[19]

Table 1: Ranked list of data quality metrics.

2.4. Geospatial Linked Data Quality Metrics

In this section, we introduce the quality metrics implemented within the *GeoKnow* project[13] in order to measure the quality of linked geospatial datasets. Note that, the GeoKnow project was dealing with geospatial data in *general* and not POIs. These data quality measures have informed our decisions on exactly what type of linked data-specific quality metrics are relevant for the SLIPO components.

- **Number of properties per class.** This metric calculates how many distinct predicates exist for instances of a class. This metric just needs a dataset as input. The output is an integer per class representing the distinct predicates that are used in statements where the subject is an instance of the class.
- **Number of instances per class.** This metric calculates how many distinct instances exist for each class. This metric can be used to weigh the importance of a class in a dataset.
- **Average surface area per class.** This metric calculates the average surface contained in polygons for each class. This metric is important to relativize the number of instances of some class. A class representing continents has only a few instances, but the covered surface is much bigger than that of a class representing a city.
- **Number of intersecting classes per instances.** This metric calculates how many instances have more types that are only in this class. This is important to represent how specific the current class is. In a very specific class, outliers are more significant than in a general class.
- **Average number of points per class.** This metric represents the average of points per class. For each instance of the current class, the metric computes how many points are linked from this instance. This metric is important to differentiate between classes representing multi-point objects and those representing one point.

[13] http://geoknow.eu

- **Average number of polygons per class.** This metric reports the average number of polygons within a class of a geospatial dataset.
- **Average distance between point sets which represent the same resource.** This metric computes the average distance between polygons which represent the same resource in two linked datasets.
- **Class Coverage.** This metric was introduced in [24] and determines how well the instance data conform to `rdf:class` (class for short), i.e., how well a specific class is covered by the different instances of that class.
- **Weighted class coverage.** [24] proposed a mechanism to compute this measure, by considering the weighted sum of the previous class coverage measure.
- **Dataset structuredness.** The overall structuredness or coherence of a dataset is defined (as proposed by [24]) as the weighted of the two previous measures of class coverage and weighted class coverage.

3. SLIPO Quality Metrics

In this section, we provide an overview of the quality metrics implemented at each step of the POI data integration lifecycle supported in SLIPO. The underlying idea of behind the POI data integration in SLIPO is to apply Linked Data technologies, which are ideally suited to handle the inherent geospatial, thematic, and semantic ambiguities of POIs. Hence, existing POI data assets need first to be transformed into RDF, so that individual POI profiles can be interlinked, fused, and enriched. This takes place in successive steps that progressively increase the size and/or the quality of the POI data throughout a virtuous cycle, implementing an iterative workflow as shown in Figure 1.

The process begins with a *transformation* stage. This assumes as input POI data collected from heterogeneous and diverse data sources (*proprietary, open, crowdsourced*), having different attribute schemata and formats. The spatial, temporal, and thematic attributes in the input data are transformed into RDF triples conforming to a common, vendor-agnostic, well-defined, yet agile and extendable POI ontology. Hence, schema mappings from attributes of the original schemata to the classes and properties of this ontology are applied.

Subsequent stages are applied in the Linked Data domain against the previously transformed RDF data comprising an iterative, step-wise workflow that first increases the size and then the quality of POIs. This forms a virtuous cycle that begins by expanding POI coverage, completeness, and richness, delivering data of greater size. Then, it focuses on increasing the quality of the POI data, fusing these intermediate results and enforcing appropriate quality assurance algorithms. This inherently reduces the size of data in absolute numbers, but increases their value. This process can be repeated in the same manner, iteratively increasing the size and then refining to increase quality, as many times as required. For example, an expert user can introduce additional data sources, apply different rules, focus on other types of metadata, etc. Such an iterative workflow involved in the POI data integration stages includes in addition to the transformation stage, the interlinking, fusion and enrichment (See Figure 1).

In the next sections, we introduce our quality metrics implemented in each of the SLIPO components to assure the quality of the generated POIs during the POI data integration lifecycle. In particular, we produce quality assurance metrics at each step of a data integration lifecycle, which are used by:

- The user of the SLIPO toolkit at the end of the integration process to assess the quality of produced results,
- The subsequent tools in the workflow,
- An expert user aiming to optimize the configurations of each individual tool for the particular data integration job, or
- The software itself to provide learning capabilities for different integration tools.

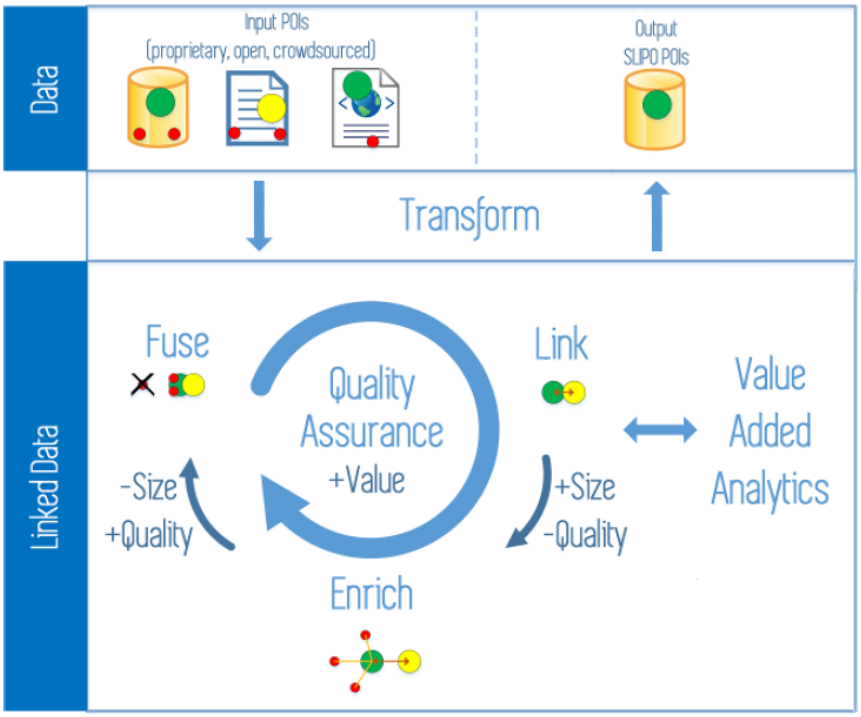


Figure 1: The POI data integration lifecycle.

3.1. TripleGeo

Transformation of conventional POI and third-party datasets into RDF is an essential part of the POI data integration lifecycle, since it enables their subsequent processing (interlinking, fusion, enrichment) as linked data. Towards this goal, our transformation software **TripleGeo**[14] offers advanced capabilities for accessing a large variety of data formats that include world-renowned DBMSs and geospatial de facto file formats and converting them to RDF resources, as documented in Deliverable D2.2 [26].

In short, TripleGeo is designed as a spatially-aware *Extract-Transform-Load* (ETL) tool, enabling users to:

- *Extract* spatial data from a source;
- *Transform* this data into a triple format and geometry vocabulary prescribed by the target RDF store;
- *Load* resulting triples into the target RDF store.

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

The software can take as input not only de facto geographical files (e.g., ESRI shapefiles), but may also access spatial tables hosted in major DBMSs (e.g., Oracle Spatial or PostGIS databases). TripleGeo always preserves data integrity and provides consistent, well-defined geospatial and thematic information complying with an underlying POI ontology.

Upon termination of a transformation process, TripleGeo provides quality statistics and metadata regarding its execution in a JSON file, as listed in Figure 2 for an example involving a transformation task against POI data in Greece. Such statistics and metadata may be utilized in the SLIPO Workbench to provide insightful visualizations (maps, charts, etc., as depicted in Figure 3) about a transformed POI dataset that takes part in a data integration workflow.

More specifically, this JSON file reports performance measurements and quality indicators regarding the data listed in three main categories:

- **Attribute Statistics**: For each attribute in the original input dataset, a count of NOT NULL values on this attribute is given. This reflects the amount of such values that have been actually submitted for transformation. However, due to the underlying POI ontology employed in SLIPO [25], note that the number of resulting triples may be inflated depending on the specified mappings [26]. For the transformation example shown in Figure 3, note that the value counts per attribute differ, as only NOT NULL values are transformed into triples. Based on those attribute statistics, the SLIPO Workbench application provides charts that illustrate the distribution of such attribute values and can assist users in detecting issues about the data. For example, all POIs in this example have names in Greek (attribute: NAME_GR), but not all of them have a name in English as well (attribute: NAME_ENG).
- **Spatial Extent**: A Minimum Bounding Rectangle (MBR) as computed by the spatial extent that covers all transformed geometries. Note that this rectangle is always reported in WGS84 coordinates, irrespective of the spatial reference system (CRS) of input and output datasets. As illustrated in Figure 3, for a transformation example concerning POI data in Greece, this box shows the spatial coverage of the POI features that have been transformed and are included in the resulting RDF file.
- Execution metadata include the following items:
    - **Count of input records** (for structured data) or features (for semi-structured data). This concerns the amount of POIs detected in the original dataset. Note that this value includes every POI feature with a geometry and a distinct identifier in the original input dataset, as specified in the configuration settings of TripleGeo.
    - **Count of output triples** indicates the amount of statements in the resulting RDF file. This file contains a complete RDF graph that can be used in subsequent stages of data integration (e.g., interlinking) and its constituent triples comply with the POI ontology.

- **End-to-end execution time** (in milliseconds) that includes the cost of accessing and fetching the input data, its transformation cost, as well as the cost of writing output to file(s). This serves as an indication about the overall processing cost regarding transformation of the original dataset.

In addition, some informative metadata are also given to keep track of the results across the data integration task as well as the transformation mode employed for the particular POI dataset:

- Path to the output file containing the resulting RDF triples;
- The RDF serialization of the output triples; and
- The transformation mode employed in the execution.


```
▼ Attribute Statistics:
    NAME_ENG:                          72355
    PHONE:                             72373
    ADD_NUMB:                          24034
    DATA_SOURCE:                       72373
    EMAIL:                             9400
    OBJECTID:                          72373
    NAME_GR:                           72373
    CATEGORY_URI:                      72373
    TK:                                42826
    ADDR_GR:                           45553
    FAX:                               12617
    ADDR_ENG:                          21205
    WEBSITE:                           17660
▼ MBR of transformed geometries (WGS84):
    X_max:                             29.596572257377158
    Y_max:                             41.746653712939526
    X_min:                             19.374669876151085
    Y_min:                             34.80414849796334
▼ Execution Metadata:
    Output triple count:               3313133
    Input record count:                72373
    Execution time (ms):               14240
    Output file:                       "./test/output/get-pois_v07.nt"
    Output serialization:              "N-TRIPLES"
    Transformation mode:               "STREAM"
```

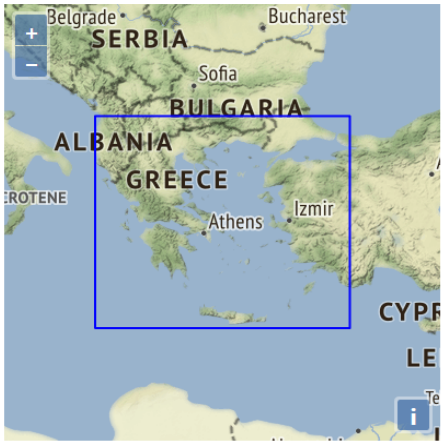

Figure 2: Metadata resulting after transformation to RDF of a POI dataset with TripleGeo.

KPI File : get_pois_full_metadata.json

Name	Value
Attribute Statistics.NAME_ENG	72355
Attribute Statistics.PHONE	72373
Attribute Statistics.LONGITUDE	72373
Attribute Statistics.ADD_NUMB	24034
Attribute Statistics.DATA_SOURCE	72373
Attribute Statistics.EMAIL	9400
Attribute Statistics.OBJECTID	72373
Attribute Statistics.NAME_GR	72373
Attribute Statistics.CATEGORY_URI	72373
Attribute Statistics.TK	42826

Previous Page 1 of 3 Next

Bounding Box



Attribute Statistics

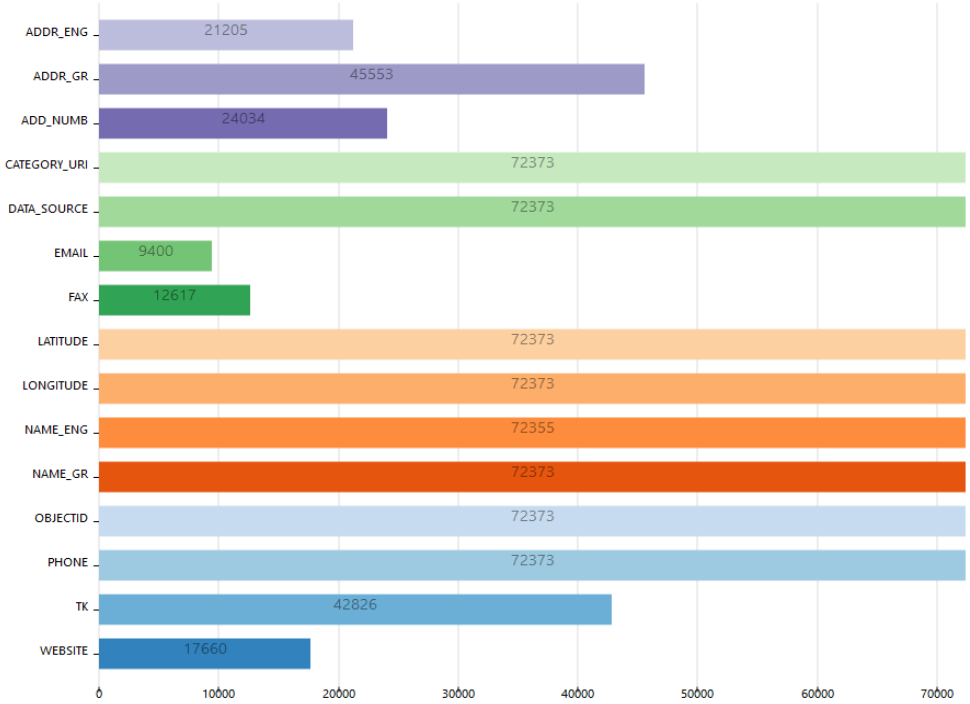


Figure 3: Quality measures visualized in SLIPO Workbench after POI transformation with TripleGeo.

3.2. LIMES

The aim of interlinking POI datasets is to develop scalable approaches for integrating massive heterogeneous, and incomplete POI data at a world-scale. Towards this goal, our state-of-the-art interlinking software **LIMES**[15] offers advanced capabilities for interlinking POI datasets including spatial, topological and temporal relation discovery (as documented in Deliverable D3.1 [27]). In particular, LIMES receives as input two RDF POI datasets conforming to the SLIPO ontology. Thus, LIMES's input POI data are first transformed by TripleGeo into the proper RDF format and schema. Further, apart from the input POI datasets, LIMES requires as input a configuration file containing the LIMES configuration parameters.

The generated links of LIMES are distributed into two files based on the **links' confidence**. The link confidence is a score computed by LIMES as a *quality indicator* for each single link it generates. To this end, the links with high confidence scores are stored into the *accept* file and links with medium confidence scores are stored into the *review* file. The output of LIMES is essential for running other SLIPO tools as shown in the POI data integration cycle (Figure 1). In addition to the two aforementioned output files of LIMES generated links, LIMES produces a JSON document featuring a multitude of *quality metrics* computed based on the *generated links* at the end of the interlinking process, including quantitative and qualitative metrics of the interlinking process. This JSON document can then also be used by the SLIPO Workbench to provide useful visualizations and summaries of these metrics. An exemplary JSON document featuring the quality metrics for an execution of the LIMES configuration developed for the DINUC A2 use case [29] is given in Figure 4 and will be used as a running example.

3.2.1. LIMES Quantitative Metrics

The quantitative metrics are quality metrics regarding the amount of input, output and links sizes and interlinking process runtime. The currently implemented quantitative quality metric in LIMES includes:

- **Mapping Time.** It measures the program runtime in milliseconds. i.e., the time LIMES needs to generate the links as mapping from source to target resources. Depending on the input sizes, a high runtime for a relatively small input data size value might signal that the configuration should be optimized. The mapping time of under two seconds measured in the running example (see Figure 4) can be regarded as very good pertaining to the input sizes of 3,737 and 510.
- **Input sizes.** A count of the triples read by LIMES from the configured data sources. i.e., the size of both the source and target datasets. This quality measure is used to verify that that LIMES

[15] Software publicly available at: https://github.com/SLIPO-EU/LIMES

reads the expected number of resources from the input source and target datasets. The ratio of input size to the run time is an indicator of the efficiency of the linking process.

- **Output sizes.** A count of the links generated by LIMES. In particular, two sets of links are generated: The first set of links are all above the *"accept threshold"* with high confidence. The second set of links is the ones above the "verify threshold" and below the accept threshold with medium confidence. Note that, both accept and verify thresholds are configuration parameters of LIMES. This quality indicator is used to judge the appropriateness of the configured link specification and confidence levels. In our running example (see Figure 4) there are no links with a confidence score under the accept and above the verify threshold but 222 links with a confidence above the accept threshold. The underlying link specification can therefore be regarded as appropriate, as we are interested in discovering the `owl:sameAs` relationship between POIs in the TomTom and OSM datasets from Bucharest and we expect to (1) have at most 510 links and (2) that there are a reasonable number of unique POIs per dataset.

3.2.2. LIMES Qualitative Metrics

The qualitative metrics are indicators for the quality of the links generated by the interlinking process. They are abbreviated PRF which stands for precision (P), recall (R) and F-measure (F). We employ two methods of computing PRF: (1) using a standard definition and (2) using our own extension.

For each of these methods, PRF is computed twice; first for the set of all generated links and then for the subset of links whose confidence is on the "accept" level. In our running example in Figure 4 we can see that all metrics for "*all*" and "*acceptance*" are the same, which does make sense because, as previously stated, there are no links with a confidence score under the accept and above the verify threshold and hence the two sets of links are identical.

3.2.2.1. Standard PRF metrics

These quality metrics measure the quality of generated links against a prespecified gold standard set of links. Note that the user of LIMES needs to specify the gold standard links in order to enable LIMES to include this set of quality metrics in the output. As there exists *no gold standard* for the interlinking in DINUC A2, the running example in Figure 4 does not include these metrics.

- **Precision**. Precision measures the specificity of the link specification as the ratio between true positive links. i.e., found links by LIMES, which are present also in the gold standard. A very specific link specification yields a high precision, while a low precision signals an underspecified link specification.
- **Recall**. Recall measures the sensitivity of the link specification as the ratio between true positive links, i.e., in the gold standards, which are found by LIMES. A very sensitive link specification yields a high recall, while a low recall signals an under-sensitive (and thus possibly over-specified) link specification.

- **F-measure**. F-measure aims to provide an overall measure of the accuracy by combining precision and recall using the harmonic mean.

3.2.2.2. Pseudo-PRF metrics

The pseudo-PRF metrics are extensions to the standard PRF metrics which do not require the presence of a gold standard and thus can be used without additional input from the user. They will be always present in the output JSON file of LIMES. The basic assumption behind these pseudo measures is that symmetrical one-to-one links exist between all the resources in source and target datasets. Note that this assumption does only hold partially for our running example in Figure 4, which explains the low performance in pseudo-recall and pseudo-F-measure.

- **Pseudo-Precision**. Pseudo-precision computes the fraction of links that stand for one-to-one links. The pseudo-precision PP is defined as the fraction of links that stand for one-to-one links.
- **Pseudo-Recall**. The pseudo-recall computes the fraction of the total number of resources (i.e. from both source and target datasets) that are involved in at least one link.
- **Pseudo-F-measure**. The pseudo-F-measure PFM is the harmonic mean of pseudo-precision and pseudo-recall.


```
{
  "mappingTime" : 1430,
  "inputSizes" : {
    "source" : 3737,
    "target" : 510
  },
  "outputSizes" : {
    "verification" : 0,
    "acceptance" : 222
  },
  "pseudoPRF" : {
    "acceptance" : {
      "precision" : 0.768018,
      "recall" : 0.080292,
      "f-measure" : 0.145385
    },
    "all" : {
      "precision" : 0.768018,
      "recall" : 0.080292,
      "f-measure" : 0.145385
    },
  }
}
```


Figure 4: LIMES JSON output for DINUC A2

3.3. FAGI

Fusion consists in receiving two datasets containing POIs and their attributes, as well as a set of links linking POI entities between the two datasets, and producing a third, final dataset, which contains

consolidated descriptions of the linked POIs. The fusion process is executed after the transformation and interlinking steps in the POI data integration lifecycle. The SLIPO toolkit utilizes the FAGI software framework in order to handle the fusion process. FAGI handles effectively the merging of the linked entities, i.e., the production, for each set of linked entities, of a richer, more correct and more complete description of the entity, minimizing manual effort from the user.

An important challenge of this process is the quality assessment of the fusion results. Towards this goal, FAGI supports the extraction of quality indicators and statistics, both in the beginning and the end of the fusion process, in order to be able to validate the quality of the fusion results, as well as compare the quality between different workflows.

The user is able to review several statistics on the input, linked POI datasets, before performing fusion on them (*pre-fusion statistics*), as well as on the output, fused data (*post-fusion statistics*). The goal of the former is to allow the integrator to obtain an overview of the data at hand, which may help the definition and configuration of the validation/fusion rules. The goal of the latter is to assist the user in the examination/validation of the fusion results, and potentially guide a possible re-configuration and re-execution of the fusion process.

Specifically, FAGI produces a JSON file that contains a set of statistics about the individual input datasets, statistics related to the linked POI entities, as well as indicators about the frequency of important attributes of the POIs.

3.3.1. Pre-fusion statistics

The statistics generated on individual input datasets are:

- Number of POI entities in each input dataset.
- Total number of triples in each input dataset, i.e., total number of properties for all POIs.
- Total numbers of empty and non-empty triples in each input dataset.
- Average number of properties per POI in each input dataset.
- Average number of empty and non-empty properties in each input dataset.
- Average number of categories(tags) per POI in each input dataset.
- Total number of POIs that have a specific property in each input dataset.
- Number of empty and non-empty values a specific property in each input dataset.

The statistics related to linked POIs of the input datasets are:

- Ratio of linked POIs to total number of POIs in each input dataset.

- Total number of triples in each input dataset (i.e., total number of properties for all POIs), corresponding only to linked POIs.
- Total numbers of empty and non-empty triples in each input dataset, corresponding only to linked POIs.
- Average number of properties per POI in each input dataset, corresponding only to linked POIs.
- Average number of empty and non-empty properties in each input dataset, corresponding only to linked POIs.
- Total number of POIs that have a specific property in each input dataset, regarding only linked POIs.
- Total numbers of empty and non-empty values for a specific property in each input dataset, regarding only linked POIs.
- Average number of categories(tags) per POI in each input dataset, regarding only linked POIs.
- Number of POI name property values from dataset A that are longer (longer literals) than the names of the corresponding (linked) POIs from dataset B (also the inverse indicator).
- Number of POI phone property values from dataset A that are longer than the names of the corresponding (linked) POIs from dataset B (longer phone strings imply more proper phone format, e.g., containment of full country/exit codes) (also the inverse indicator).
- Number of fully matching address streets between linked POIs in the two datasets.
- Number of fully matching address numbers between linked POIs in the two datasets.

3.3.2. Post-fusion statistics

The statistics generated after the end of the fusion process are:

- Number of fused POIs vs. initial links, i.e., the number of POI links that were not rejected by FAGI and participated in the fusion process.
- Number of rejected POI links vs. initial links, i.e., the number of POI links that were eventually rejected by FAGI.
- Number of fusion actions that: Kept left value; Kept right value; Concatenated left and right value and kept as one; Kept both values as separate properties; Kept longest value
- For each fusion rule that was defined in the fusion specification, the number of times it was executed and produced a fused POI.
- For each link validation rule that was defined in the validation specification, the number of times it was executed.

- Number/percentage of fused POIs that were marked as ambiguous (and thus require further examination/validation) vs. the number of initial links.
- Number/percentage of rejected POIs that were marked as ambiguous (and thus require further examination/validation) vs. the number of initial links.

Additionally, FAGI assigns an interlinking score (*when available from the input*) and calculates a *fusion gain* and a *fusion confidence* metric for the linked POIs that participated in the fusion process. These quality metrics -in contrast to the above indicators- are integrated into the RDF data in order to be exploited in cases of multiple fusion steps inside the same data integration workflow.

- **Fusion-gain.** The number of added attributes in the fused POI. This metric is computed as: $\frac{p-l}{n+m}$,

where *p* is the total number of the resulting fused attributes, *l* is the number of the common attributes between the source POIs and *n, m* represents the number of total attributes of the respective source POIs.

- **Fusion-confidence.** The confidence of the fusion process between a pair of linked POIs. The fusion confidence metric is calculated as the average of the similarities between existing values of the most important POI attributes (name, address, phone, geometry):

$$AVG(sim(name_A, name_b), sim(address_a, address_b), numSim(phone_a, phone_b), geoSim(geo_a, geo_b))$$

where "sim" is a textual similarity, "numSim" is an edit distance similarity of normalized values of phone numbers and "geoSim" is the normalized orthodromic distance by the max distance of the linked POIs.

3.4. DEER

Enrichment is one of the main parts of any data integration process. In SLIPO, enrichment focuses on POI entities that are characterized by a set of major properties (*e.g., name, coordinates and category*) as well as potentially several additional properties (*e.g., address, telephone, email, rating.*). Enrichment considers one or more input dataset(s) containing POIs. The goal of enrichment is to produce one or more enriched dataset(s), containing better descriptions of the input POIs based on information retrieved from external, third-party RDF data sources (e.g., SPARQL endpoints, DBpedia). That is, each POI entity in the final, enriched dataset must be described by a set of RDF triples that have been derived by merging the initial description for the POI with those generated via various enrichment operations. Note that, some enrichment approaches can define a set of triples to be removed from the original POI descriptions. Those removed set of triples are either wrong or inaccurate. The enrichment process can also replace inaccurate triples with ones with correct values. Considering the big picture of the POI

integration lifecycle in Figure 1, the enrichment process is tightly interconnected with validation and quality assurance. To this end, the enrichment process needs to incorporate several mechanisms to assess the quality of the proposed enrichment operations and their results.

In SLIPO, we use our state-of-the-art enrichment software DEER[16] for POI enrichment. DEER incorporates many approaches for performing efficient enrichment among POI resources [28]. DEER receives as input one or more RDF POI dataset(s) conforming to the SLIPO ontology. Thus, DEER's input POI data are first transformed by TripleGeo, into the proper RDF format and schema. Moreover, DEER input datasets, are (or may be) linked via LIMES prior to be enriched by DEER. Further, DEER requires as input a configuration file containing the DEER configuration parameters. DEER's output consists in one or more files containing the enriched versions of the input datasets. For assuring the quality of the enrichment process, we implemented an analytics feature built into DEER's standard output, a separate JSON file containing the quality metrics of all the enrichment operators DEER uses in the enrichment process. In particular, the JSON file includes a set of general quality metrics shared by all enrichment operators, in addition to a set of enrichment-operator-specific quality indicators. We use the enrichment configuration for DINUC A2 to generate the example output of quality metrics in Annex 5.1.

3.4.1. DEER's General Quality Metrics

- **Runtime.** The time DEER took to execute the given enrichment operator in milliseconds. This quality indicator is used to control and improve the overall runtime. In the example we see that most of the enrichment operators have runtimes below one seconds with the exception of (1) the `#fullInput` operator and (2) the `#dereferencing` operator taking about three and six seconds respectively. However, this is expected as (1) is reading over 400,000 triples from disk and (2) is sending multiple queries to DBpedia, retrieving over 20,000 triples in the process.
- **Input sizes.** The number of triples DEER reads from each of the incoming edges of the given enrichment operator in the enrichment graph. Can be used together with output sizes to detect problems in the configuration of the enrichment operator.
- **Output sizes.** The number of triples written to each of the outgoing edges of the given enrichment operator in the enrichment graph. Can be used together with input sizes to detect problems in the configuration of the enrichment operator.

3.4.2. Specific Measurements

Currently, only the *LinkingEnrichmentOperator* implements the specific quality measure of:

- **Number of Discovered Links.** This quality indicator counts the number of links generated by this enrichment operator.

[16] Software publicly available at: https://github.com/SLIPO-EU/DEER

4. Evaluation

In this Section, we evaluate the SLIPO Toolkit components regarding quality in a real-world scenario involving POI datasets from two different sources. These data sources differ in their respective schema, content, volume, and of course quality. We demonstrate how a user can get insights about these features with the particular measures and statistics issued after each stage of a data integration workflow. We first present some basic information about the two input POI datasets, and then discuss the quality indicators and statistics derived from SLIPO components involved in successive steps of the data integration workflow.

4.1. Evaluation Setup

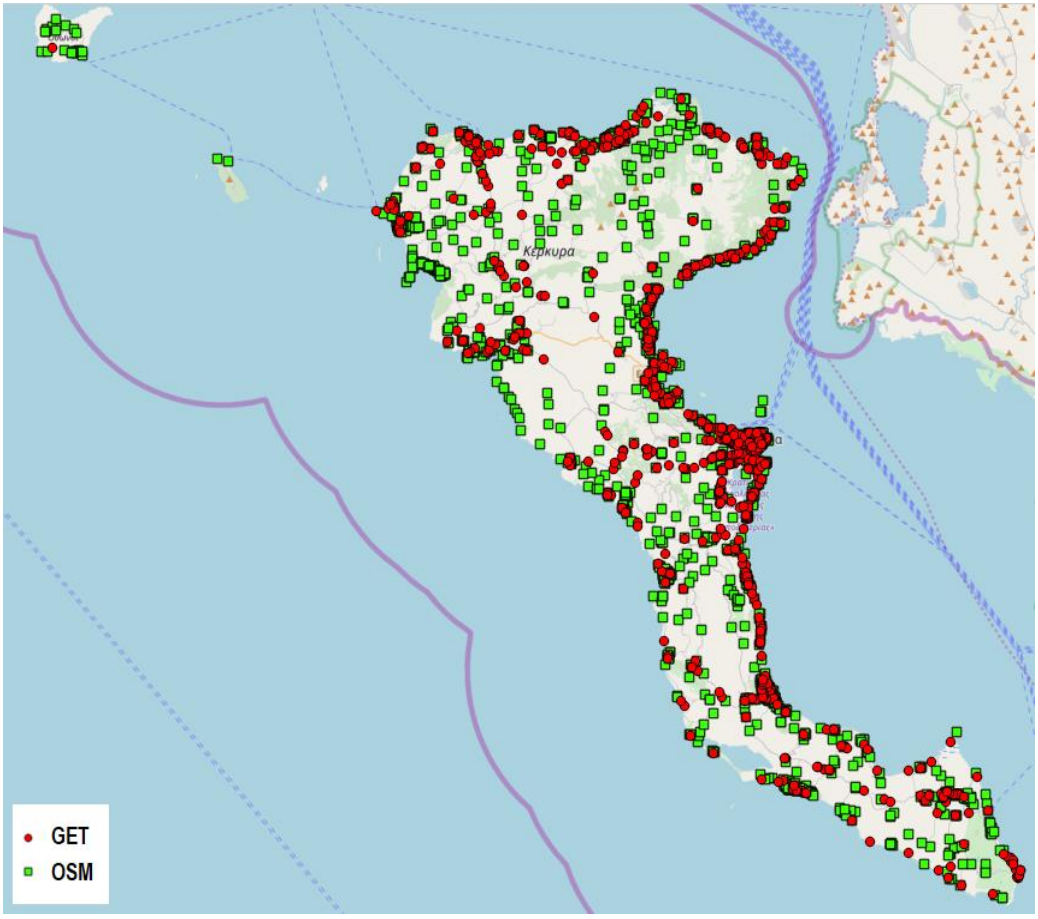


Figure 5: Original POIs from two data sources on Corfu Island, Greece.

In this evaluation scenario, we follow a data integration workflow applied on two different datasets containing POIs on Corfu Island, Greece (Figure 5). The two data sources have different schema, content, and quality, with one being crowdsourced (OSM) and the other offered by a commercial data provider (GET). Through SLIPO, the user can define in a few minutes a simple data integration workflow that delivers a single dataset with more POIs (e.g., missing POIs from OSM), richer geometries (e.g., polygons from OSM), and more accurate information (e.g., updated telephone numbers from OSM). More specifically:

1) *GET dataset*[17], within lists 1043 POIs with their point lon/lat coordinates in WGS1984, name, full address (i.e., street, house number, postal code), as well as basic contact details (phone, email, website). POI names are given in Greek and in English.

2) *OSM dataset*, which is extracted from the OpenStreetMap database[18] and contains 1814 fuel stations. Each record includes full geometry georeferenced in WGS1984, so position of POIs is not only based on lon/lat point coordinates, but it is more detailed for many of them (i.e., the geometry of a POI as a polygon or multipolygon). In addition, many more thematic attributes are available per record: phone and fax numbers, email address, website, opening hours, full address (i.e., street, house number, postal code, town, country), as well as international name and classification of the POI.

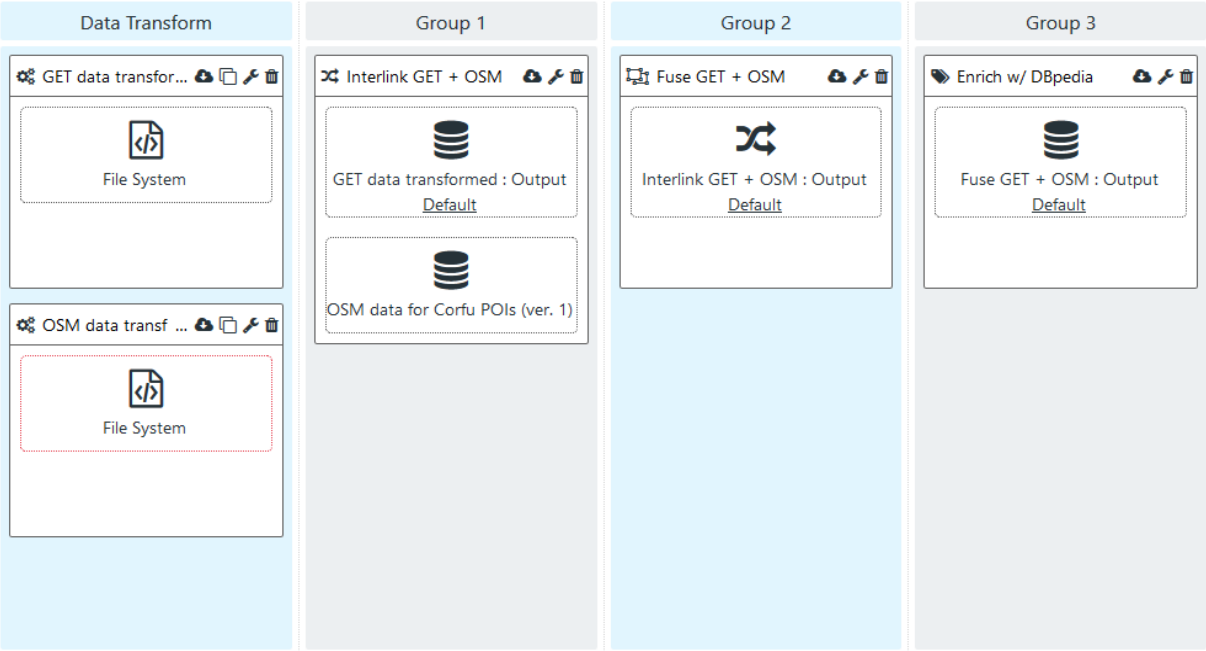


Figure 6: Data integration workflow designed in SLIPO Workbench for integrating two POI datasets in Corfu.

[17] https://www.getmap.eu

[18] https://www.openstreetmap.org

We demonstrate the quality measures available from the various components in the SLIPO Workbench application through a typical data integration workflow. As illustrated in Figure 6, this workflow involves the following stages:

- *Transformation* of the two POI datasets to RDF according to user-specified mappings and parameters. Two instances of TripleGeo are used to perform these transformations.
- *Interlinking* the two POI datasets in order to identify matching POI entities. An instance of LIMES takes in the previously transformed RDF datasets and issues links between them.
- *Fusion* of the two datasets into a new one according to a fusion strategy using FAGI for combining spatial and thematic properties based on the previously identified links between RDF entities.
- *Enrichment* of the fused dataset with additional information extracted from *DBpedia*[19].

Next, we describe each stage and discuss the quality measures obtained after its execution.

4.2. Quality Measures per Stage

4.2.1. Quality Measures after Transformation

As illustrated in Figure 6, this workflow involves two separate instances of TripleGeo to transform each of the two input datasets into RDF. Quite importantly, the properties of each transformation task must be specified, in this case thanks to predefined profiles for most common transformation tasks (e.g., OSM data). As documented in Deliverable D2.2 [26], this concerns all details for the transformation of a given dataset, including the mappings and a classification scheme as required for the proper conversion of input POI data into RDF triples according to the SLIPO ontology for POIs.

Once each transformation task is finished, the resulting RDF triples are written into a file in N-TRIPLES serialization so that they can be used in the subsequent stages. Regarding quality, TripleGeo issues statistics in JSON format regarding performance measurements and quality indications, as discussed in Section 3.1. In this example, the two JSON files for the transformed GET dataset and the transformed OSM dataset are listed in Figure 7. Note that the same results can be intuitively illustrated as Figure 3.

Regarding the transformed GET dataset, it is evident from the results that all POIs indicate their identifiers, names and location (longitude/latitude) coordinates, whereas they are also classified into a specific category. However, address specifications are missing for many of them, as only 405 (out of a total 1043) POIs indicate a street name, and even less (255) a house number as well. Furthermore, only

[19] https://wiki.dbpedia.org/

388 POIs indicate their postal code (TK), and 117 include their address in internationalized (Latin) characters (ADDR_ENG). This means that address specifications are incomplete and cannot properly serve for POI matching in the subsequent interlinking stage. The same holds for the remaining attributes, which mostly concern contact details. Indeed, many POIs have phone information (477 of them) or provide a website (374), but much less indicate a fax number (196) or an email address (171). The fact that so many missing values exist in those thematic attributes underscores the possible advantages from data integration with other sources.

```
Attribute Statistics:
    NAME_ENG:                          1043
    PHONE:                             477
    LONGITUDE:                         1043
    ADD_NUMB:                          255
    DATA_SOURCE:                       1043
    EMAIL:                             171
    OBJECTID:                          1043
    NAME_GR:                           1043
    CATEGORY_URI:                      1043
    TK:                                388
    ADDR_GR:                           405
    FAX:                               196
    LATITUDE:                          1043
    ADDR_ENG:                          117
    WEBSITE:                           374
MBR of transformed geometries (WGS84):
    X_max:                             20.12061473617188
    Y_max:                             39.89663121109356
    X_min:                             19.387085091143003
    Y_min:                             39.36674700884863
Execution Metadata:
    Output triple count:               30162
    Input record count:                1043
    Execution time (ms):               876
    Output file:                       "get_pois_v02_corfu_2100.nt"
    Output serialization:              "N-TRIPLES"
    Transformation mode:               "STREAM"
```

(a) GET

```
Attribute Statistics:
    osm_id:                            1814
    country:                           4
    image:                             4
    website:                           124
    city:                              60
    LONGITUDE:                         1814
    LENGTH:                            421
    postcode:                          269
    DATA_SOURCE:                       1814
    description:                       20
    CATEGORY_URI:                      1814
    int_name:                          47
    alt_name:                          5
    AREA:                              421
    phone:                             116
    housenumber:                       24
    street:                            62
    name:                              1814
    opening_hours:                     55
    wikipedia:                         19
    fax:                               27
    LATITUDE:                          1814
    email:                             57
MBR of transformed geometries (WGS84):
    X_max:                             20.1075647
    Y_max:                             39.8833243
    X_min:                             19.378324
    Y_min:                             39.3641665
Execution Metadata:
    Output triple count:               30447
    Input record count:                1814
    Execution time (ms):               1633
    Output file:                       "osm20_pois_corfu.nt"
    Output serialization:              "N-TRIPLES"
    Transformation mode:               "STREAM"
```

(b) OSM

Figure 7: Quality indications regarding transformation of two POI datasets using TripleGeo.

Regarding the transformed OSM dataset, the situation is different. First, the amount of POIs available in this dataset is considerably larger (1814 POIs). All of them indicate their identifier, name, and location coordinates. However, 421 of them have detailed (polygon) geometries, as indicated by the derived attributes concerning their perimeter (LENGTH) and area (AREA). Contact information (phone, fax, email, website) is poorer compared to the GET data, and the same holds for address specifications (street name, house number, postal code, etc.). Given that this is a crowdsourced dataset, this should be expected, as OSM data is not particularly rich in Greece compared to other countries in Europe (e.g., Germany, Sweden). Still, this OSM dataset provides some extra information in additional thematic attributes concerning opening hours, Wikipedia links, pointers to image files, and detailed descriptions for POIs, which are entirely missing from the other dataset (GET). Even though this is available for rather few POIs, extracting such information would be certainly advantageous for data integration.

Overall, these quality results hint that interlinking should be based primarily on geometry proximity and name similarity, as those properties exist in all POIs. Using other properties (e.g., phone numbers) for POI matching has risks, as too many missing values indicate that links may not be identified. However, those thematic properties (address components, contact details) can be used for filtering the links and validating potential fusion actions between POIs during later stages, most notably during fusion with FAGI, as will be discussed in Section 4.4.


```
{
  "mappingTime" : 1939,
  "inputSizes" : {
    "source" : 1043,
    "target" : 1814
  },
  "outputSizes" : {
    "verification" : 827,
    "acceptance" : 218
  },
  "pseudoPRF" : {
    "acceptance" : {
      "precision" : 0.752294,
      "recall" : 0.114806,
      "f-measure" : 0.199210
    },
    "all" : {
      "precision" : 0.328230,
      "recall" : 0.240112,
      "f-measure" : 0.277340
    },
  }
}
```


Figure 8: LIMES quality metrics.

4.3. Quality Measures after Interlinking

After the transformation of both datasets to RDF, interlinking is applied in order to get links between matching POIs from the GET and OSM datasets. As these will subsequently form the basis for the fusion process, it is important to assure the quality of the generated links.

In Figure 8, we can see the quality metrics produced by LIMES after interlinking as discussed in Section 3.2. The quantitative analysis shows that the interlinking task was finished within 2 secs for input sizes of 1,043 x 1,814 triples, which is sufficiently fast. A total of 1,045 links were found of which 218 were above the "accept" confidence level.

As there is no gold-standard dataset for this particular task, we rely on the pseudo-PRF metrics for the qualitative analysis. Recall that precision reflects the portion of the generated links which can be regarded as 1-to-1 correspondences, i.e., how well they approximate an identity relation. We want to identify 1-to-1 correspondences between POIs to be used for fusion and thus the precision is our primary quality metric for this task. As we can see, the accepted links (i.e., the ones with confidence above the "accept" threshold) achieve a precision of 0.75. Comparing this with the precision over all links of 0.33, we can conclude that most links below the "accept" confidence level would require a close manual inspection while the links above the "accept" confidence level are good enough to be fed directly into fusion without the need for the manual control.

4.4. Quality Measures during Fusion

Following the interlinking process, the user can define rules and set the fusion configuration by getting some insights from statistics produced by FAGI.

4.4.1. Pre-fusion exploration

As described in Section 3.3, the user is able to explore several statistics on the input POI datasets, but most importantly in conjunction with the interlinking output.

As we see in Figure 9, the total POIs are much more than the linked, which means that the two datasets *do not overlap* to a great degree. In order to confirm the interlinking behavior, we take a look at the output file of the category frequencies produced by FAGI. By looking at Table 2, we can verify that most POIs belong to different main categories. Specifically, most POIs from OSM are assigned to "eat/drink", "settlements" and "shop" category and most POIs from GET contain mainly "tourism", "health" and "business" categories, with the "accommodation" being the main common category that most pairs are expected to be identified as same.

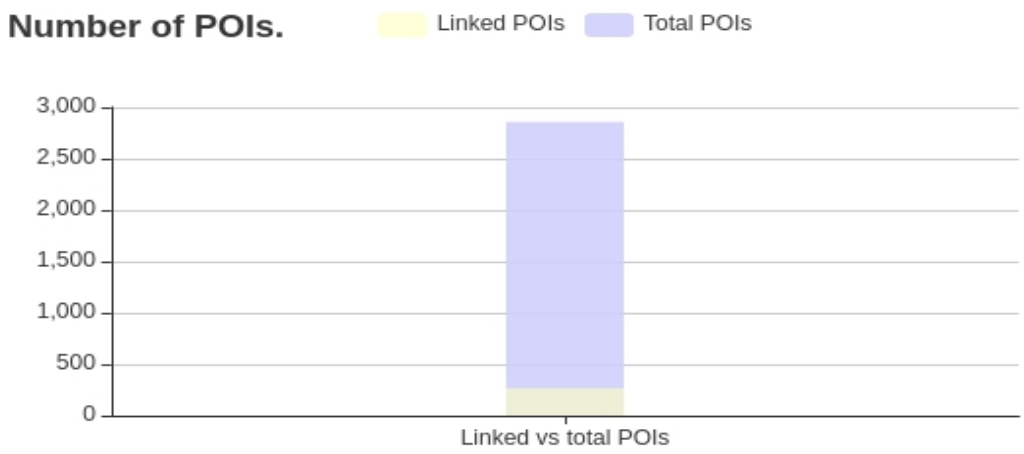


Figure 9: Linked POIs against the total POIs from both datasets.

As a result, a fusion strategy that combines all POIs to a composite dataset seems more appropriate. In this way, the fused dataset, except from containing richer content regarding geometries, extra thematic attributes and filled missing values on the linked POIs, will also greatly increase regarding the number of POIs it contains.

In order to define suitable fusion rules on the linked POIs, we explore some basic attributes that concern particularly the linked POIs between the two datasets (Figure 10). As we see, the POIs from GET have more values to offer regarding these attributes (phone number, website, e-mail, address) and we consider these values more accurate than the crowdsourced OSM dataset. As a conclusion, we would like to retain the values of the basic attributes from GET (and only fill missing values from OSM) and keep all other extra attributes (like opening hours, image links, *wikipedia* links etc.) from the OSM dataset.

4.4.2. Post-fusion statistics and quality metrics

After deciding on the rule specification, we are ready to execute the fusion process. During this step, some of the links are getting rejected (28 links in particular) based on validation rules that also take into consideration more properties (phone, address) in the similarity check. This process ends by keeping only the links that were deemed reliable. As we see in Figure 11, the fused dataset contains more than 2700 POIs, which is almost 3 times the size of the first dataset. Regarding the fused properties we can see at Figure 12 the number of properties resulted in the fused dataset, which exceeds the number of properties of each of the input datasets.

Category Name	GET	OSM
accommodation	260	276
automotive	85	56
business	93	9
eat/drink	64	560
education	44	32
health	107	25
landuse	7	16
public	84	9
religious	67	76
service	84	9
settlements	-	253
shop	88	211
sport	16	30
tourism	118	149
transport	10	83

Table 2: POI category frequencies of input datasets

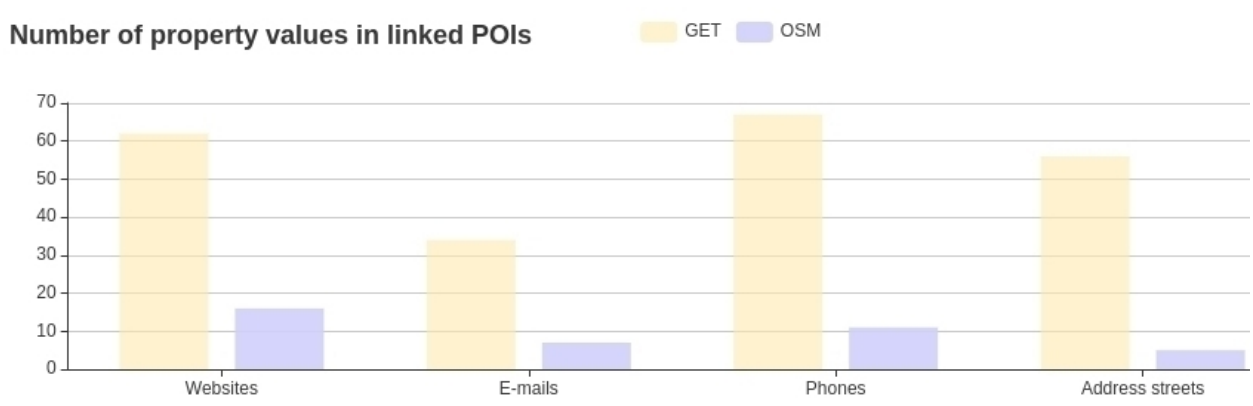


Figure 10: Basic attributes of linked POIs of the two datasets

In addition to the post-fusion statistics, FAGI outputs the two quality measures of fusion gain and fusion confidence as explained in Section 3.1**Error! Reference source not found**.. Since these measures are calculated for each fused pair of POIs, the final output of FAGI indicates an average fusion confidence, as well as an average and maximum fusion gain. In this scenario, the average confidence is 0.73, which is acceptable, considering our relaxed criteria for linking and the high heterogeneous datasets. Regarding attribute gain, integrated results denote an average 18% increase in properties per POI and in some cases we see an attribute increase over 30%, which is noteworthy, since the crowdsourced content in OSM for this area is not very rich. Still, the most important outcome of this integration is that the final dataset includes *more than 2700 POIs, far more than the initial ~1000 POIs of GET*.

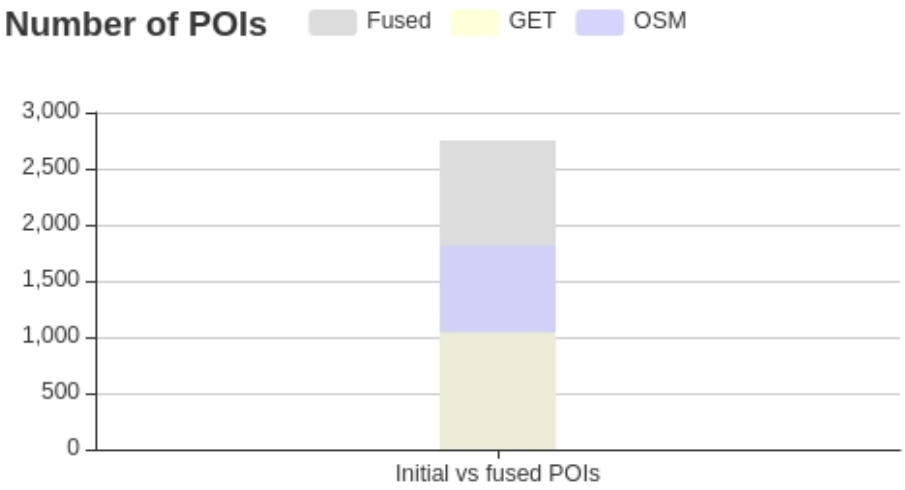


Figure 11: POIs in fused dataset vs input POIs

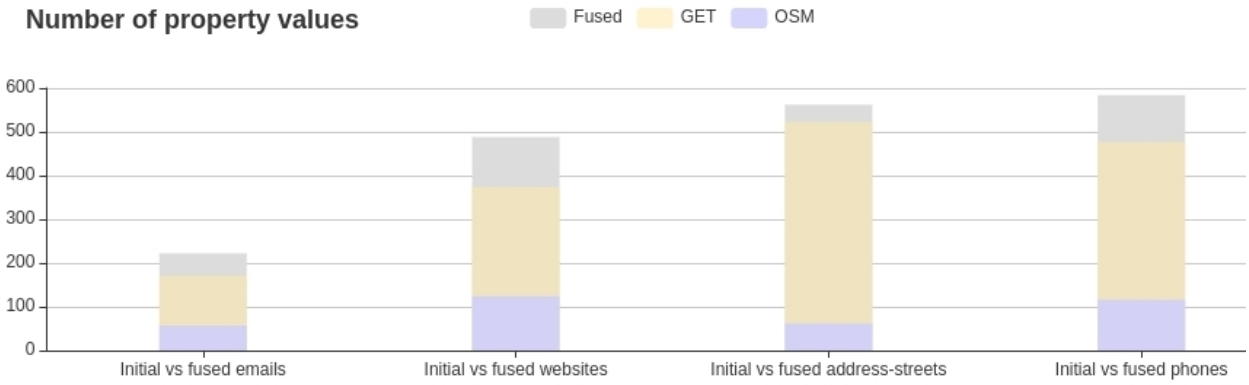


Figure 12: Number of basic property values in fused vs input datasets

4.5. Quality Measures after Enrichment

The goal of enrichment for this task is to increase the attractivity of POIs for tourism. To this end, we use DBpedia as external source to enrich each POI with a link to the nearest Corfu landmark as well as its distance from the POI in kilometers.

We present the output of DEER analytics (introduced in Section 3.4) in Annex 5.2, where we can see that we successfully found 558 connections between POIs in the fused input dataset and Corfu landmarks from DBpedia. The whole enrichment is done in under 3 seconds. It has to be noted that currently, DBpedia does only include 4 landmarks for the municipality of Corfu. However, this is just an issue of DBpedia not DEER. More landmarks could be found either using future releases of DBpedia with more landmarks in Corfu or by using other external data sources other than DBpedia.

References

[1] James Abello, Panos M Pardalos, and Mauricio GC Resende. Handbook of massive data sets, volume 4. Springer, 2002.

[2] Parameter Barbara. Data quality for GIS. In Report Tuft University, 2007.

[3] Enrico Bertini, Andrada Tatu, and Daniel Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. Visualization and Computer Graphics, IEEE Transactions on, 17(12):2203– 2212, 2011.

[4] Anna T Boin and Gary J Hunter. What communicates quality to the spatial data consumer. In Proceedings of the 7th international symposium on spatial data quality (ISSDQ 2007), Enschede, The Netherlands, 2007.

[5] M. Caprioli, A. Scognamiglio, G Strisciuglio, and E Tarantino. Rules and standards for spatial data quality in GIS environments. In Proc. 21st Int. Cartographic Conf. Durban, South Africa 10–16 August 2003, 2003.

[6] M. Caprioli and E. Tarantino. Standards and quality in gis contexts. Multidimensional approaches and new concepts in SIM, Paris, 2003.

[7] Didier Cherix, Ricardo Usbeck, Andreas Both, and Jens Lehmann. Crocus: Cluster-based ontology data cleansing. In Proceedings of the 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice, 2014.

[8] Rodolphe Devillers, Yvan Bédard, and Robert Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within gis. Photogrammetric Engineering & Remote Sensing, 71(2):205–215, 2005.

[9] Rodolphe Devillers, Marc Gervais, Yvan Bédard, and Robert Jeansoulin. Spatial data quality: from metadata to quality indicators and contextual end-user manual. In OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management, pages 21–22, 2002.

[10] Edward. Spatial data quality and transportation applications. In 5th FIG Regional Conference. 2006.

[11] GIS ESRI's. Implementing iso data quality standards using esri's gis data reviewer. 2004.

[12] Martin Ester, Hans-Peter Kriegel, Jo¨rg Sander and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, volume 96, pages 226–231, 1996.

[13] Michael F Goodchild and Keith C Clarke. Data quality in massive data sets. In Handbook of massive data sets, pages 643–659. Springer, 2002.

[14] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. LUBM: A benchmark for OWL knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web, 3(2–3):158 – 182, 2005.

[15] Shirlee Knight and Janice Burn. Developing a framework for assessing information quality on the world wide web. Informing Science, 8, 2005.

[16] Weise Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. Journal of Visual Communication and Image Representation, 22(4):297–312, 2011.

[17] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: based on eye-tracking data. Circuits and Systems for Video Technology, IEEE Transactions on, 21(7):971–982, 2011.

[18] Peter Mooney, Padraig Corcoran, and Adam C Winstanley. Towards quality metrics for openstreetmap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 514–517. ACM, 2010.

[19] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. Communications of the ACM, 45(4):211–218, 2002.s

[20] Bastian Quilitz and Ulf Leser. 5021:524–538, 2008.

[21] Patrick Stickler. Cbd-concise bounded description. W3C Member Submission, 3, 2005.

[22] Howard Veregin. Data quality parameters. Geographical information systems, 1:177–189, 1999.

[23] Jing Feng Xia. Metrics to measure open geospatial data quality in. Issues Sci. Technol. Librar, 2012.

[24] Songyun Duan, Anastasios Kementsietsidis, Kavitha Srinivas, and Octavian Udrea. Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pages 145–156. ACM, 2011.

[25] SLIPO Project. Deliverable D2.1: POI Models and Identifiers Registry.

[26] SLIPO Project. Deliverable D2.2: Mapping specification and POI transformation service.

[27] SLIPO Project. Deliverable D3.1: Initial release of the POI interlinking software.

[28] SLIPO Project. Deliverable D3.2: Initial release of the POI enrichment framework.

[29] SLIPO Project. Deliverable D5.1: Specification of Pilots

5. Annex

5.1. DEER Analytics Output Example

```
http://deer.aksw.org/vocabulary/#dbp_buildings:
run time:        707ms
output sizes:    ( 196 )

http://deer.aksw.org/vocabulary/#dbp_malls:

run time:        706ms
output sizes:    ( 43 )

http://deer.aksw.org/vocabulary/#fullInput:
run time:        3070ms
output sizes:    ( 406163 )

http://deer.aksw.org/vocabulary/#malls:
input sizes:     ( 406163 )
run time:        413ms
output sizes:    ( 256 )

http://deer.aksw.org/vocabulary/#linkMalls:
input sizes:          ( 256 43 )
#discovered links:    50
run time:             178ms
output sizes:         ( 349 )

http://deer.aksw.org/vocabulary/#non_malls:
input sizes:     ( 406163 )
run time:        827ms
output sizes:    ( 29559 )

http://deer.aksw.org/vocabulary/#linkNonMalls:
input sizes:          ( 29559 43 )
#discovered links:    2625
run time:             225ms
output sizes:         ( 32227 )

http://deer.aksw.org/vocabulary/#distanceToMalls:
input sizes:     ( 32227 )
run time:        38ms
output sizes:    ( 34852 )

http://deer.aksw.org/vocabulary/#mergeLinks:
input sizes:     ( 349 34852 )
run time:        47ms
output sizes:    ( 34989 )
```

```
http://deer.aksw.org/vocabulary/#linkBuildings:
input sizes:          ( 34989 196 )
#discovered links:    3792
run time:             238ms
output sizes:         ( 38977 )

http://deer.aksw.org/vocabulary/#distanceToBuildings:
input sizes:     ( 38977 )
run time:        23ms
output sizes:    ( 42678 )

http://deer.aksw.org/vocabulary/#dereferencing:
input sizes:     ( 42678 )
run time:        6336ms
output sizes:    ( 66956 )

http://deer.aksw.org/vocabulary/#mergeFull:
input sizes:     ( 406163 66956 )
run time:        590ms
output sizes:    ( 461321 )

http://deer.aksw.org/vocabulary/#output_node:
input sizes:    ( 461321 )
run time:       301ms
```

5.2. DEER Corfu Analytics Output

```
http://deer.aksw.org/vocabulary/#dbp_places:
run time:        409ms
output sizes:    ( 31 )


http://deer.aksw.org/vocabulary/#fullInput:
run time:        828ms
output sizes:    ( 77212 )


http://deer.aksw.org/vocabulary/#filter:
input sizes:     ( 77212 )
run time:        236ms
output sizes:    ( 14577 )


http://deer.aksw.org/vocabulary/#linkPlaces:
input sizes:          ( 14577 31 )
#discovered links:    558
run time:             281ms
output sizes:         ( 15166 )


http://deer.aksw.org/vocabulary/#distanceToPlaces:
input sizes:     ( 15166 )
run time:        14ms
output sizes:    ( 15724 )
```

```
http://deer.aksw.org/vocabulary/#mergeFull:
input sizes:     ( 77212 15724 )
run time:        108ms
output sizes:    ( 87434 )


http://deer.aksw.org/vocabulary/#output_node:
input sizes:    ( 87434 )
run time:       142ms
```