DELIVERABLE D3.6

Final 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 36, 31/12/2019
Actual Submission Date	31/12/2019
Work Package	WP3 - Data Integration and Quality Assurance
Tasks	3.5 Quality Assurance
Type	Report
Approval Status	Submitted for approval
Version	0.7
Number of Pages	49
Filename	D3.6_Final_Report_on_Quality_Assurance_Metrics.pdf

Abstract

This report presents the final 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 assessment of the proposed quality assurance metrics on a big data integration scenario as executed in the SLIPO Workbench.

History

version	date	reason	revised by
0.1	20/05/2019	Deliverable structure and initial content	Kevin Dreßler
0.2	10/09/2019	Updates in all sections	Kostas Patroumpas, Abdullah Fathi Ahmed, Pauline Baudens, Kai Barenscher, Thodoris Vakkas
0.3	01/10/2019	Updated section 2	Kevin Dreßler
0.4	07/10/2019	Updated section 3.4 (DEER)	Kevin Dreßler
0.5	21/11/2019	Updated structure of the document	Kevin Dreßler, Kostas Patroumpas
0.6	10/12/2019	Included paragraphs on optimization	Mohamed Ahmed Sherif
0.7	20/12/2019	Included evaluation section	Kostas Patroumpas, Spiros Athanasiou
0.8	30/12/2019	Final additions and editing	Kevin Dreßler
1.0	31/12/2019	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
INFAI	Abdullah Fathi Ahmed	ahmed@informatik.uni-leipzig.de
INFAI	Mohamed Ahmed Sherif	sherif@informatik.uni-leipzig.de
INFAI	Kevin Dreßler	dressler@informatik.uni-leipzig.de
TomTom	Pauline Baudens	Pauline.Baudens@tomtom.com
WIGeoGIS	Kai Barenscher	kb@wigeogis.com
GET	Thodoris Vakkas	tvakkas@getmap.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 final implementation of quality assurance metrics, indicators and processes for POI data integration. We begin by giving an overview of the related work in this area. Afterwards, we present the data quality assurance metrics we chose and included 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 will critically assess the performance and utility of our quality metrics on the example of a large-scale integration process.

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
EPSG	European Petroleum Survey Group
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

In this section, we first discuss the broad goals of POI data quality assurance in general, before we introduce the specific measures, indicators and processes deployed within the SLIPO project in the following section.

As has been discussed in [4], POI data quality is of paramount importance for industrial use cases.

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 [10]. 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 (POIs) are especially prone to errors because they contain data from multiple providers and use different assumptions about structure and semantics of data [9]. 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 [9]. In recent years, the concern for POI data quality has increased due to a number of factors including [7]:

- 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.

When dealing with big amounts of POI data, there should also be a concern for efficient computation of quality assurance metrics, which we incorporated as one of the original goals for Task 3.5. In particular, *"we intended to develop RDF-specific sampling approaches for big POI data sets that will allow selecting characteristic subsets for which quality measures can be computed efficiently while remaining good approximations of the real quality core of the original data set."*[1] RDF sampling means to pick an appropriate, i.e., relevant [26], subset of the RDF graph. However, as we already focused on runtime optimization in selecting and implementing the POI quality assurance metrics used in our POI-integration tools (i.e., LIMES, DEER and FAGI), we found that the computational overhead of applying

[1] From the SLIPO project proposal.

any of the state-of-the-art RDF sampling algorithms [27-31] outweighs the speed-up that applying our metrics on sampled RDF graphs would offer. Therefore, we decided to not implement any sampling approaches contrary to our initial intent. As a consequence of this, the quality assurance metrics that we report represent the data quality *exactly*, while using a sampling approach would only yield approximations of POI data quality metrics.

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[2] provides a series of standards that deal with various aspects of POI-related geospatial data quality. In particular, **ISO 19115-1:2014**[3], **ISO 19113:2002**[4], **ISO 19114:2003**[5] and the technical specification **ISO/TS 19138**[6] 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.

---

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

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

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

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

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

The Spatial Data Transfer Standard (SDTS7), 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**[8]), the European Committee for Standardization Technical Committee 287 (**CEN/TC 287**[9]), Canadian Geospatial Data Infrastructure (**CGDI**[10]), National Spatial Data Infrastructure (**NSDI**[11]), Digital Geographic Information Exchange Standards **(DIGEST**[12]), and Spatial Archiving and Interchange Format (**SAIF**[13]).

2.2. Geospatial Data Quality Scientific Contributions

The work in [14] explores measurement standards on the quality of open Geospatial data. The main purpose of [14] 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. [12] 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. [13] provides a detailed discussion of various spatial data quality components. Boin et al. [3] 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) [6] allows the management of geospatial data quality and the communication of the quality information using indicators that can be analyzed at different levels of

---

[7] http://www.fgdc.gov/metadata/csdgm/02.html

[8] http://www.opengeospatial.org/

[9] http://www.centc287.eu/index.php/education-a-training/education-a-training-centc-287

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

[11] http://www.fgdc.gov/nsdi/nsdi.html

[12] https://www.dgiwg.org/digest/

[13] 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) [5] 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. [8] discusses the implementation of various ISO spatial Data Quality standards. Mooney et al. [11] 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 contributions [3, 16, 17] discuss the various spatial Data Quality metrics and their implementations.

In the year since the first report on quality assurance metrics [20], more recent academic work related to geospatial data quality has been published. In particular, Talhofer et al. [23] suggest multiple fine grained models for multi-criteria analysis of geospatial data quality that are specifically useful for military command and control systems. Anderson et al. [24] survey the applicability of a set of extrinsic and intrinsic geospatial quality measures for peer-produced data during disasters. [25] suggests a general semantic quality analysis framework to encode the differing and specific needs of OpenStreetMap (OSM) (sub)communities, such as the Wheelmap[14].

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.

[14] https://wheelmap.org

- **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 recent 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, 32, 33, 34]
1	Consistency	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2, 32, 33, 34]
1	Completeness	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11 ,2, 32, 33, 34]
2	Lineage	ISO/TC 211, SDTS, [13, 4, 5, 8, 11, 34]
3	Resolution	ISO/TC 211, [19,2,5,32]
4	Currency	[13, 4, 2, 34]
5	Integrity/Believability	[19, 23, 33]
6	Accessibility	[19, 23]
6	Discoverability	[19, 23]
7	Validity	[19, 23]
8	Objectivity	[12]

Rank	Metric	Citation
8	Relevancy	[12]
8	Security	[12]
8	Sufficiency	[12]
8	Compatibility	[12]
8	Legibility	[12]
8	Repurposing	[12]
8	Transparency	[12]
8	Visualization	[12]
8	Value-Added	[12]
8	Reputation	[12]
8	Verifiability	[12]

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[15] 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.

[15] http://geoknow.eu

- **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.
- **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 [15] 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.** [15] 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 [15]) as the weighted of the two previous measures of class coverage and weighted class coverage.

3. SLIPO Quality Assurance 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.

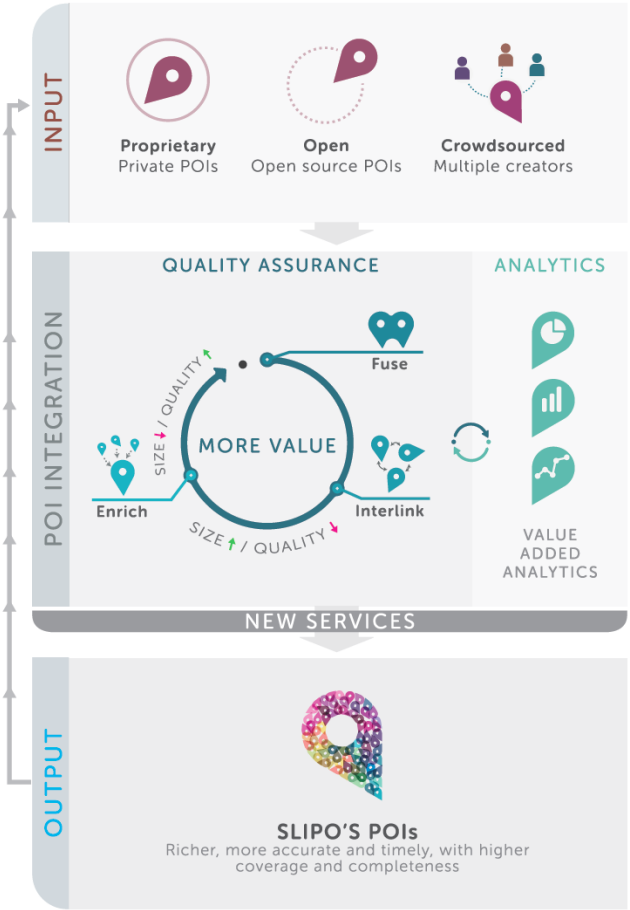

Figure 1: The POI data integration lifecycle.

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**[16] 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 D2.4 [17].

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

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

- *Extract* spatial data from a source;
- *Transform* this data into RDF triples according to a geospatial vocabulary (e.g., GeoSPARQL);
- *Load* resulting triples into a target RDF store.

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 [16], note that the number of resulting triples may be inflated depending on the specified mappings [17]. For the transformation example shown in Figure 3, note that the value counts per attribute differ, since NOT NULL values are only 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 available in the resulting RDF file. In case the user suspects that the spatial extent is not correct, she may check again certain properties of the input data (e.g., CRS, geometry attribute) and the configuration settings used for TripleGeo.
- **Execution metadata** include the following items:

- **Input record count** indicates the total number of *records* (for *structured* data) or *features* (for *semi-structured* data), so it 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 input records excluded.** Some input POI features may not be transformed into RDF either because of filters specified by the user (e.g., a *spatial* area of interest or a specific kind of POIs based on *thematic* criteria such as their category) or due to errors (e.g., invalid geometries). So, this integer value represents the total number of POIs excluded from transformation.
- **Input records transformed** indicates the number of POIs that have been transformed into RDF. If the input dataset contains N POIs, and E of them have been excluded from transformation due to filtering criteria or errors, then the number of transformed POIs is $T = N - E$.
- **Output CRS** indicates the EPSG code of Coordinate Reference System (CRS) of all geometries in the resulting RDF graph.
- **Output triple count** 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.
- **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 end-to-end execution time serves as an indication about the overall processing cost regarding transformation of the original dataset.

If users find that less features or attributes have been transformed, they can check again the configuration settings or attribute mappings used during transformation to RDF and possibly change them before repeating the transformation task.

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 (GRAPH, STREAM, RML, REVERSE).

```
▼ Attribute Statistics:
    ADDR_ENG:                          21205
    ADDR_GR:                           45553
    ADD_NUMB:                          24034
    ASSIGNED_CATEGORY:                 72373
    CATEGORY_URI:                      72373
    DATA_SOURCE:                       72373
    EMAIL:                             9400
    FAX:                               12617
    LATITUDE:                          72373
    LONGITUDE:                         72373
    NAME_ENG:                          72355
    NAME_GR:                           72373
    OBJECTID:                          72373
    PHONE:                             72373
    TK:                                42826
    TRANSLIT:                          72373
    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:
    Input records transformed:         72373
    Output triple count:               2589403
    Input record count:                72373
    Execution time (ms):               162202
    Output CRS:                        "EPSG:4326"
    Output file:                       "/var/local/triplegeo/output/get-pois_v07.nt"
    Output serialization:              "N-TRIPLES"
    Input records excluded:            0
    Transformation mode:               "STREAM"
```

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

KPI File : get_pois_full_metadata.json


```
"metadata" : { 3 items
  "Attribute Statistics" : {...} 17 items
  "MBR of transformed geometries (WGS84)" : {...} 4 items
  "Execution Metadata" : { 9 items
    "Input records transformed" : 72373
    "Output triple count" : 2589403
    "Input record count" : 72373
    "Execution time (ms)" : 162202
    "Output CRS" : "EPSG:4326"
    "Output file" : "/var/local/triplegeo/output/get-pois_v07.nt"
    "Output serialization" : "N-TRIPLES"
    "Input records excluded" : 0
    "Transformation mode" : "STREAM"
  }
}
```


Bounding Box

Attribute Statistics

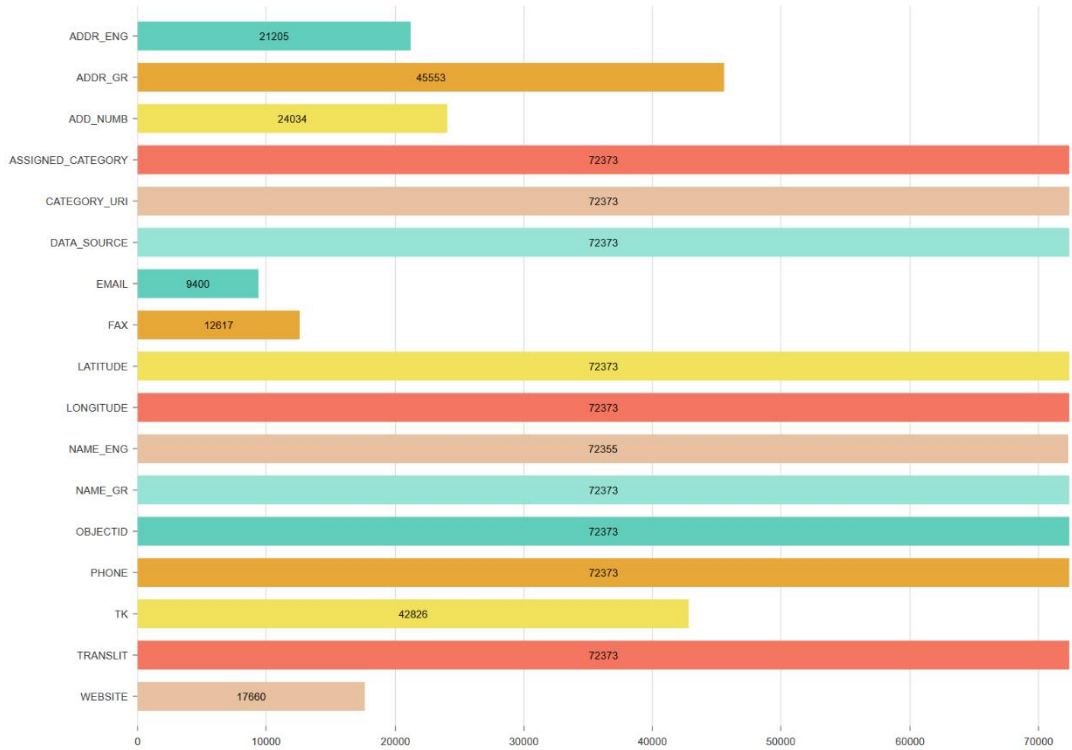

Figure 3: Quality measures visualized in SLIPO Workbench after POI transformation with TripleGeo

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**[17] offers advanced capabilities for interlinking POI datasets including spatial, topological and temporal relation discovery (as documented in Deliverable D3.1 [18]). 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 [21] 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) however is not indicative for further optimization.
- **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 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.

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

- **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.

Obviously, all of the quantitative metrics are easy to compute, as they only involve computing a time interval and counting quantities.

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. The implementation of the precision involves computing set intersections and counting quantities, which can be solved efficiently using hash tables and require no additional optimization.
- **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. The implementation of the recall involves computing set intersections and counting quantities, which can be solved efficiently using hash tables and require no additional optimization.

- **F-measure**. F-measure aims to provide an overall measure of the accuracy by combining precision and recall using the harmonic mean. The F-measure is implemented as the harmonic mean of precision and recall and therefore also does not require any additional optimizations.

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. The implementation of the pseudo precision involves computing set intersections and counting quantities, which can be solved efficiently using hash tables and require no additional optimization.
- **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. The implementation of the pseudo recall involves computing set intersections and counting quantities, which can be solved efficiently using hash tables and require no additional optimization.
- **Pseudo-F-measure**. The pseudo-F-measure PFM is the harmonic mean of pseudo-precision and pseudo-recall. The Pseudo-F-measure is implemented as the harmonic mean of precision and recall and therefore also does not require any additional optimizations.

```
{
  "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.

3.3.3. Fusion Quality Metrics

FAGI retains in its output the interlinking score (*when available from the input*) and also calculates two additional metrics, namely an *attribute gain* and a *fusion confidence* for the linked POIs that participated in the fusion process. These quality metrics -in contrast to the aforementioned statistics- are integrated into the RDF data in order to be exploited in cases of multiple fusion steps inside the same data integration workflow. In particular, for each fused POI, FAGI calculates:

- **Attribute gain** indicates 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, whereas n, m represent the number of total attributes of the respective

source POIs. For instance, an attribute gain of 0.4 for a given fused POI means that it was complemented with 40% additional attribute values.

- **Fusion confidence** indicates 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. Confidence values close to 1 indicate almost perfect match, so the user is informed that such fused results are more trustworthy than others with lower confidence.

In addition, FAGI issues *aggregated quality* estimates about the entire fused dataset:

- **Average attribute gain.** This is the average *attribute gain* calculated from all the POIs that participated in the fusion process.
- **Maximum attribute gain.** This indicates the maximum attribute gain observed among all POIs in the resulting fused dataset. Although this estimate refers to a single POI and may diverge from the average attribute gain, it still gives an idea of the best result obtained by the fusion process.
- **Average fusion confidence.** This is the average *fusion confidence* calculated from all the POIs that were fused during the whole process.
- These indicators are utilized both internally in FAGI (similarity measures, learning mechanisms) and as output for the end user, for further inspection and manual validation of the fused results.

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[18] for POI enrichment. DEER incorporates many approaches for performing efficient enrichment among POI resources [19]. 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. We improved the analytics feature introduced in [20], which now outputs a separate JSON file containing (1) SLIPO-specific POI quality metrics of all the enrichment operators DEER uses in the enrichment process and (2) global SLIPO-specific POI quality metrics. Moreover, this quality assurance report contains a set of general, enrichment-operator-specific quality indicators. In order to separate the SLIPO-specific QA measures from the core module of DEER, we implemented a *DEER SLIPO Analytics Plugin*.

3.4.1. SLIPO-Specific Quality Metrics

Our focus in SLIPO lies clearly on POI data while DEER is a general-purpose tool for linked data enrichment. To bridge this gap, we implemented SLIPO-specific quality assurance metrics as a plugin to the core DEER application.

This plugin computes POI-centric quality metrics and constitutes a major improvement in utility and applicability compared to our previous general approach as presented in the first version of this report [20]. While contributing configurations to the SLIPO pilots [21], it became clear that the previously identified metrics were (1) too general and (2) hard to interpret for novice users. We therefore aimed to move towards metrics that focus more on the actual value of the enrichment process coordinated by DEER from the POI perspective.

- **Global Metrics.** The global metrics are computed with respect to a global input node and a global output node. These correspond to the input file that DEER receives from FAGI within the SLIPO setting and the enriched main output file which DEER generates as its end result.
    - **Total Number of POIs.** A typical configuration of DEER within SLIPO will aim to improve data quality and utility by enriching *existing POIs* with external data which is then conformed to the SLIPO ontology. While it would also lie within the possibilities of the DEER software, the addition of *new POIs* can be solved much better using the fusion capabilities of FAGI. Therefore, we can assume that the **total number of POIs** will be constant within the scope of the SLIPO system. Rather than to convey a notion of POI data quality on its own, we use the total number of POIs as an easily understood,

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

intuitive reference for the other global POI-centric metrics. This metric is implemented using a filtering iterator and counting with Apache Jena, runs in O(n) and therefore demands no further optimization.

- **Number of Enriched POIs.** While we can fine-tune each of the enrichment operators in DEERs configuration to get the most out of the enrichment process, DEER will typically not be able to enrich all of the source datasets POIs due to (1) lack of information coverage from our external sources and (2) lack of 100% accuracy w.r.t. the disambiguation of same-as relationships when similar POIs are densely packed within a small area, as can easily be the case in urban areas. Note that these pitfalls can influence each other, leading to suboptimal results in the worst case. Therefore, the number of enriched POIs – seen in relation to the total number of POIs – can be used to evaluate (1) external data quality and (2) appropriateness of critical parameters within the configuration of the enrichment process. This metric is implemented using simple SPARQL count queries within Apache Jena, runs in O(n log n) and therefore demands no further optimization.
- **Maximum & Average Number of Triples Added by Enrichment.** Even if POIs get enriched by DEER, they are not *enriched equally* for same reasons as we just listed. Varying availability of some attributes can result in sparse enrichment, which might not be tolerable for some use cases. However, when we assume that the availability of attributes underlies a random distribution, a large POI dataset is likely to have at least one POI with a lot of enriched attributes. Therefore, we introduce the **maximum** and **average number of triples added by enrichment** quality metrics, which can tell the user about the *potential* of the enrichment process on the one hand and the *realization* of that potential under the current data availability situation and enrichment parametrization on the other hand. These metrices are implemented within the same loop as the *number of enriched POIs* metric and therefore also demands no further optimization.
- **Relative Information Gain.** While the interpretations of the previous metrics are more on the qualitative side, this metric completes the perspective on POI data quality after enrichment by adding a quantitative evaluation of the amount of information value that is added to a dataset. It is computed as *100 / total number of triples in source * total number of triples in target – 100*. This metric is implemented using the size methods of Apache Jena Model and therefore runs in O(n), which demands no further optimization.

- **Operator Metrics.** The main takeaway of the experience of our initial definition of quality assurance metrics for enrichment in [20] was that a fine-grained evaluation of each enrichment operators general characteristics was too technical as only the developers of DEER could interpret them and the other partners had no intuition about their meaning. However, operator metrics are still included for advanced users that want to have more insights into the

enrichment process. They are computed on-the-fly while executing the enrichment and range in O(1) to O(n) complexity, which is why they also do not necessitate additional optimization.

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.

4. Evaluation

In this Section, we evaluate the SLIPO Workbench as well as the SLIPO Toolkit components regarding data quality in a real-world scenario involving large POI datasets. The SLIPO Workbench, as thoroughly discussed in [22], integrates all the pieces from the SLIPO Toolkit into a modern, user-friendly web app. As its functional centerpiece, the Workbench offers the design and execution of POI data integration workflows using intuitive graphical interfaces. In order to encourage iterative improvements of the POI data integration workflows and therefore their results data quality, the SLIPO Workbench implements dedicated views to monitor the quality assurance metrics for each of the integrated SLIPO Toolkit components.

These views consist of up to three GUI components which complement one another:

- Tree view of a raw JSON *quality assurance report* as generated by each of the SLIPO Toolkit components;
- Configurable *visualization* (in bar or pie charts) of selected quality metrics from the JSON quality assurance report; and
- Map view which shows the *spatial extent* (bounding box) of all the POIs in a given output POI file from one of the SLIPO Toolkit components that produce such outputs.

Next, we will first present some basic information about the data sources used and the specific experimental setup of the quality assurance evaluation. We then assess the utility and efficiency of particular measures and statistics issued by each of the SLIPO Toolkit components used within the data integration workflow. All the experiments are defined and carried out within the SLIPO Workbench.

4.1. Experimental Setup

4.1.1. Datasets

We used POIs extracted from OpenStreetMap (OSM) as datasets in this evaluation. More specifically, we made use of a real-world OSM dataset with 7447691 POIs across Europe used in Deliverable D2.4 [17]. After filtering of OSM tags, we isolated particular attributes concerning POIs as listed in Table 2. These attributes include basic information that identifies each POI (location, name, type), but also extra attributes, such as *address* and *contact* information, as well as other complementary attributes. Note that geometric information in the collected POIs concerns not only *Points* (in longitude/latitude coordinates) according to the OGC geometry types, but also *LineStrings*, *MultiLineStrings*, *MultiPolygons*, as well as *Geometry Collections*. As POI names in OSM may be written in different languages, alphabets, phonetic representations, etc., this hinders data integration tasks. To overcome

this issue (as detailed in Deliverable D5.2 [21]), we performed transliteration of name strings from any language into Latin, so this *transliterated value* is henceforth considered as the POI name. The resulting real-world data is henceforth called *dataset **B***.

We also synthetically created another *dataset **A***, by taking advantage of the *generator utility*[19] included in TripleGeo, which enables creation of synthetic POI data. More specifically, this generator took the real-world dataset ***B*** as seed, and for each POI it made the following alterations:

- *Displacement of locations*. Original POI geometries in dataset ***B*** were randomly *translated* by parameters *dx*=0.001, *dy*=0.001 decimal degrees along the *x*, *y* axes. Thus, the geometry of each POI in the resulting dataset ***A*** is found in small distance from the corresponding POI in dataset ***B***. This small displacement makes it possible to identify such matching POIs in a data integration task.
- *Altered POI names*. The transliterated names obtained from dataset ***B*** are changed by erasing one random character with a probability of 50% from string names having more than 5 characters. This creates POI names in dataset ***A*** having strong similarity with those in dataset ***B***, hence facilitating their matching.
- *Eliminated thematic attributes*. In dataset ***A*** only the basic attributes (listed in the first column of Table 2) were retained from dataset ***B***. This was a deliberate decision, as the synthetically created dataset ***B*** will be used as the base dataset in this integration task. Our objective was to integrate as much information as possible from ***A***, so that eventually the output data will be as rich as ***B*** in terms of the number of available properties and *NOT NULL* values in as many properties as possible. In other words, the output dataset should retain information from dataset ***A*** but also restore information originally provided in dataset ***B***.

Overall, dataset ***A*** has exactly the same number of POIs as dataset ***B***, but it contains less attributes and has displaced geometries and distorted name values with respect to ***B***.

Basic attributes	Address-related attributes	Contact-related attributes	Other attributes
geometry	street name	phone number	international name
OSM identifier	house number	fax number	category
name	postal code	email address	country
type	city	webpage	opening hours

Table 2: Attributes extracted from OpenStreetMap concerning POIs

[19] https://github.com/SLIPO-EU/TripleGeo/blob/master/src/eu/slipo/athenarc/triplegeo/extra/SyntheticDataGenerator.java

4.1.2. Data Integration Workflow

The data integration workflow for this scenario is designed in the SLIPO Workbench, as illustrated in Figure 5, and consists of the following steps:

- *Transformation*. Two instances of TripleGeo are employed, respectively for transforming OSM datasets A and B into RDF. Mappings from the original OSM attributes to the SLIPO ontology were already available from the DINUC A.2 and B.1, i.e., the use cases involving OSM data as described in Deliverable D5.2 [21].
- *Interlinking*. An instance of LIMES is invoked to identify matching POIs between the two transformed RDF datasets *A* and *B*. After several tests, the following similarity metric was used:

  AND(*Geo_Mean*(a.wkt,b.wkt)|0.85, *qgrams*(a.translit,b.translit)|0.45)

specifying that POI entities *a*, *b* coming respectively from dataset ***A***, ***B*** should have a mean spatial similarity above 0.85, which is synonymous to a mean spatial distance of about 176m between all the points belonging to their respective geometries. Moreover, ***A*** and ***B*** should have a *qgrams* similarity of their transliterated names above 0.45. The threshold for accepting links was set to 0.2, which was purposely relaxed in order not to miss valid links.

- *Fusion*. An instance of FAGI takes the previously detected links as well as the two transformed RDF datasets and produces a unified output dataset that contains all POIs from both inputs. A predefined fusion configuration is used that prescribes a set of rules for deciding how the values for each property of the linked POIs will be fused. Given that both datasets have a similar number of POI (i.e., identical *coverage*), the fusion output mode selected for this case is *SLIPO_default_aa_mode*. This actually produces a fused output that retains all POIs from dataset A, but also includes any extra information (more detailed geometry, longer literal values, additional properties) from dataset B in order to improve the number of attribute values available per POI (*completeness*).
- *Enrichment*. An instance of DEER is applied against the fused dataset using DBpedia as the SPARQL endpoint from which it attempts to find extra properties per POI. The configuration is similar to those utilized in DINUCs A.2 and B.1 for Deliverable D5.2 [21]. Since DEER requires a fast endpoint to run in an acceptable time for large input data sizes, a private DBpedia endpoint was set up and used in this test.

Once all stages were defined and properly configured, we executed this POI data integration workflow in the SLIPO workbench. We collected all data, metadata, and statistics issued by each component and next we discuss these results with particular emphasis on quality assurance metrics as estimated by each step in the overall POI integration process.

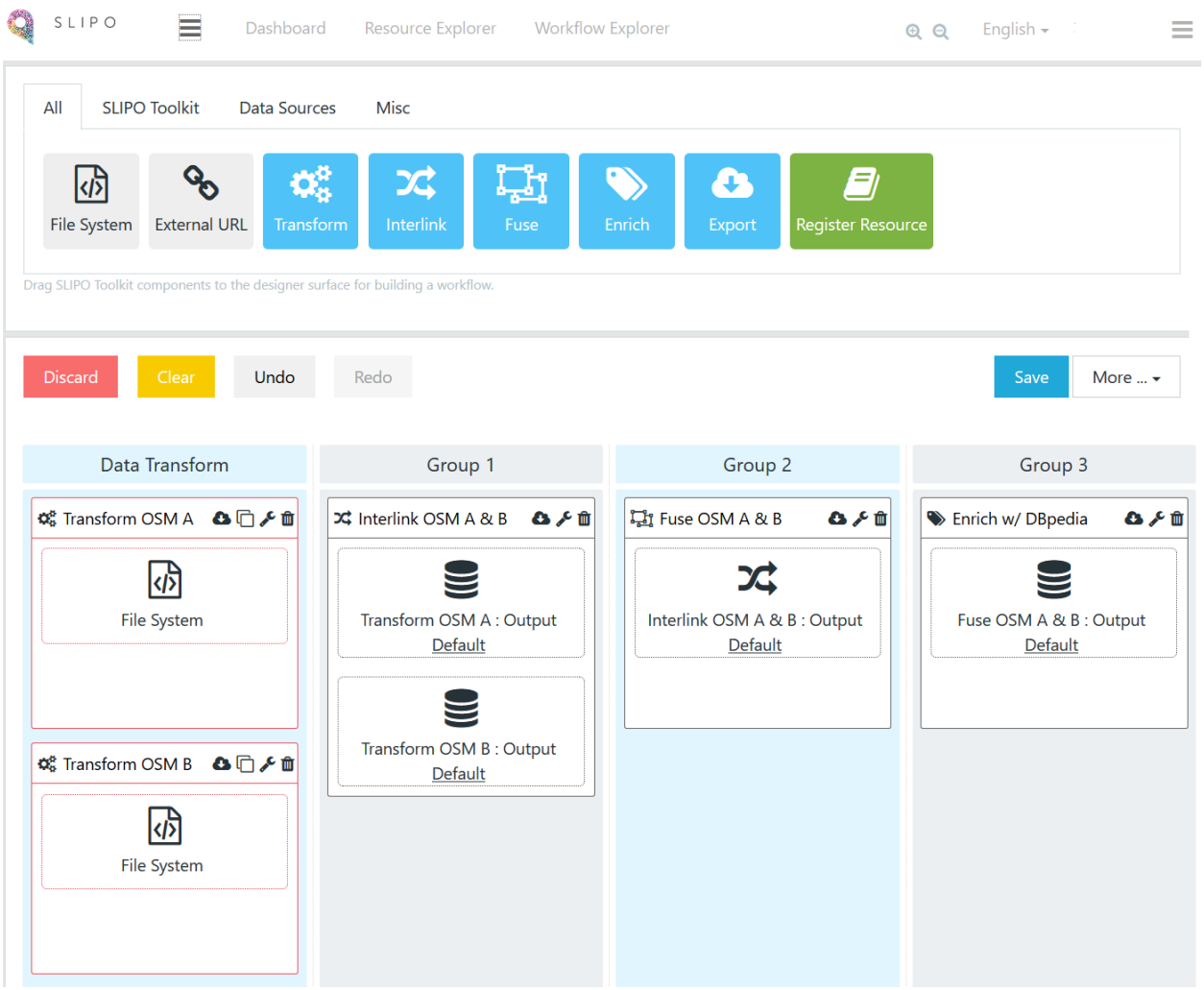

Figure 5: Integration workflow specified in the SLIPO Workbench

4.2. Assessment Results

In this Section, we discuss the quality assessment results issued after each stage of the POI data integration task.

4.2.1. Quality Metrics after Transformation

This task involves two transformation steps, respectively for OSM datasets ***A*** and ***B***. Once a transformation task is complete, a JSON file containing several quality metrics is issued. As discussed in Section 0, this JSON offers statistics about the number of attribute values transformed, the spatial extent of the POI geometries and several execution metadata. Since each such task is executed in isolation, we discuss their quality metrics separately:

KPI File : osm_pois_europe_A_metadata.json


```
"metadata" : { 3 items
  "Attribute Statistics" : { 8 items
    "ASSIGNED_CATEGORY" : 7447691
    "CATEGORY_URI" : 7447691
    "DATA_SOURCE" : 7447691
    "lat" : 7447691
    "lon" : 7447691
    "name" : 7447576
    "osm_id" : 7447691
    "translit" : 7447580
  }
  "MBR of transformed geometries (WGS84)" : { 4 items
    "X_max" : 98.915028
    "Y_max" : 81.02901750000001
    "X_min" : -144.6272057
    "Y_min" : -35.1205432
  }
  "Execution Metadata" : { 9 items
    "Input records transformed" : 7447691
    "Output triple count" : 156400496
    "Input record count" : 7447691
    "Execution time (ms)" : 700855
    "Output CRS" : "EPSG:4326"
    "Output file" : "/var/local/data-1/slipo/output/osm_pois_europe_A.nt"
    "Output serialization" : "N-TRIPLES"
    "Input records excluded" : 0
    "Transformation mode" : "STREAM"
  }
}
```


Figure 6: Quality metrics obtained after transformation of OSM dataset ***A***

- *OSM dataset* ***A***. From the JSON file issued after transformation (Figure 6), the user is able to readily inspect how many POI, attribute values and geometries have been transformed. In this case, no original POIs have been rejected during transformation, so all data from every original attribute has been successfully transformed into RDF triples as visualized in Figure 7. Note that this chart also includes extra properties (like DATA_SOURCE, CATEGORY_URI) that do not exist in the original dataset but are generated during transformation and added to the transformed result. In addition, the spatial extent of the transformed features is shown on map (Figure 8). Observe that this MBR encloses not only Europe, but also large parts of other continents. The reason is that dataset A includes POIs in overseas territories of several European countries in the Indian Ocean (e.g., Reunion), the Atlantic (Guadeloupe) and the Pacific (New Caledonia), hence the unusually enlarged extent.

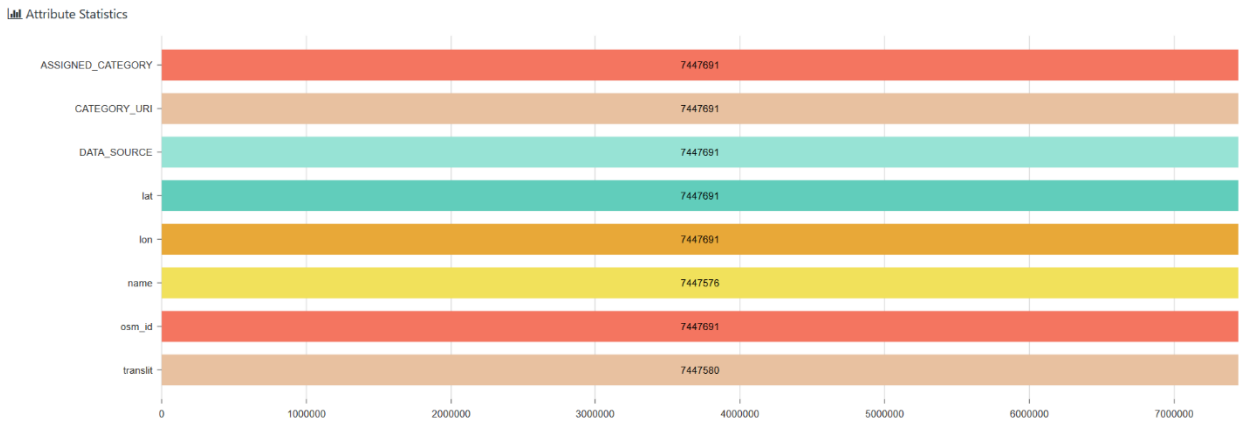

Figure 7: Attribute statistics regarding transformation of OSM dataset ***A***

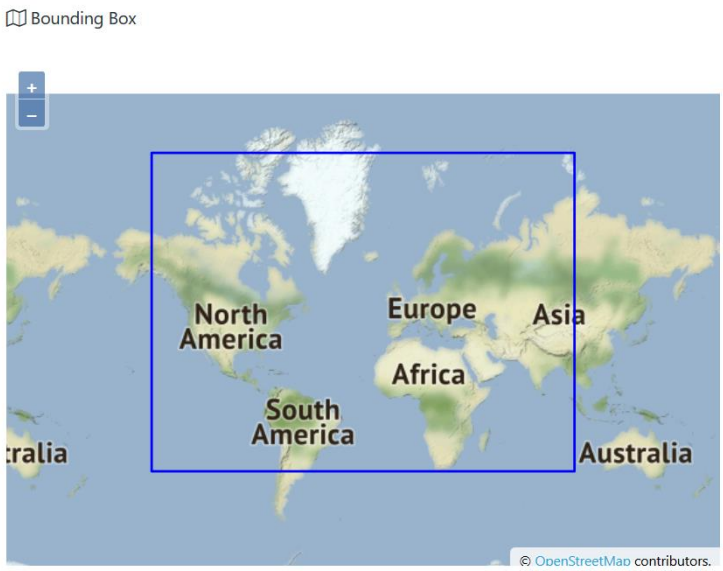

Figure 8: Spatial extent of POI features transformed from OSM dataset ***A***

- *OSM dataset **B***. From the JSON file issued after transformation (Figure 9), it is evident that the amount of values transformed per attribute differs widely for this dataset. Also based on the chart illustrated in Figure 10, the user easily identifies that several attributes (including identifier, name, categorization, and location) have been fully transformed. However, some attributes have yield relatively fewer triples (e.g., addresses, phone numbers) and certain attributes very few triples (like those for fax numbers or email addresses). This is a clear indication that the completeness of the original dataset is limited. However, since the other dataset ***A*** does not contain even this limited extra information, the user may decide whether integration with dataset ***B*** is meaningful and is expected to add any extra value to the POI data. Finally, the user is also able to inspect the spatial extent of the transformed features (Figure 11). Again, notice that the box is enlarged for the same reason as in dataset ***A***.

KPI File : osm_pois_europe_B_metadata.json


```
"metadata" : { 3 items
  "Attribute Statistics" : { 19 items
    "ASSIGNED_CATEGORY" : 7447691
    "CATEGORY_URI" : 7447691
    "DATA_SOURCE" : 7447691
    "TRANSLIT" : 7447569
    "city" : 966636
    "country" : 629459
    "email" : 147692
    "fax" : 81112
    "housenumber" : 1052218
    "int_name" : 78846
    "lat" : 7447691
    "lon" : 7447691
    "name" : 7447576
    "opening_hours" : 445802
    "osm_id" : 7447691
    "phone" : 560273
    "postcode" : 1191023
    "street" : 1179736
    "website" : 708213
  }
  "MBR of transformed geometries (WGS84)" : { 4 items
    "X_max" : 98.914028
    "Y_max" : 81.0280175
    "X_min" : -144.6282057
    "Y_min" : -35.1215432
  }
  "Execution Metadata" : { 9 items
    "Input records transformed" : 7447691
    "Output triple count" : 170351568
    "Input record count" : 7447691
    "Execution time (ms)" : 685124
    "Output CRS" : "EPSG:4326"
    "Output file" : "/var/local/data-1/slipo/output/osm_pois_europe_B.nt"
    "Output serialization" : "N-TRIPLES"
    "Input records excluded" : 0
    "Transformation mode" : "STREAM"
  }
}
```


Figure 9: Quality metrics in JSON after transformation of OSM dataset *B*

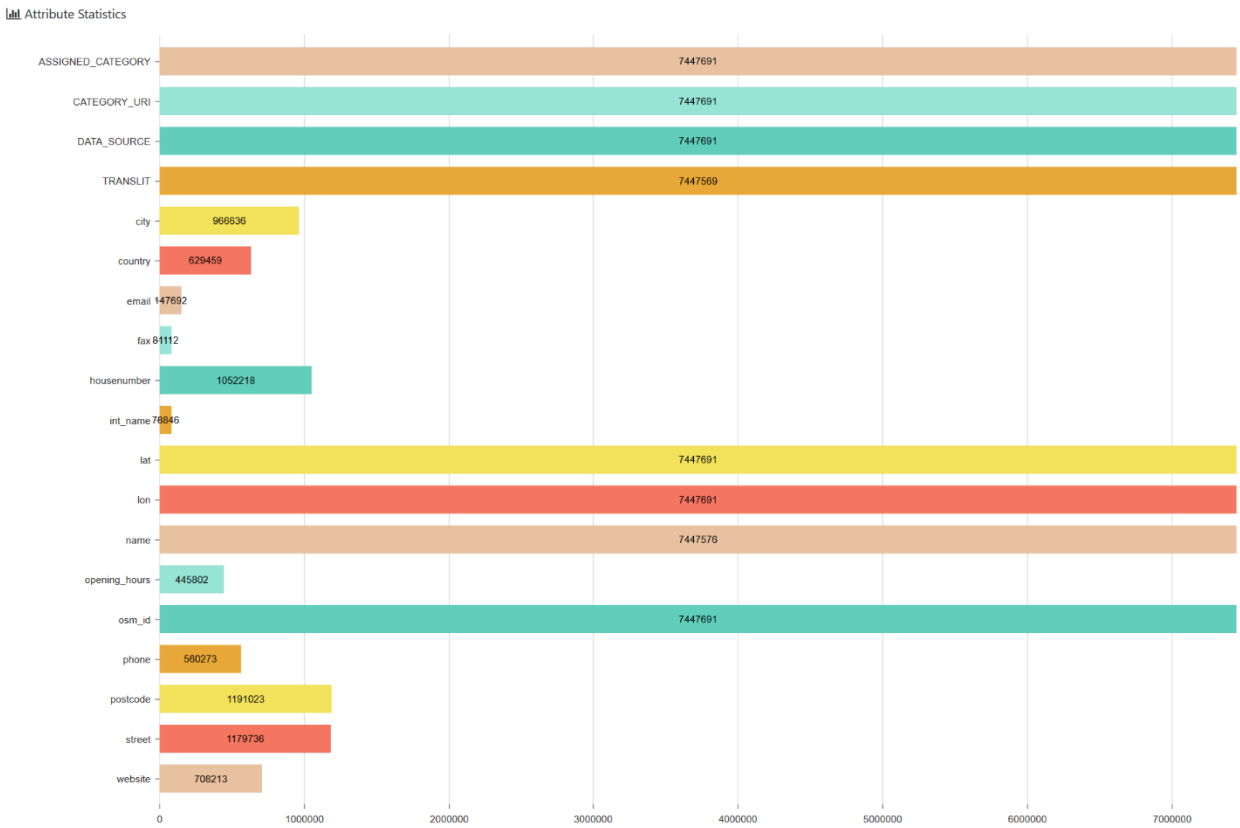

Figure 10: Attribute statistics regarding transformation of OSM dataset B

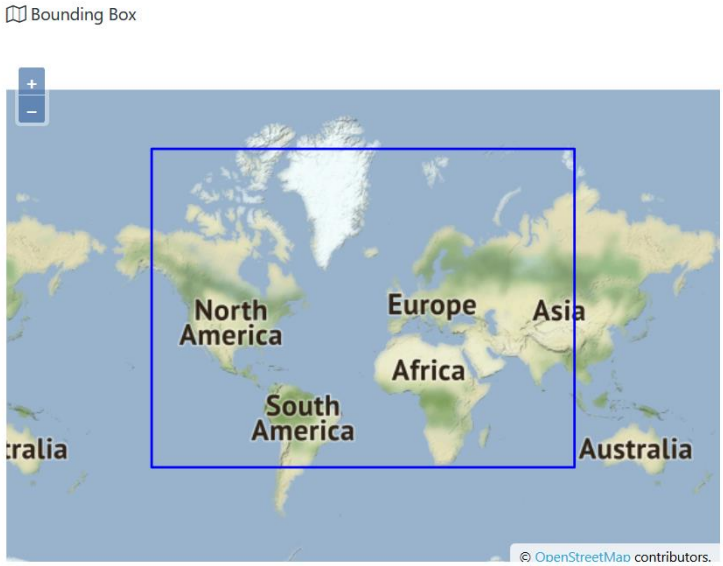

Figure 11: Spatial extent of POI features transformed from OSM dataset B

4.2.2. Quality Metrics after Linking

This task involved the interlinking of both transformed datasets with the Links Specification (LS) as defined in Section 4.1.2.

Concerning the quantitative metrics, it is worth mentioning that the number of POIs considered for linking is slightly less than the original ones available in the input datasets. Indeed, since the similarity metric for identifying links involves both the *geographic location* and the *transliterated names* of POIs, only those having both properties can be considered. Since some POIs do not have names (122 for *source* dataset ***A*** and 111 for *target* dataset ***B***), they are excluded from the interlinking process, as indicated by the input sizes listed in Figure 12.

Concerning the qualitative metrics, we can see a Pseudo-Recall of almost 1.0, meaning that we found at least one link for almost all POIs in the input datasets. Comparing this with the Pseudo-Precision of about 0.73, we can reason that for a considerable portion of POIs in the source dataset ***A*** we have found multiple links to POIs in the target dataset ***B***. However, note that we value recall more than precision within the scope of the SLIPO workbench as we can filter and resolve duplicate links in the subsequent fusion phase. Therefore, these quality assurance results clearly show that our link specification was able to meet the requirements of this task.

KPI File : links_10M.json


```
"metadata" : {  4 items
  "mappingTime" : 4372484
  "inputSizes" : {  2 items
    "source" : 7447569
    "target" : 7447580
  }
  "outputSizes" : {  2 items
    "verification" : 0
    "acceptance" : 10155319
  }
  "pseudoPRF" : {  2 items
    "acceptance" : {  3 items
      "precision" : 0.733307
      "recall" : 0.999918
      "f-measure" : 0.846107
    }
    "all" : {  3 items
      "precision" : 0.733307
      "recall" : 0.999918
      "f-measure" : 0.846107
    }
  }
}
```


Figure 12: Quality metrics in JSON after linking of OSM datasets ***A*** and ***B***

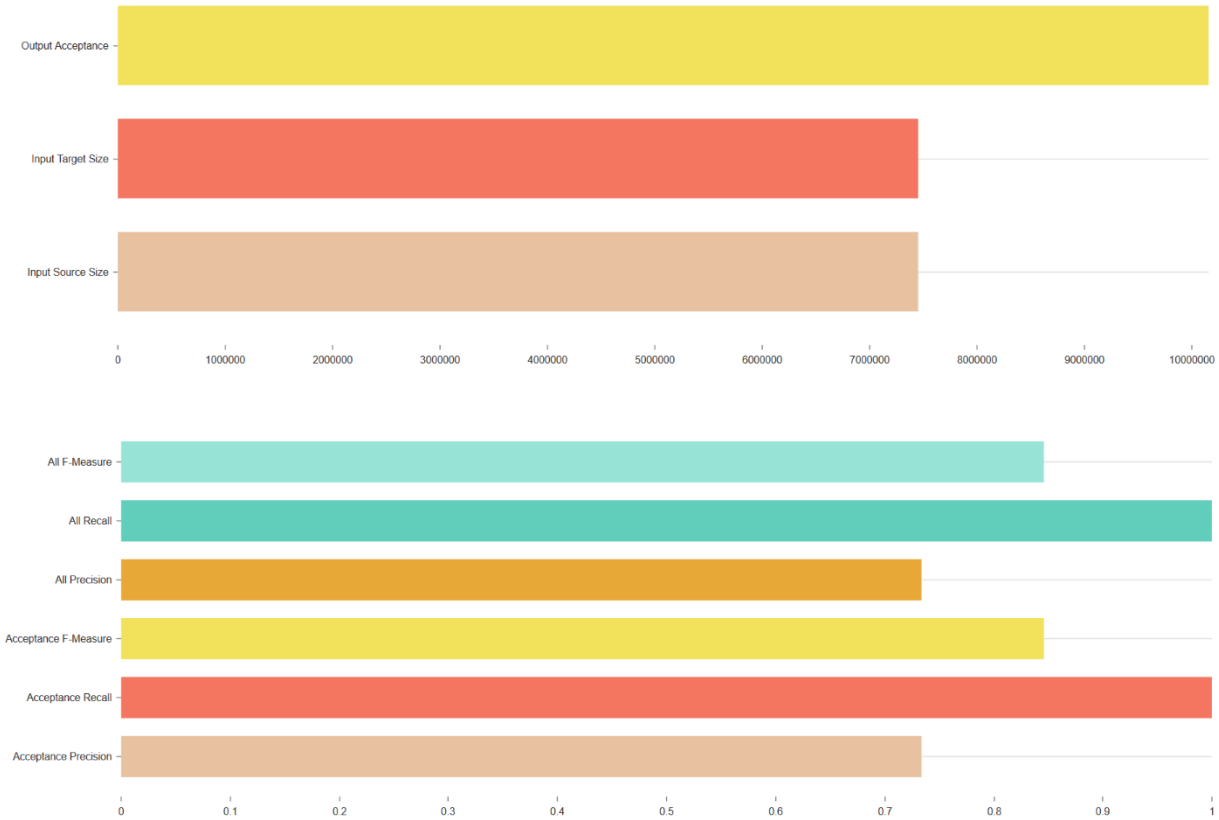

Figure 13: Visualization of quality metrics from linking of OSM datasets ***A*** and ***B***

4.2.3. Quality Metrics after Fusion

As explained in Section 4.1.2, the fusion process in this workflow takes the previously identified links and enhances the completeness of input dataset ***B*** with as many extra properties and values from dataset ***B***. Figure 14 lists the JSON file (split in two columns) with all statistics and quality metrics measured for the entire datasets. These values can be either *ABSOLUTE* numbers (e.g., number of POIs, number of links) or *PERCENTAGE* measurements (e.g., attribute gain, fusion confidence) as indicated in each item. For each category, the user is able to visualize these statistics. As illustrated in Figure 15 for absolute metrics, our fusion software FAGI considers all contents of the transformed datasets A and B and every available link between them as obtained after their interlinking using LIMES. No link is rejected during their validation, and also note that all links are unique. However, the total number of fused POIs is slightly less (7446988) than those in original base dataset ***B*** (7447691). Indeed, 703 POIs are missing from the fused dataset as they failed during their validation due to geographic discrepancies. Indeed, FAGI computes distances between candidate pairs of POIs in order to verify they proximity and accordingly determine their fusion into a single POI. When such candidates are near the

poles of the Earth or their geometries are complex (in particular, multipolygons or geometry collections), orthodromic distances cannot be computed and these POIs are excluded from any further consideration.

To estimate the *overall quality* of the fused POIs, FAGI offers the *aggregated quality metrics* discussed in Section 3.3.3. As each fused POI is always issued with an indication of attribute gain and fusion confidence, Figure 16 illustrates a visualization of their aggregated values for the entire fused data (in *percentage*). In particular, *average attribute gain* is rather low (7%) in this scenario, meaning that each POI in dataset ***B*** after fusion with dataset ***A*** gets only 7% extra attributes on average. But this should be expected, given that original OSM data has poor thematic information at global scale. For some of these extra attributes (e.g., addresses, phone numbers) it offers relatively few *NOT NULL* values, and for certain attributes (like those for fax numbers or email addresses) very little information as illustrated in Figure 10. However, *maximum attribute gain* among all POIs is 41%, so there are POIs that can attain many more extra properties than the average. In use cases like this, POI data integration may be worthwhile as it enhances the amount of available POI characteristics.

But overall, it is very important that *fusion confidence* is 90% *on average*. Since this is very close to a confidence value of 100% that signifies perfect match between the pair of POIs which resulted in a fused POI, it is evident that the quality of the fused dataset is very high. This offers to users a valuable indication of whether to trust the results of this process. Note that each fused POI also contains detailed measures regarding its attribute gain and confidence value, so users are able to inspect these results, filter them according to criteria (e.g., reject those fused POIs of confidence less than 50%), and possibly change the configuration for fusion and repeat the process.

KPI File : fusion_merged_stats.json


```
"metadata" : { 11 items
  "executionTime" : { 3 items
    "label" : "Execution time (ms)"
    "type" : "ABSOLUTE"
    "items" : [ 3 items
      0 : { 2 items
        "label" : "Dataset load"
        "value" : "180138"
      }
      1 : { 2 items
        "label" : "Fusion"
        "value" : "2322804"
      }
      2 : { 2 items
        "label" : "Statistics"
        "value" : "36232"
      }
    ]
  }
  "averageConfidence" : { 3 items
    "label" : "Average Confidence"
    "type" : "PERCENT"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Average Confidence"
        "value" : "0.9"
      }
    ]
  }
  "averageGain" : { 3 items
    "label" : "Average Gain"
    "type" : "PERCENT"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Average Gain"
        "value" : "0.07"
      }
    ]
  }
  "maxGain" : { 3 items
    "label" : "Max Gain"
    "type" : "PERCENT"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Max Gain"
        "value" : "0.41"
      }
    ]
  }
  "initialLinks" : { 3 items
    "label" : "Initial links"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Initial links"
        "value" : "10155319"
      }
    ]
  }
  "uniqueRejectedLinks" : { 3 items
    "label" : "Unique vs rejected links"
    "type" : "ABSOLUTE"
    "items" : [ 2 items
      0 : { 2 items
        "label" : "Unique links"
        "value" : "10155319"
      }
      1 : { 2 items
        "label" : "Rejected links"
        "value" : "0"
      }
    ]
  }
  "uniqueLinks" : { 3 items
    "label" : "Unique links"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Unique links"
        "value" : "10155319"
      }
    ]
  }
  "initialPOIsA" : { 3 items
    "label" : "Initial POIs in A"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "POIs in A"
        "value" : "7447691"
      }
    ]
  }
  "initialPOIsB" : { 3 items
    "label" : "Initial POIs in B"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "POIs in B"
        "value" : "7447691"
      }
    ]
  }
  "fusedPOIs" : { 3 items
    "label" : "Fused POIs"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 2 items
        "label" : "Fused POIs"
        "value" : "7446988"
      }
    ]
  }
  "poisInFinalDataset" : { 3 items
    "label" : "POIs in final dataset"
    "type" : "ABSOLUTE"
    "items" : [ 1 item
      0 : { 3 items
        "label" : "POIs in final dataset"
        "type" : "ABSOLUTE"
        "value" : "7446988"
      }
    ]
  }
}
```


Figure 14: Quality metrics in JSON after fusion of OSM datasets ***A*** and ***B***

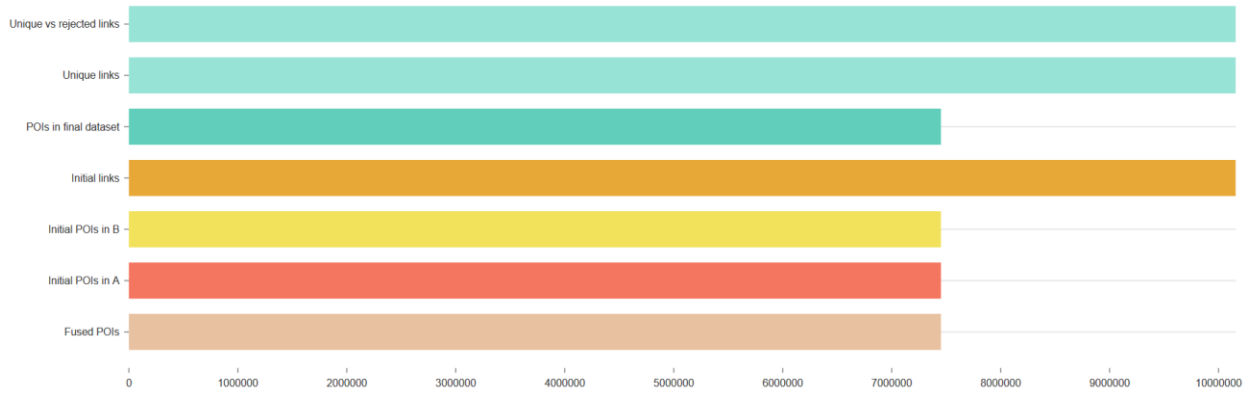

Figure 15: Visualization of *absolute* quality metrics after fusing OSM datasets ***A*** and ***B***

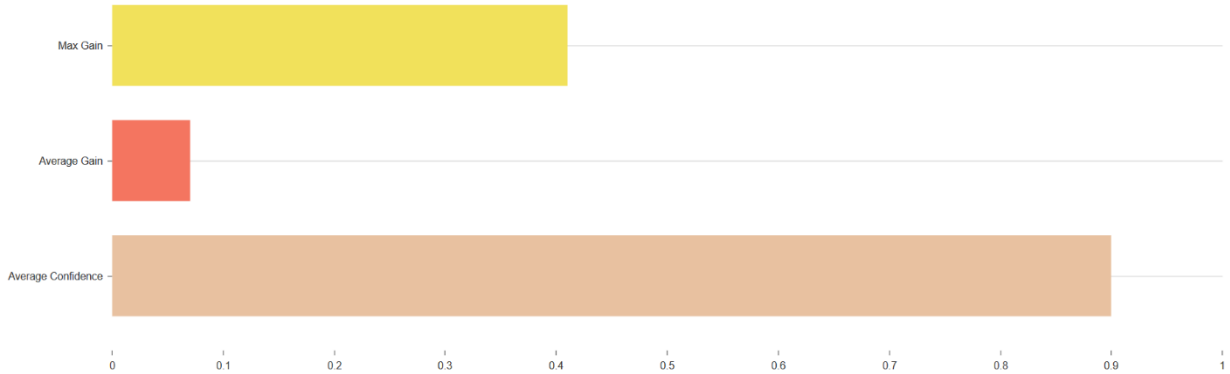

Figure 16: Visualization of *percentage* quality metrics after fusing OSM datasets ***A*** and ***B***

4.2.4. Quality Metrics after Enrichment

The enrichment we performed on the fused data consisted of finding the nearest buildings from DBpedia in a radius of 2km for each POI and then dereferencing a number of interesting properties, such as number of floors and parking slots. As (1) the buildings available in DBpedia are not equally distributed over the continent and (2) not all buildings contain all of the properties that we wanted to dereference, not all of the POIs could benefit from the enrichment process, as we can see in Figure 17. In particular, about 16.5% of all POIs get enriched as illustrated in the pie chart in Figure 17. Note that the total number of POIs considered for enrichment is exactly those (7446988) obtained after fusion. In other words, our enrichment process is lossless. Overall, the total number of properties that were added during enrichment amount to 1.62% of all properties in the final enriched dataset.

KPI File : deer-analytics.json


```
"metadata" : { 2 items
  "globalStats" : { 6 items
    "enrichedPOIs" : 1232743
    "maxEnrichedTriples" : 17
    "totalEnrichedTriples" : 6444048
    "totalPOIs" : 7446988
    "enrichmentPercentage" : 1.62
    "avgEnrichedTriples" : 1
  }
}
```


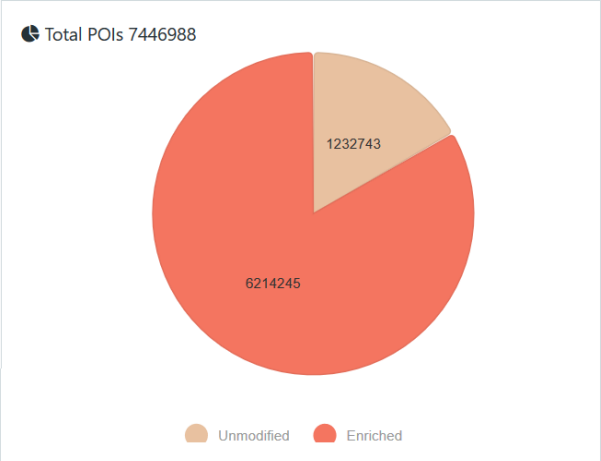

Figure 17: Quality metrics in JSON after enrichment of the fused dataset

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] 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.

[4] 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.

[5] Rodolphe Devillers, Yvan Bedard, 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.

[6] Rodolphe Devillers, Marc Gervais, Yvan Bedard, 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.

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

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

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

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

[11] 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.

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

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

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

[15] 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.

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

[17] SLIPO Project. Deliverable D2.4: Final Mapping and Transformation Software.

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

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

[20] SLIPO Project. Deliverable D3.5: Initial report on Quality Assurance Metrics, Indicators and Processes.

[21] SLIPO Project. Deliverable D5.2: Pilot Report.

[22] SLIPO Project. Deliverable D1.4: Final SLIPO System.

[23] Talhofer V., Hošková-Mayerová Š., Hofmann A. (2019) The Influence of Quality of Geospatial Data on the Decision-Making Processes. In: Quality of Spatial Data in Command and Control System. Studies in Systems, Decision and Control, vol 168. Springer, Cham

[24] Jennings Anderson, Robert Soden, Brian Keegan, Leysia Palen & Kenneth M. Anderson (2018) The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters, International Journal of Human–Computer Interaction, 34:4, 295-310

[25] Homburg T., Boochs F. (2019) Situation-Dependent Data Quality Analysis for Geospatial Data Using Semantic Technologies. In: Abramowicz W., Paschke A. (eds) Business Information Systems Workshops. BIS 2018. Lecture Notes in Business Information Processing, vol 339. Springer, Cham

[26] Rietveld, Laurens, et al. "Structural properties as proxy for semantic relevance in RDF graph sampling." International Semantic Web Conference. Springer, Cham, 2014.

[27] Cheng, G., Tran, T., & Qu, Y. (2011, October). RELIN: relatedness and informativeness-based centrality for entity summarization. In International Semantic Web Conference (pp. 114-129). Springer, Berlin, Heidelberg.

[28] Gunaratna, K., Thirunarayan, K., & Sheth, A. (2015, February). FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

[29] Thalhammer, A., Lasierra, N., & Rettinger, A. (2016, June). LINKSUM: using link analysis to summarize entity data. In International Conference on Web Engineering (pp. 244-261). Springer, Cham.

[30] Gunaratna, K., Yazdavar, A. H., Thirunarayan, K., Sheth, A., & Cheng, G. (2017, August). Relatedness-based multi-entity summarization. In IJCAI: proceedings of the conference (Vol. 2017, p. 1060). NIH Public Access.

[31] Wei, D., & Liu, Y. (2019). ESA: Entity Summarization with Attention. arXiv preprint arXiv:1905.10625.