SLIPO

Scalable Linking and Integration of Big POI data

Data Sharing and Integration,
European Big Data Value Forum,
Versailles, 17/11/22

Giorgos Giannopoulos
Athena RC

What are POIs?

- **Points of Interest** (POIs)
  - Locations that exhibit a certain interest
- The **concept** of a POI is quite **broad**
  - Points, Polygons; Metadata; Relations to each other; Spatial, temporal, and/or thematic contexts
- The **foundation** of our physical and digital Economy (*expansive value chain*)
  - **What is there?**
  - **How do I get there?**
  - **Where do I find something?**

Why is POI integration challenging?

The challenge with POIs

- **Ambiguity; no common identifiers**
  - Lack of standardization in models, formats, identifiers
- **Labor-intensive**
  - Long update cycles
  - Fragmented
  - Quality; manual validation
- **More and better POI data at a fraction of the cost for the entire value-chain**
  - **Cross-border**
  - **Cross-domain**

Same POI, different geometries (points, polygons).

Different POIs, same location.

POI (point) within another POI (polygon).

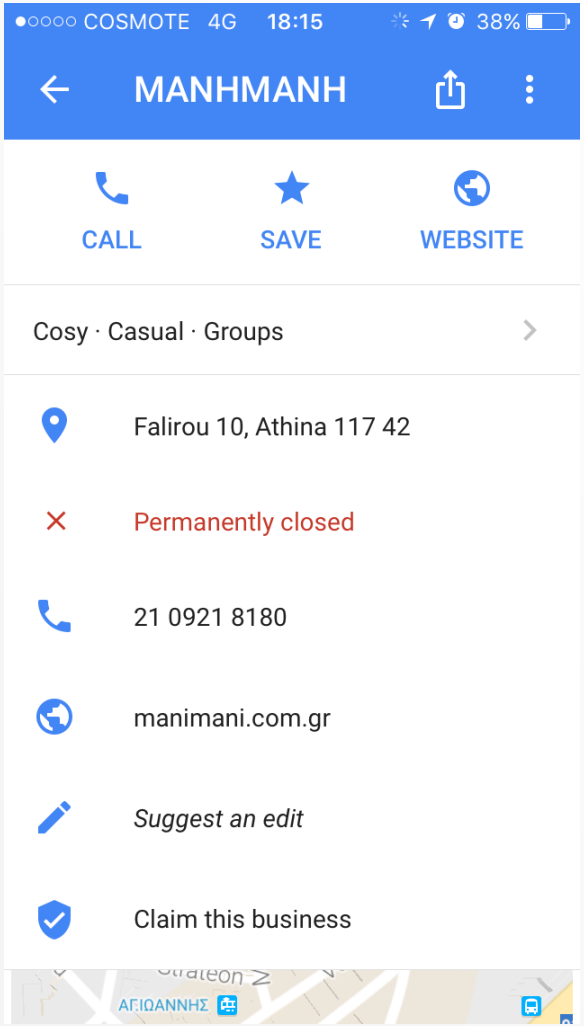
Diffeent POIs belonging to the same chain.

Defunct POI.

Same POI, change of type.

Obstacles

- Quality
  - Erroneous, conflicting, or outdated data
- Coverage, completeness, richness
  - POI categories, attributes or geographic areas
- Interoperability
  - Diverse data formats and schemas; ad-hoc solutions
- Volume
  - World-scale: no complete, commercial-grade solutions
- Data sharing and trust
  - **Proprietary** data are closely-guarded
  - Quality-assured integration of open data a challenge/opportunity



Happens to the best of them!

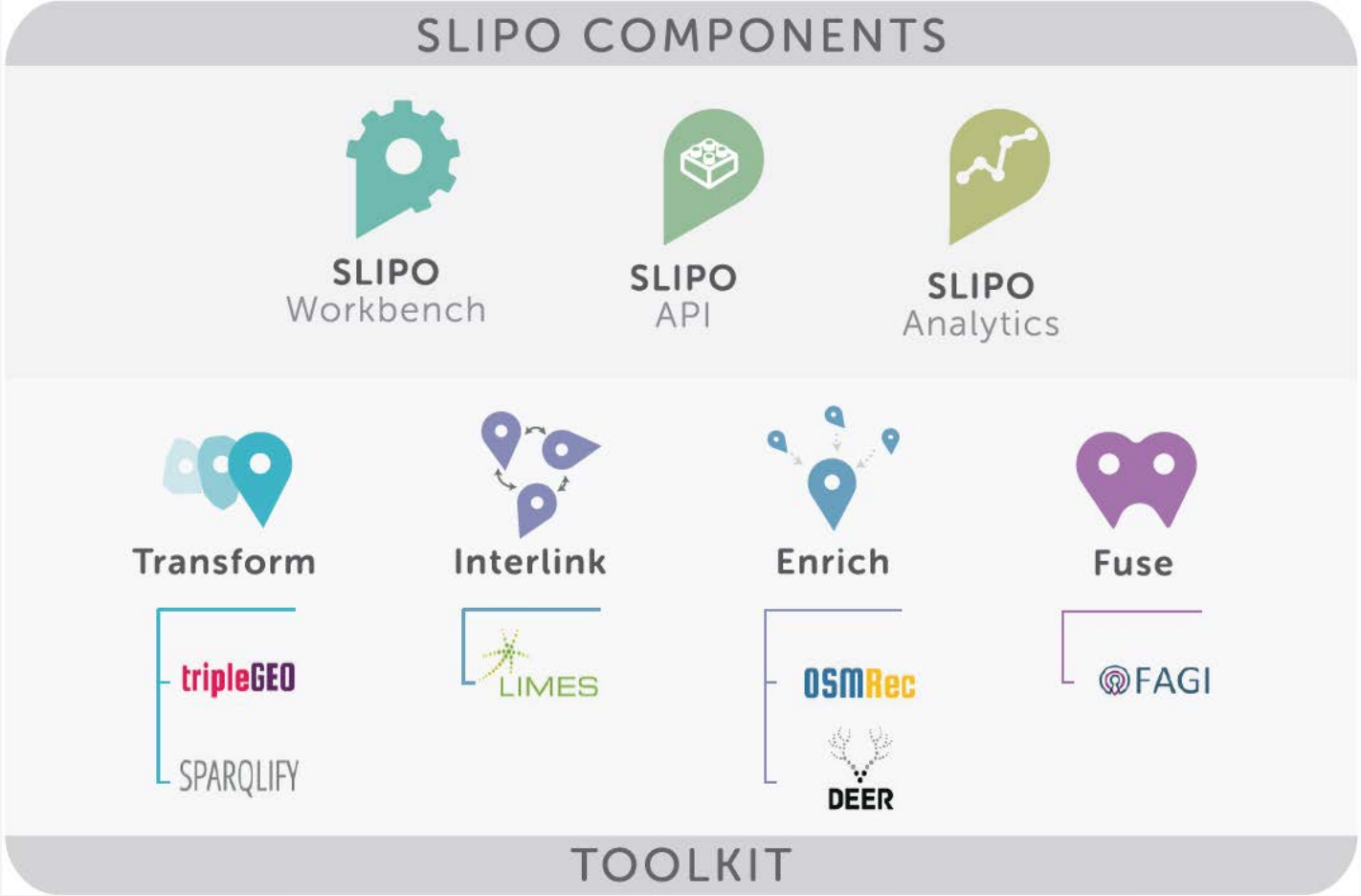
Geo-marketing Use Case

- Typical B2B scenario
  - Help a *company* (supplier) to identify *competitor* (suppliers) companies and new *customers (sellers)* in an area
- Integration tasks
  - Compare company's **customer database** with external sources to identify existing and new customers and update
    - Harvest; Transform; Interlink; Fuse
  - Compare company's **competitor database** with external sources enrich their metadata (e.g., customers of the competitors)
    - Harvest; Transform; Interlink; Enrich
  - Spatial statistics on the areas
    - Value added services

SLIPO

- **Software, models & processes for scalable and quality-assured POI integration**
  - **Transform** POIs into RDF
  - **Interlink** POIs from different datasets
  - **Enrich** POIs with additional metadata
  - **Fuse** Linked POI data
  - Assess the **quality** of POI data
- **Results**
  - SLIPO Toolkit; SLIPO APIs
  - SLIPO Ontology; Open POI data

Transfer integration into the Linked Data domain


SLIPO COMPONENTS
SLIPO Workbench
SLIPO API
SLIPO Analytics
Transform
tripleGEO
SPARQLIFY
Interlink
LIMES
Enrich
OSMRec
DEER
Fuse
FAGI
TOOLKIT

Transformation

- **TripleGeo**
  - Major performance/scalability advancements
    - **Currently more than x10 speed-up**
    - **Data partitioning scheme for parallel processing**
  - Extended set of supported POI input sources and formats
    - **Oracle; PostGIS; MySQL; SQL Server; IBM DB2; SpatiaLite**
    - **Shapefiles; KML; GML**
    - **CSV; GPX; GeoJSON; OSM XML**
  - RML mappings to the SLIPO Ontology
- **SLIPO Ontology**
  - Support of concepts (classes/properties) from industrial/open data sources

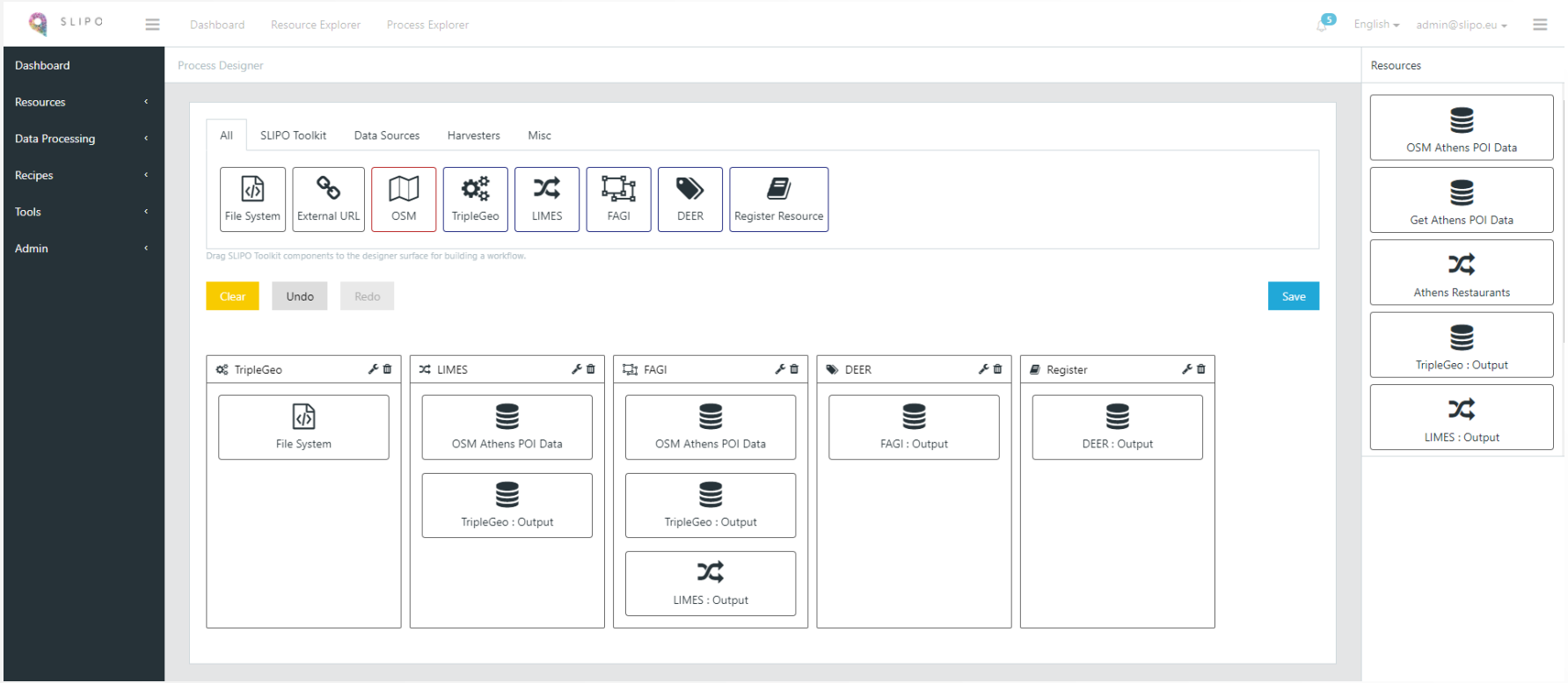
Interlinking

- **LIMES**
  - Extended functionality for textual and spatial matching on several POI attributes
  - Major performance/scalability advancements
    - **Parallelization of POI interlinking**
    - **Up to 10 times speed-up**
  - Topological linking of POIs
    - **Linking of 2M POIs in 20min**
    - **x800 faster than current competition**
  - Integration with SoA distributed processing frameworks
    - **Apache Flink, Apache SPARK**

Fusion and Enrichment

- **FAGI**
  - Focus on data quality and correctness
    - **Dataset-tuned fusion actions**
    - **Extended functionality for comparing similarity of POI attributes → Fusion automation**
  - Scalability
    - **Prototype distributed version of FAGI**
- **DEER**
  - Performance/scalability advancements
    - **Parallelization of the core enrichment algorithm**
  - Automation of enrichment
    - **Self-configuration of enrichment operators**
    - **Genetic programming algorithms for enrichment specification**

SLIPO Workbench (v0.8)



Thank you!

SLIPC

The SLIPO Solution

- Scalable, complete solutions for POI integration
  - Open, interoperable software for performing transformation, interlinking, enrichment and fusion of Big POI data
  - In parallel, ease the (unavoidable) manual effort by incorporating semi-automatic processes when possible
- Schema standardization and data exchange
  - Work on common schemas and vocabularies for representing POI data
  - Develop mapping and transformation software for most conventional POI formats
  - Exploit the potential of Linked Data technologies and standards

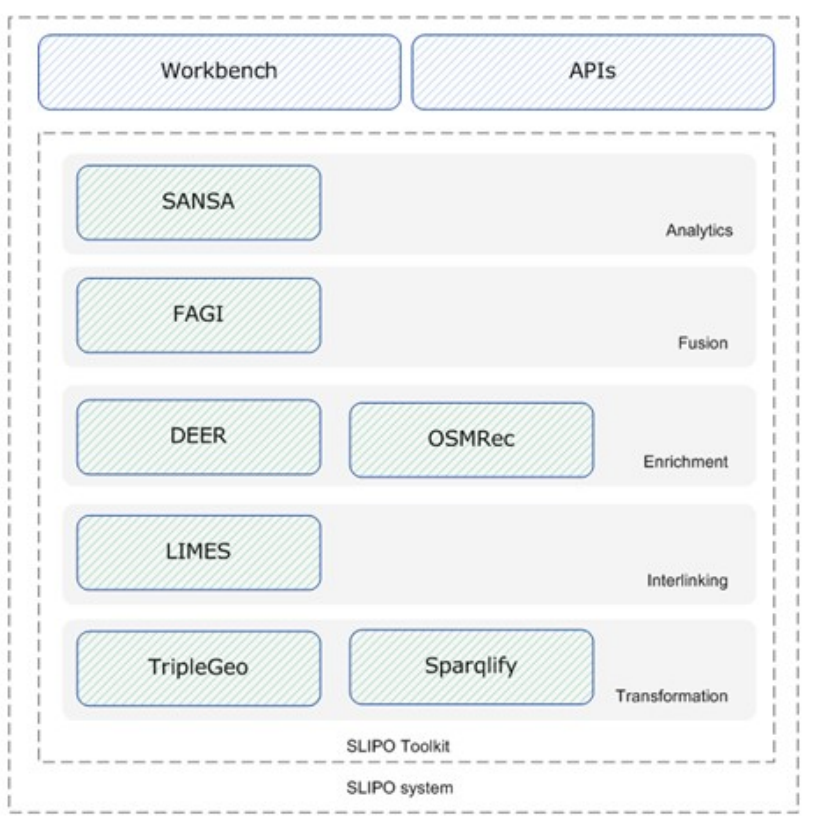
The SLIPO Solution

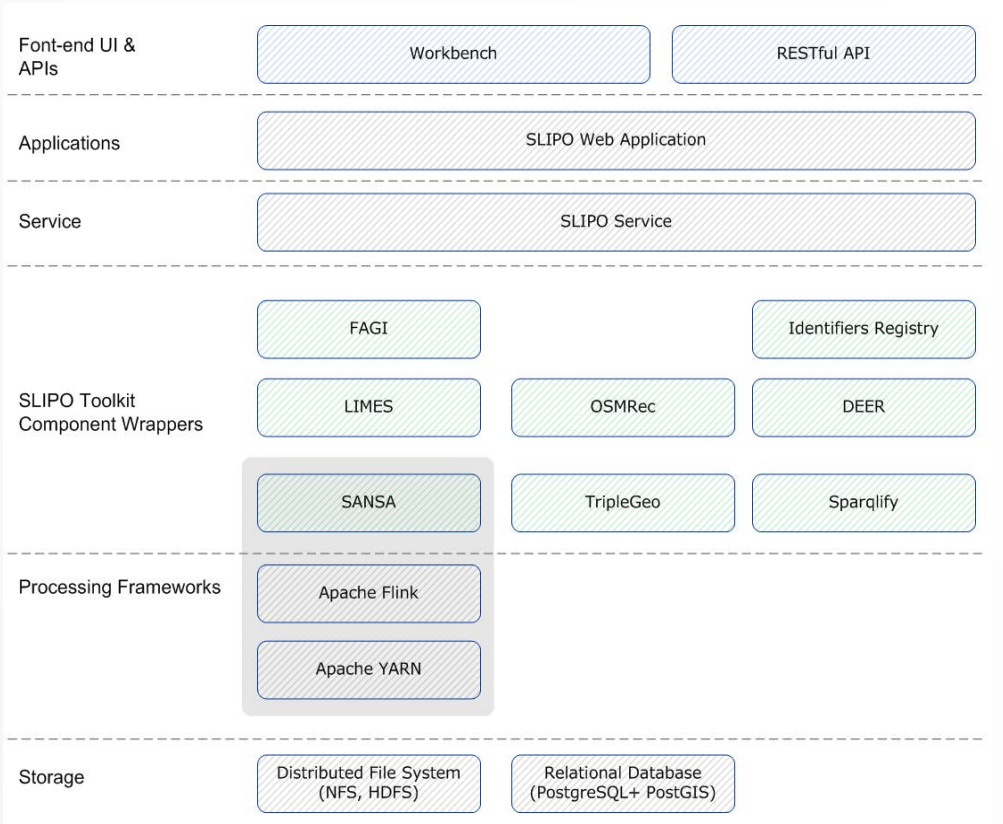
- Data quality
  - Incorporate quality assurance into every step of the POI integration process
    - Produce data-specific and task-specific quality indicators
    - Allow the stakeholder to evaluate the quality of both the data and the integration processes
  - Explore and exploit the data at hand
    - Build algorithms, metrics and configurations based on the individual characteristics of POI datasets
    - Fine-tune and generalize the developed tools with extensive experimentation on the data at hand
    - Continuous feedback from the stakeholders

The SLIPO Solution

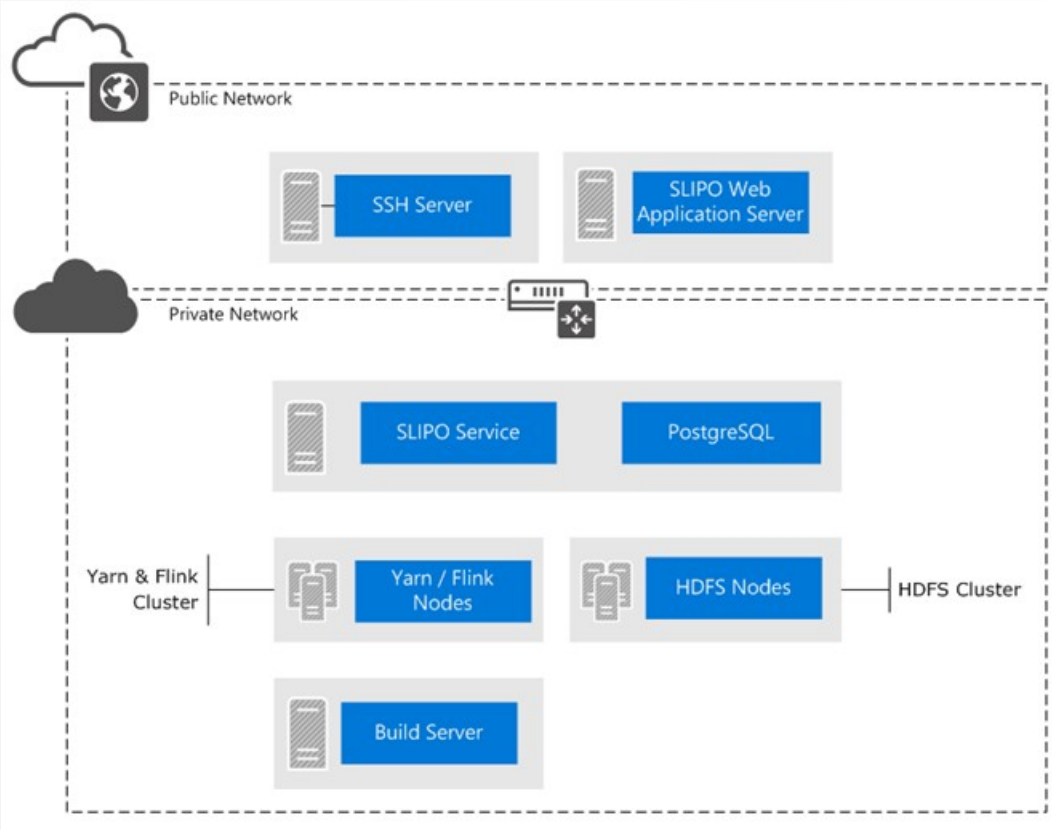
- Data coverage, completeness, richness
  - Achieved through the integration of more data sources and data types
    - Quality assurance
    - Automation of interlinking and enrichment
    - Agile POI schema that can incorporate new data types
- Data sharing and trust issues
  - Facilitate the exploitation of other data sources, while minimizing quality
  - Communicate the value of opening/sharing data, by leveraging the quality of existing open data sources and software

SLIPO Architecture





SLIPO Deployment



Requirements and KPIs

- **Requirements elicitation**
  - Web survey; Discussion panels; 1-1 interviews
  - 5 user stories; 13 use cases
- **KPIs**
  - Absolute POI **number**
  - **Coverage** in a given region
  - **Accuracy** (geospatial, attributes)
  - **Enrichment** (geometry, attributes, external sources)
  - **Timeliness**
  - **Update cycle** (time/effort)

More, better, fresh, accurate POIs at a fraction of the cost

Insights

- **Data quality first**
  - Accuracy of coordinates (geometries) and categories
  - Automation of the integration process is required, but not at the expense of quality (maintain QA via manual validation)
- **Coverage, completeness and attribute richness second**
  - POI updating/timeliness currently hindered
  - Interlinking of POIs often needs to take into account/combine several POI attributes
- **Performance is important but not critical**
  - Already orders of magnitude faster than manual/current SOA
  - Ensure world-scalability; relaxed time-constraints