Ontology Building vs Data Harvesting and Cleaning for Smart-city Services

Pierfrancesco Bellini, Monica Benigni, Riccardo Billero, **Paolo Nesi**, Nadia Rauch
Dipartimento di Ingegneria dell'Informazione, DINFO
Università degli Studi di Firenze
Via S. Marta 3, 50139, Firenze, Italy
Tel: +39-055-4796567, fax: +39-055-4796363
DISIT Lab
http://www.disit.dinfo.unifi.it *alias* http://www.disit.org , paolo.nesi@unifi.it

Proc. of the 20th International Conference on Distributed Multimedia Systems, Pittsburgh, USA, August 2014

Smart-City axes

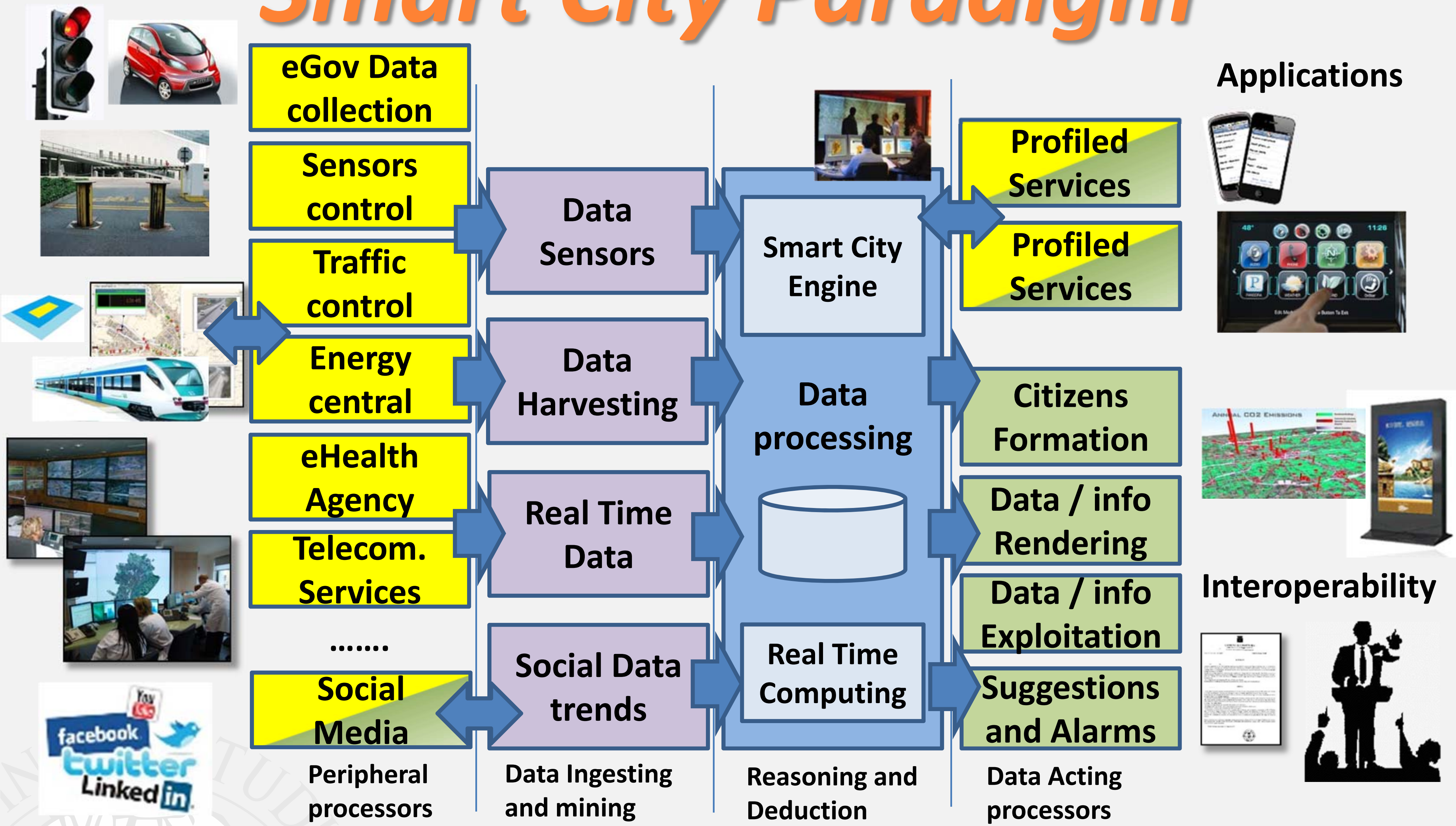
- Smart Health
- Smart Education
- Smart Mobility
- Smart Energy
- Smart Governmental
  - Smart economy
  - Smart people
  - Smart environment
  - Smart living
- Smart Telecommunication

- **Cities produce a HUGE amount of data every day**
  - **'Static' data**
    - Road graph
    - Bus/train graph
    - Services
    - ...
  - **Dynamic (real time) data**
    - Weather conditions
    - Traffic conditions
    - Pollution status
    - Bus/train positions
    - Parking status
    - People flows
    - ...
  - **Open/Private Data**

Smart-City

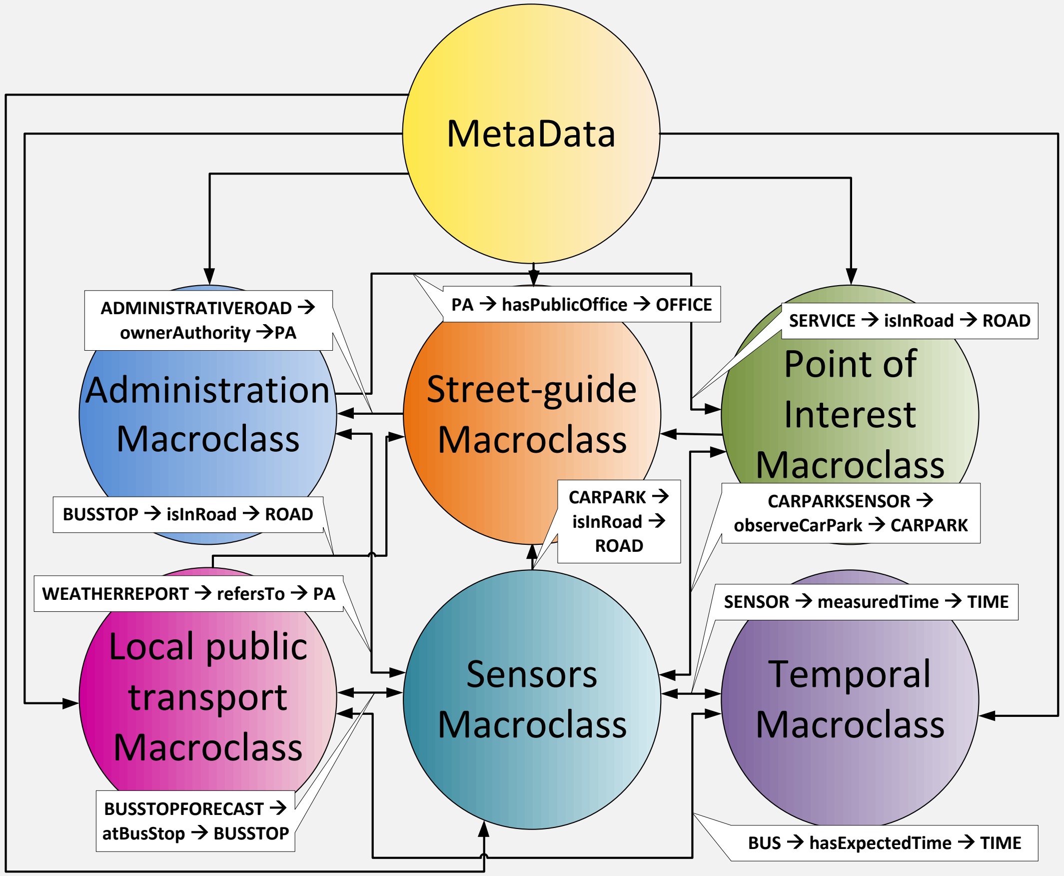
- ***Main Aim***
  - Provide a platform able to ingest and take advantage a large number of the above data, big data:
    - ***Exploit data integration and reasoning***
    - ***Deliver new services and applications to citizens,*** Leverage on the ongoing Semantic Web effort
- ***Problems & Challenges***
  - Data are provided in many different formats and protocols and from many different institutions, different convention and protocols, a different time, …. !
  - Data are typically not aligned (e.g., street names, dates, geolocations, tags, …). That is, they are n**ot semantically interoperable**
  - resulting a big data problem: volume, velocity, variability, variety, …..

Smart City Paradigm



Smart-city Ontology

- The data model provided have been mapped into the ontology, it covers different aspects:
  - Administration
  - Street-guide
  - Points of interest
  - Local public transport
  - Sensors
  - Temporal aspects
  - Metadata on the data



Smart-city Ontology

- ***Administration:*** structure of the general public administrations (*Municipality*, *Province* and *Region*) also includes *Resolutions* (ordinance issued by administrations, may change the viability, infrastructural works, schedule for RTZ, etc.)
- ***Street-guide:*** formed by entities as *Road*, *Node*, *RoadElement*, *AdministrativeRoad*, *Milestone*, *StreetNumber*, *RoadLink*, *Junction*, *Entry*, *EntryRule*, *Maneuver*,... represents the entire road system of the region, including the permitted maneuvers and the rules of access to the limited traffic zones. Based on OTN (Ontology of Transportation Networks) vocabulary
- ***Points of Interest***: includes all *Services*, activities, which may be useful to the citizen and who may have the need to search for and to arrive at, commercials, public administration, Cultural, ...

Smart-city Ontology

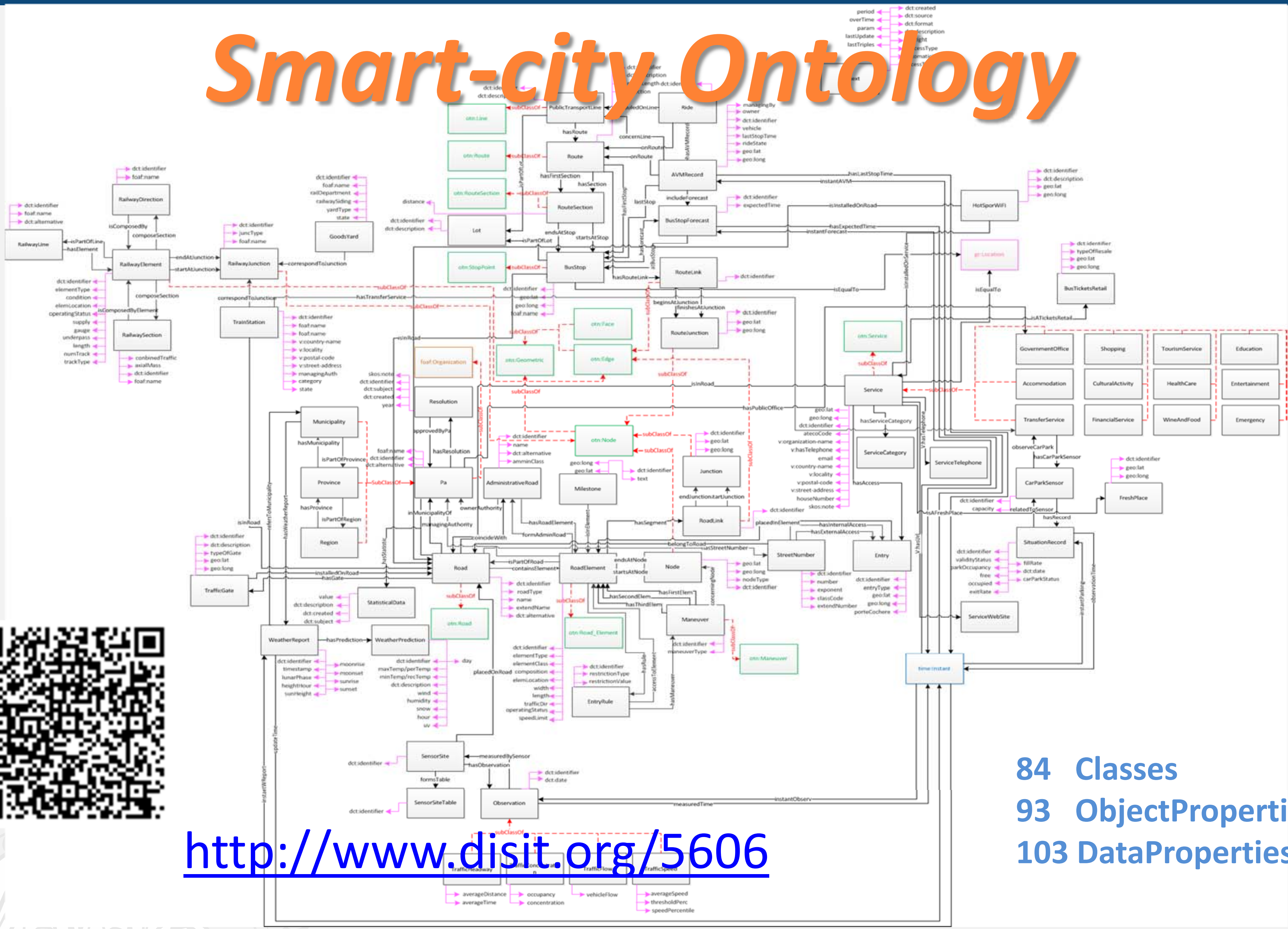
- ***Local public transport***: includes the data related to major local public transport companies as scheduled times, the rail graph, and data relating to real time passage at bus stops, real time position, ...
- ***Sensors***: data provided by sensors: currently, data are collected from various sensors (parking status, meteo, pollution) installed along some streets of Florence and surrounding areas, and from sensors installed into the main car parks of the region.
  – Plus: car sharing, bike sharing, AVM, RTZ, etc.
- ***Temporal***: that puts concepts related with time (time intervals and instants) into the ontology, so that associate a timeline to the events recorded and is possible to make forecasts. It uses time ontologies such as OWL-Time.

Smart-city Ontology

- **Metadata**: modeling the additional information associated with:
  - **Descriptor** of Data sets that produced the triples: data set ID, title, description, purpose, location, administration, version, responsible, etc..
  - **Licensing** information
  - **Process** information: IDs of the processes adopted for ingestion, quality improvement, mapping, indexing,.. ; date and time of ingestion, update, review, …;

  When a problem is detected, we have the information to understand when and how the problem has been included
- **Including basic ontologies as:**
  - DC: Dublin core, standard metadata
  - *OTN:* Ontology for Transport Network
  - *FOAF:* for the description of the relations among people or groups
  - *vCard:* for a description of people and organizations
  - *wgs84_pos:* for latitude and longitude, GPS info
  - OWL-Time: reasoning on time, time intervals
  - GoodRelations: commercial activities models

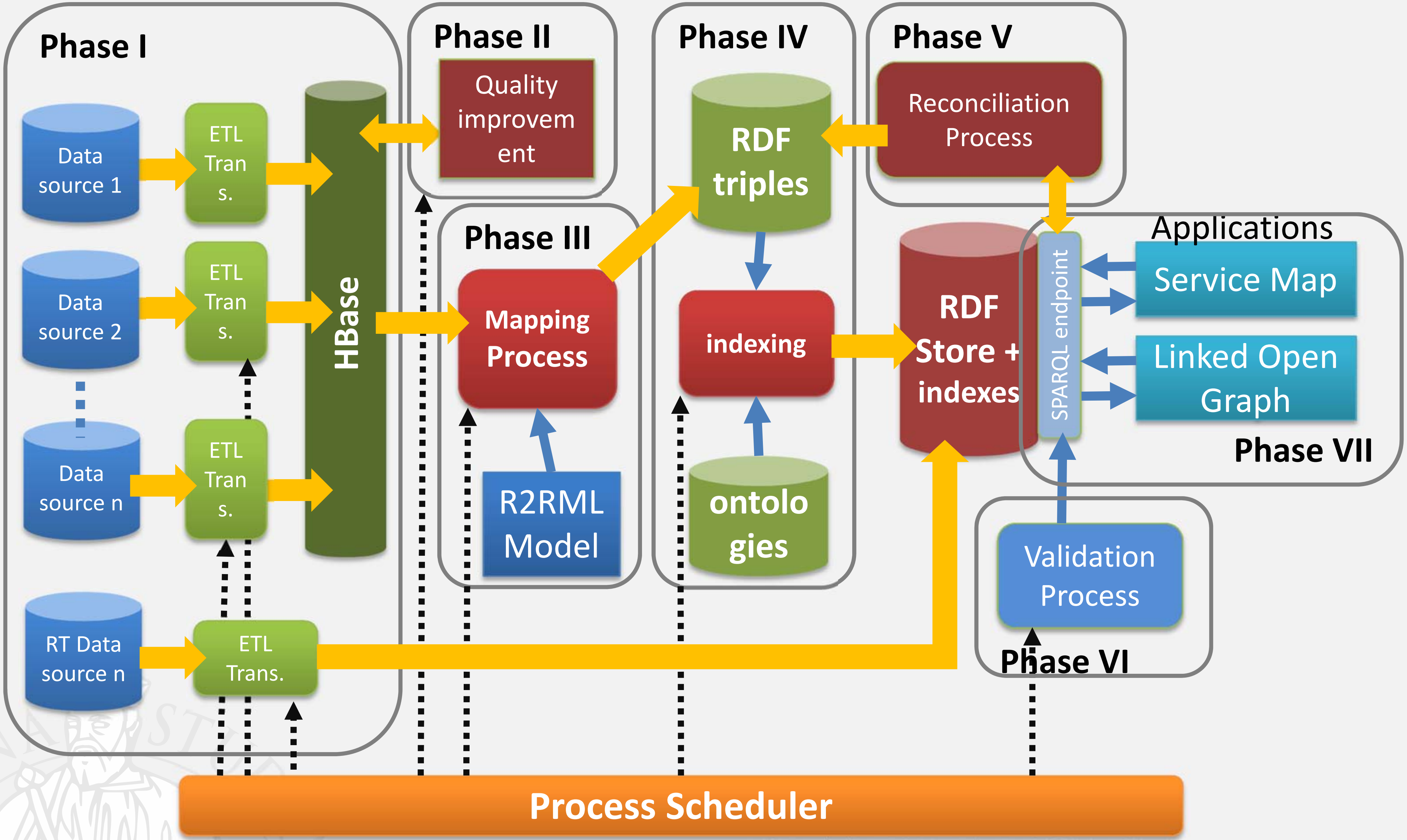
Smart-city Ontology



http://www.disit.org/5606

84 Classes
93 ObjectProperties
103 DataProperties

Data Engineering Architecture



Phase I - Data Ingestion

- **Ingesting a wide range of OD/PD**: public and private data, static, quasi static and/or dynamic real time data.
- For the case of Florence, we are addressing about **150 different data sources** of the 564 available, plus the regional, province, other municipalities, ….
- Using ***Pentaho - Kettle*** for data integration (Open source tool)
  - using specific ETL Kettle transformation processes (one or more for each data source)
  - data are stored in HBase (Bigdata NoSQL database)
- **Static and semi-static data** include: points of interests, geo-referenced services, maps, accidents statistics, etc.
  - files in several formats (SHP, KML, CVS, ZIP, XML, etc.)
- **Dynamic data** mainly data coming from sensors
  - parking, weather conditions, pollution measures, bus position, etc.
  - using Web Services.

Phase II - Data Quality Improvement

- **Problems kinds:**
  - Inconsistencies, incompleteness,..
- **Problems** on:
  - CAPs vs Locations
  - Street names (e.g., dividing names from numbers, normalize when possible)
  - Dates and Time: normalizing
  - Telephone numbers: normalizing
  - Web links and emails: normalizing
- **Partial Usage** of
  - Certified and accepted tables and additional knowledge
- **Enrichment** process may need several versions:
  - VIP names, GeoNames, etc..

Phase III - Data mapping

- Transforms the data from HBase to RDF triples
- Using **Karma Data Integration tool**, a mapping model from SQL to RDF on the basis of the ontology was created
  - Data to be mapped first temporarly passed from Hbase to MySQL and then mapped using Karma (in batch mode)
- The mapped data in triples have to be uploaded (and indexed) to the **RDF Store** (OpenRDF – sesame with OWLIM-SE)

Phase IV - Indexing

- **Periodic task** for reindexing: triples, text, space (GPS), dates, etc.
- **Indexing triples**: ontologies, all RDF files for OD, RT triples (from - to), reconciliation triples for OD, triples for enrichments, etc.
- ***If you do not index, you cannot identify all missing reconciliations***

Phase V - Data Reconciliation/alignment

- After the loading and indexing into the RDF store a dataset may be connected with the others **if entities refer to the same triples**
  - **Missed connections** strongly limit the usage of the knowledge base,
  - e.g. the services are not connected with the road graph.
- To associate each **Service** with a **Road** and an **Entity** on the basis of the street name, number and locality
- **It is not easy!** data coming from different sources

Phase V - Data Reconciliation/alignment

- **Examples**:
  - Typos;
  - Missing street number, or replaced with "0" or "SNC";
  - Municipalities with no official name (e.g. Vicchio/Vicchio del Mugello);
  - Street names and street numbers with strange characters (-, /, ° ? , Ang., ,);
  - Road name with words in a different order (e.g. Via Petrarca Francesco, exchange of name and surname);
  - Red street numbers (for shops);
  - Presence/absence of proper names in road name (e.g. via Camillo Benso di Cavour / via Cavour);
  - Number wrongly written (e.g. 34/AB, 403D, 36INT.1);
  - Roman numerals in the road name (e.g., via XXVII Aprile).

- **Steps**:
  1. *SPARQL Exact match* – match the strings as they are
  2. *SPARQL Enhanced Exact Match* – make some substitutions (Via S. Marta → Via Santa Marta, ...)
  3. *Last Word Search* – use only the last word of street name
  4. Use Google GeoCoding API
  5. Remove ‘strange chars’ (-, /, °, ? , Ang., ,) from Street number
  6. Remove ‘strange chars’ from Street name
  7. Rewrite wrong municipality names

Phase V - Data Reconciliation/alignment

Comparing different reconciliation approaches based on

- SILK link discovering language
- SPARQL based reconciliation described above

Method	Precision	Recall	F1
SPARQL –based reconciliation	1,00	0,69	0,820
SPARQL -based reconciliation + additional manual review	0,985	0,722	0,833
Link discovering - Leveisthein	0,927	0,508	0,656
Link discovering - Dice	0,968	0,674	0,794
Link discovering - Jaccard	1,000	0,472	0,642
Link discovering + heuristics based on data knowledge + Leveisthein	0,925	0,714	0,806

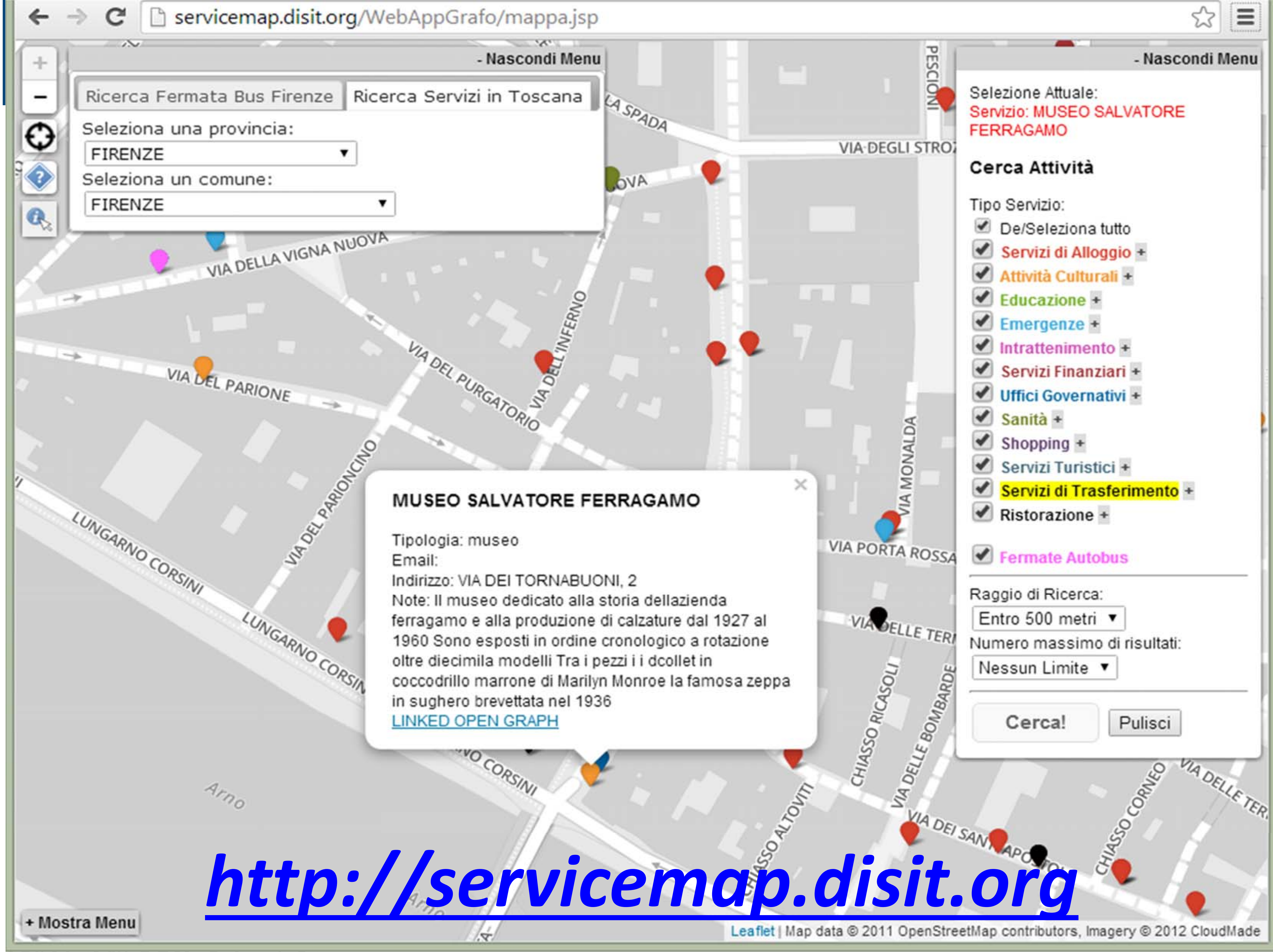
Thus automation of reconciliation is possible and produces acceptable results!!

Phase VI - Validation

- A set of queries applied automatically to verify the consistency and completeness, after new re-indexing and new data integration
  – I.e.: the KB regression testing!!!!!

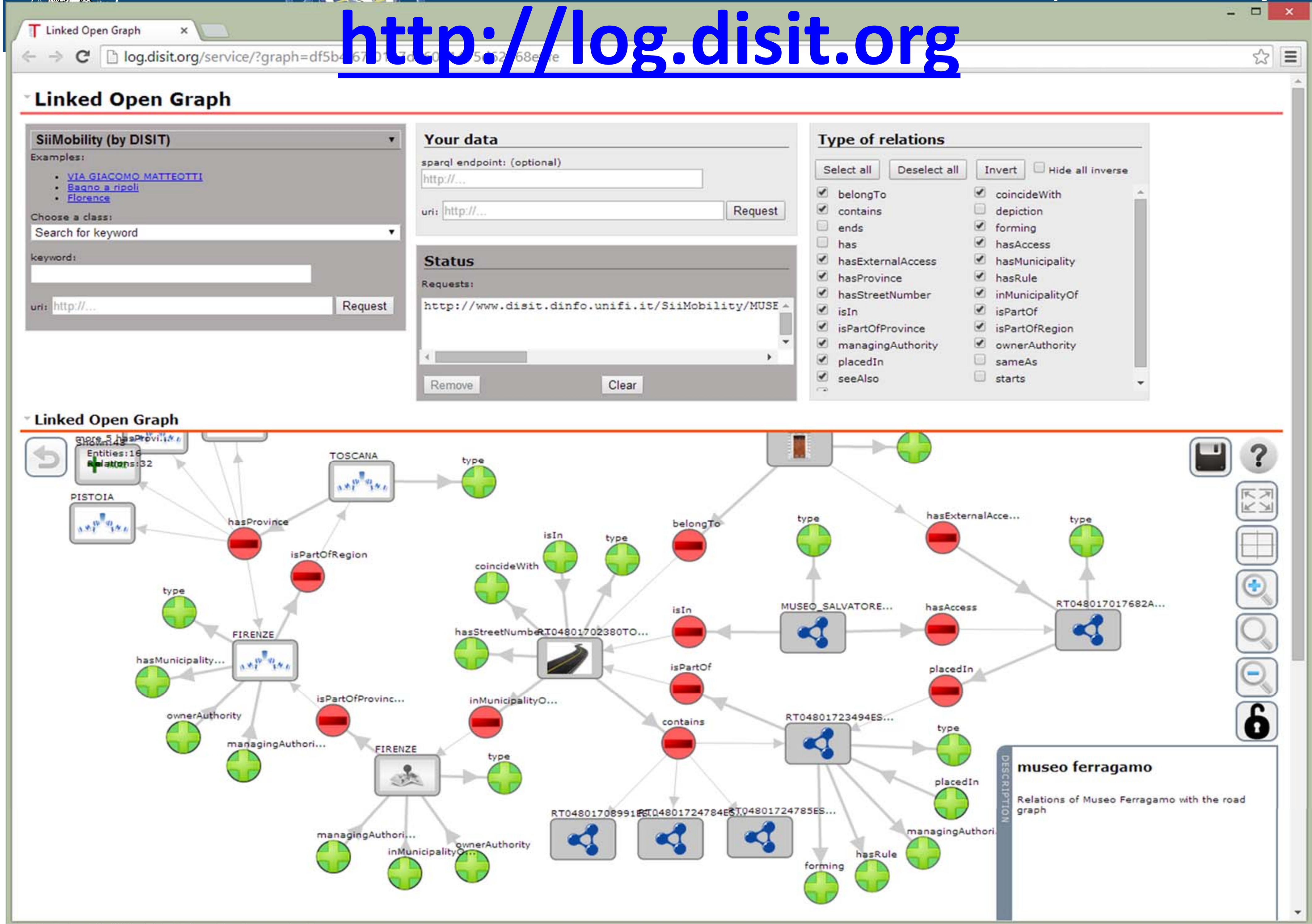
Phase VII - Data access

- Applications can access the data using the SPARQL endpoint, currently we have two applications:
  – ServiceMap (http://servicemap.disit.org) for a map based application
  – Linked Open Graph (http://log.disit.org) for browsing the data from SPARQL/Linked Data sources



http://servicemap.disit.org

http://log.disit.org



Conclusions

- ***Developed***
  - Smart-city Ontology as conceptual model for reasoning
  - platform for smart-city data ingestion and semantic interoperability processes as big data tools
  - Assessment demonstrated that automated reconciliation is possible
- Future/Ongoing activities
  - Improvement of data alignment and cleaning
  - Definition of languages and tools for reasoning
- It will be used in ***Sii-Mobility*** **project:**
  - Adding prediction algorithms
  - Adding user-generated information
  - Adding more applications using the data

References

- Caragliu, A., Del Bo, C., Nijkamp, P. (2009), Smart cities in Europe, 3rd Central European Conference in Regional Science – CERS, Kosice (sk), 7-9 ottobre 2009.
- Bellini P., Di Claudio M., Nesi P., Rauch N., "Tassonomy and Review of Big Data Solutions Navigation", Big Data Computing To Be Published 26th July 2013 by Chapman and Hall/CRC
- Vilajosana, I. ; Llosa, J. ; Martinez, B. ; Domingo-Prieto, M. ; Angles, A., "Bootstrapping smart cities through a self-sustainable model based on big data flows", Communications Magazine, IEEE, Vol.51, n.6, 2013
- Ontology of Trasportation Networks, Deliverable A1-D4, Project REWERSE, 2005 http://rewerse.net/deliverables/m18/a1-d4.pdf
- Pan, Feng, and Jerry R. Hobbs. "Temporal Aggregates in OWL-Time." In FLAIRS Conference, vol. 5, pp. 560-565. 2005.
- Embley, David W., Douglas M. Campbell, Yuan S. Jiang, Stephen W. Liddle, Deryle W. Lonsdale, Y-K. Ng, and Randy D. Smith. "Conceptual-model-based data extraction from multiple-record Web pages." Data & Knowledge Engineering 31, no. 3 (1999): 227-251.
- Auer, Sören, Jens Lehmann, and Sebastian Hellmann. "Linkedgeodata: Adding a spatial dimension to the web of data." In The Semantic Web-ISWC 2009, pp. 731-746. Springer Berlin Heidelberg, 2009.
- Andrea Bellandi, Pierfrancesco Bellini, Antonio Cappuccio, Paolo Nesi, Gianni Pantaleo, Nadia Rauch, ASSISTED KNOWLEDGE BASE GENERATION, MANAGEMENT AND COMPETENCE RETRIEVAL, International Journal of Software Engineering and Knowledge Engineering, Vol.22, n.8, 2012
- Apache HBase: A Distributed Database for Large Datasets. The Apache Software Foundation, Los Angeles, CA. URL http://hbase.apache.org.
- Pentaho Data Integration, http://www.pentaho.com/product/data-integration
- Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov, “OWLIM: A family of scalable semantic repositories”, Semantic Web Journal, Volume 2, Number 1 / 2011.
- S.Gupta, P.Szekely, C.Knoblock, A.Goel, M.Taheriyan, M.Muslea, "Karma: A System for Mapping Structured Sources into the Semantic Web", 9th Extended Semantic Web Conference (ESWC2012).
- A. Ngomo, S. Auer. “LIMES: a time-efficient approach for large-scale link discovery on the web of data”. Proc. of the 22nd int. joint conf. on Artificial Intelligence, Vol.3. AAAI Press, 2011.
- R. Isele, C. Bizer. “Active learning of expressive linkage rules using genetic programming”. Web Semantics: Science, Services and Agents on the World Wide Web 23 (2013): pp.2-15.
- Powers, D.M.W. (February 27, 2011). "Evaluation from precision, recall and F-Measure to roc informedness, markedness and correlation". Journal of Machine Learning Technologies 2 (1): 37–63.

Thank you!

http://www.disit.org/5606

http://www.disit.dinfo.unifi.it

Paolo Nesi
Dipartimento di Ingegneria dell'Informazione, DINFO
Università degli Studi di Firenze
Via S. Marta 3, 50139, Firenze, Italy
Tel: +39-055-4796567, fax: +39-055-4796363
DISIT Lab
http://www.disit.dinfo.unifi.it *alias* http://www.disit.org
paolo.nesi@unifi.it