Ingestion tools for open data (via files and services) towards DISIT Ontology for Smart City

Prof. Paolo Nesi
Department of Information Engineering
University of Florence
Via S. Marta 3, 50139, Firenze, Italy
tel: +39-055-4796523, fax: +39-055-4796363
DISIT Lab
http://www.disit.dinfo.unifi.it alias http://www.disit.org
paolo.nesi@unifi.it , http://www.disit.org/nesi

Contesto: Smart City

Sii-Mobility promuove, nell'ambito della mobilità urbana su gomma e/o su rotaia, lo sviluppo di nuove tecnologie e soluzioni ICT innovative finalizzate a migliorare l'interoperabilità dei sistemi di gestione informativi e di infomobilità urbana e metropolitana.

Collabora nel supporto fra OUC/ASL, associazioni e famiglie con disabili

Sii-Mobility

- **Title**: Support of Integrated Interoperability for Services to Citizens and Public Administration
- **Duration**: 36 months
- **Cost**: 22 Meuro
- **Objectives**:
  1. Reduction of social costs of mobility;
  2. Simplify the use of mobility systems;
  3. Developing working solutions and application, with testing methods;
  4. Contribute to standardization organs, and establishing relationships with other smart cities' management systems.

The Sii-Mobility platform will be capable to provide support for SME and Public Administrations. Sii-Mobility consists in a federated/integrated interoperable solution aimed at enabling a wide range of specific applications for private services to citizen and commercial services to SME.

- **Link**: http://www.disit.dinfo.unifi.it/siimobility.html


Commenti dei cittadini,
Social Media
Merci
AVM trasporto
Pubblico
Sensori su
trasporto Privato
Sensori,
sistema monitoraggio
Sensori
Parcheggi
UTC
Sii-Mobility
Infomobility
Monitoraggio
traffico, autostrade
Varchi
Telematici, ZTL
Emergenze,
polizia, 118
Rete
Ferroviaria
Parametri
ambientali
Servizi ed
enti
Ordinanze: eventi,
lavori pubblici, ..

Coll@bora

- **Title**: Collaborative Support for Parents and Operators of Disabled
- **Duration**: 24 months
- **Cost**: 1 Meuro
- **Objectives**: providing strong advantages for
  1. Relatives interested in facilitating relations with the management team;
  2. Associations in order to offer a better service to the families and people with disabilities by providing a collaborative support to the involved teams, but also to manage the wealth of knowledge, to support the training of the staff, etc.

  Coll@bora provides a secure collaboration tool for the teams and for the association to support the families and the disabled people.
- **Link**: http://www.disit.dinfo.unifi.it/collabora.html

Obiettivi (una parte)

- Analisi dei dati in tempo reale messi a disposizione dalla Regione Toscana
- Creazione di una fase di ingestion per l'acquisizione automatica dei dati
- Definizione di procedure dedicate alla trasformazione dei dati acquisiti in triple RDF e memorizzazione su database semantico
- Riconciliazione dei dati raccolti con quelli relativi al grafo stradale
- Creazione di una base di dati RDF con informazioni provenienti da varie fonti
- Creazione di una fase di ingestion per l'acquisizione automatica delle informazioni
- Interrogazione del database per la verifica dei collegamenti tra i diversi elementi acquisiti e facenti parte del grafo stradale

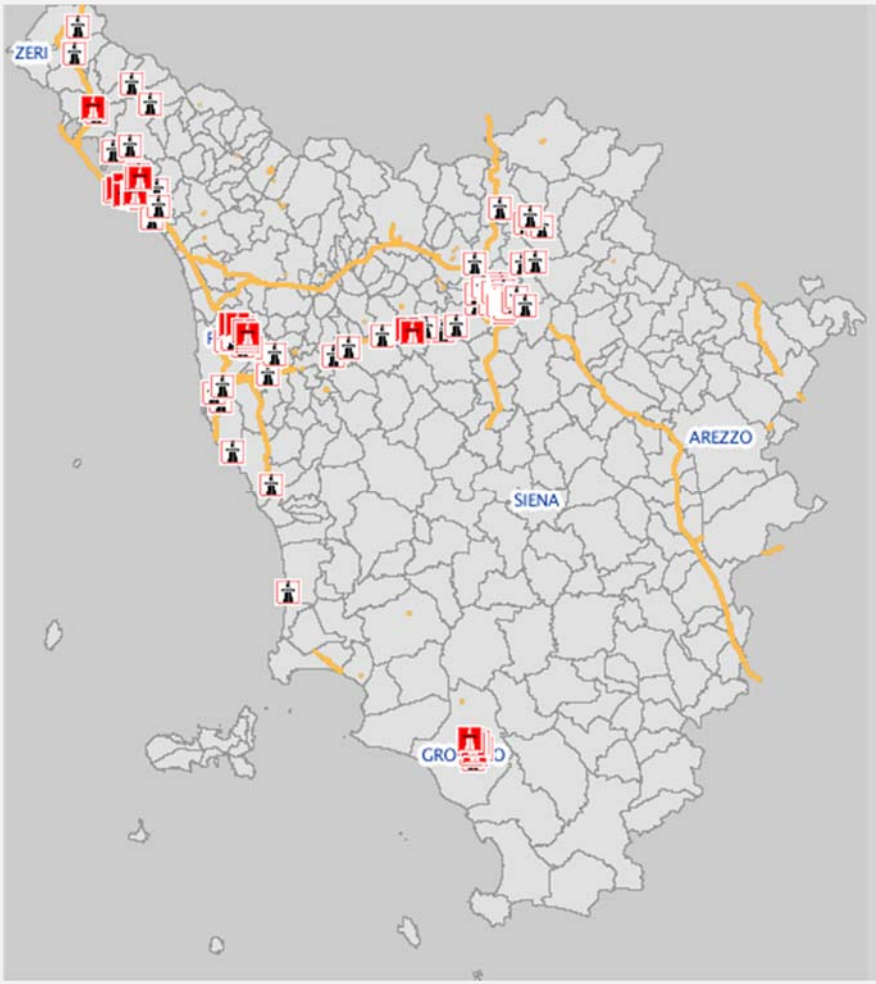
Analisi dei dati (per mobilita' ed attuali)

- I dati utilizzati per questo lavoro sono quelli messi a disposizione dal **MIIC** (Mobility Information Integration Center), un progetto della Regione Toscana che si occupa della raccolta, da enti federati, di dati relativi alla infomobilità e della loro distribuzione mediante web services.

- I web services espongono dati relativi a: traffico, stato dei parcheggi, AVM (Automatic Vehicle Monitoring), emergenze viarie e informazioni meteo. Queste ultime due tipologie sono state escluse dal lavoro per carenza di dati.

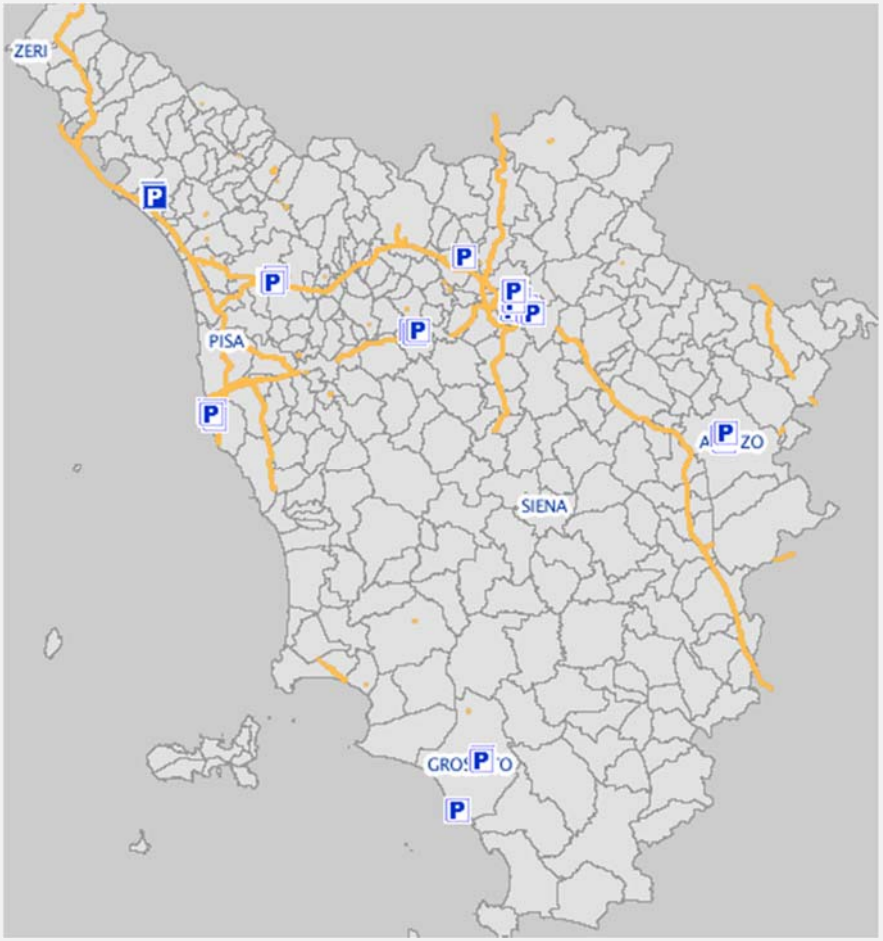
Analisi dei dati attuali MIIC

- **Sensori traffico**: dati relativi alla situazione della viabilità stradale provenienti da gestori di sistemi di rilevamento a sensori
  - Le misurazioni comprendono dati come distanza media tra veicoli, velocità media di transito, percentuale di occupazione della strada, transiti orari, ecc.
  - I sensori sono divisi in gruppi, ciascuno identificato da un codice di catalogo che viene passato come parametro in fase di invocazione del web service
  - Un gruppo è un insieme di sensori che monitora un tratto stradale
  - I gruppi producono una misurazione ogni 5 o ogni 10 minuti
  - Attualmente sono attivi 71 gruppi su tutto il territorio toscano, per un totale di 126 sensori



Analisi dei dati attuali MIIC

- **Parcheggi**: dati relativi allo stato di occupazione dei parcheggiprovenienti da gestori di aree di parcheggio.
  - Lo stato di un parcheggio viene descritto da dati come il numero di posti totali e occupati, il numero di veicoli in entrata e in uscita, ecc.
  - I parcheggi sono divisi in gruppi, ciascuno identificato da un codice di catalogo che viene passato come parametro in fase di invocazione del web service
  - Un gruppo corrisponde all'insieme di parcheggi appartenenti a un comune
  - La situazione di ogni parcheggio viene pubblicata circa ogni minuto
  - Attualmente vengono monitorati 50 parcheggi



Analisi dei dati, AVM locata in RT

- **AVM**: dati in tempo reale relativi ai mezzi del trasporto pubblico locale dell'area fiorentina dotati di dispositivi AVM
  - Monitora lo stato delle corse attive sul territorio, dove una corsa è il percorso che esegue un mezzo da un capolinea di inizio a uno di fine
  - I dati forniti sono relativi allo stato di ritardo o anticipo di un mezzo, posizione del mezzo in coordinate GPS, informazioni sull'ultima fermata effettuata e su quelle programmate, ecc.
  - Il web service viene invocato passando come parametro il codice identificativo della corsa
  - I dispositivi AVM inviano due tipi di messaggio, uno in corrispondenza di un orario programmato, solitamente ogni minuto, e uno in corrispondenza di un evento notevole come l'arrivo a una fermata, la partenza da un capolinea o l'interruzione del servizio
  - Attualmente vengono monitorate le corse relative a 8 linee del trasporto pubblico di Firenze

- **Dati statici**: dati aggiornati raramente che arricchiscono quelli in tempo reale

  - Vengono forniti dall'Osservatorio Regionale per la Mobilità ed i Trasporti mediante un portale con interfaccia grafica, al momento non è disponibile un web service che permetta il dowload automatico
  - Informazioni sulla geolocalizzazione dei parcheggi e dei sensori censiti dal MIIC (da riconciliare)
  - Dettagli aggiuntivi sulla rete di trasporto pubblico: descrizione e dettagli di linee, percorsi e fermate, geolocalizzazione con coordinate Gauss-Boaga delle fermate (da riconciliare)

Altri dati di origine

- Previsioni meteo fornite dal consorzio LaMMA
  - Formato XML
  - Informazioni sul giorno attuale: ora tramonto/alba sole, ora tramonto/alba luna, altezza sole …
  - Previsioni sul giorno attuale e sui 4 successivi: temperatura massima, temperatura percepita, pioggia, umidità, potenza raggi uv
  - Previsioni su cinque momenti della giornata: mattina, pomeriggio …
  - Un file per ognuno dei 286 comuni aggiornato 2 volte al giorno
- Servizi della regione Toscana
  - Formato CSV
  - Servizi di varia natura: banche, scuole, enogastronomia, ospedali, negozi, teatri, musei …
  - Localizzati geograficamente da indirizzo (via, numero civico) e comune di appartenenza
  - Contengono il nome del servizio, indirizzo, comune, provincia, tipologia di servizio, numero telefono, email …
  - 24 file contenenti 28560 servizi diversi

Altri Dati di origine

❑ Statistiche sul comune di Firenze

- Formato CSV
- Contengono informazioni sul comune e sulle vie di Firenze: sinistri per via, arrivi turistici, veicoli circolanti …
- Le statistiche riguardano gli ultimi cinque anni
- Analizzati sei diversi file contenenti 8100 statistiche

❑ Linea tram

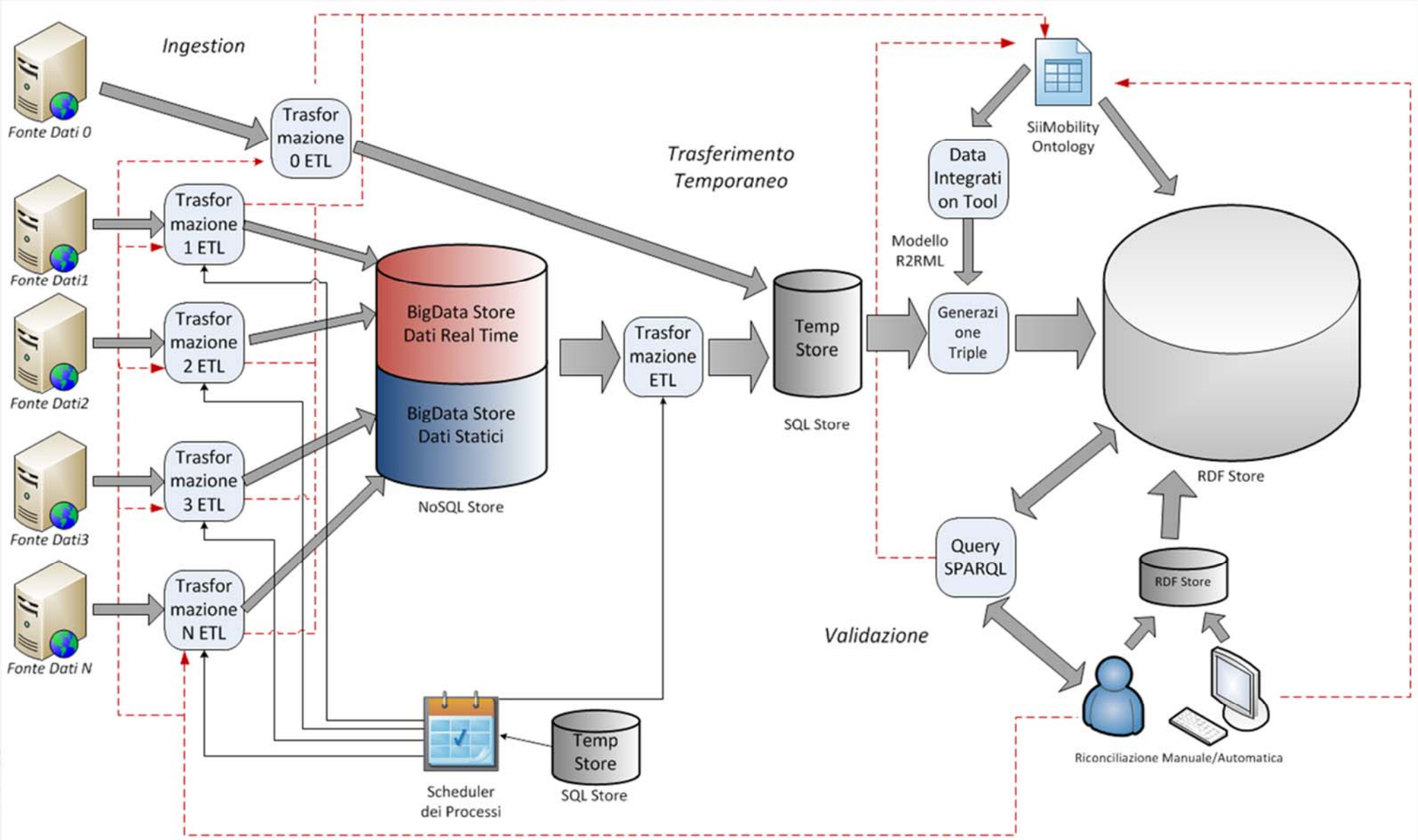
- Formato KMZ
- Contiene file in formato KML

usato per gestire dati geospaziali nei programmi Google earth e Google maps

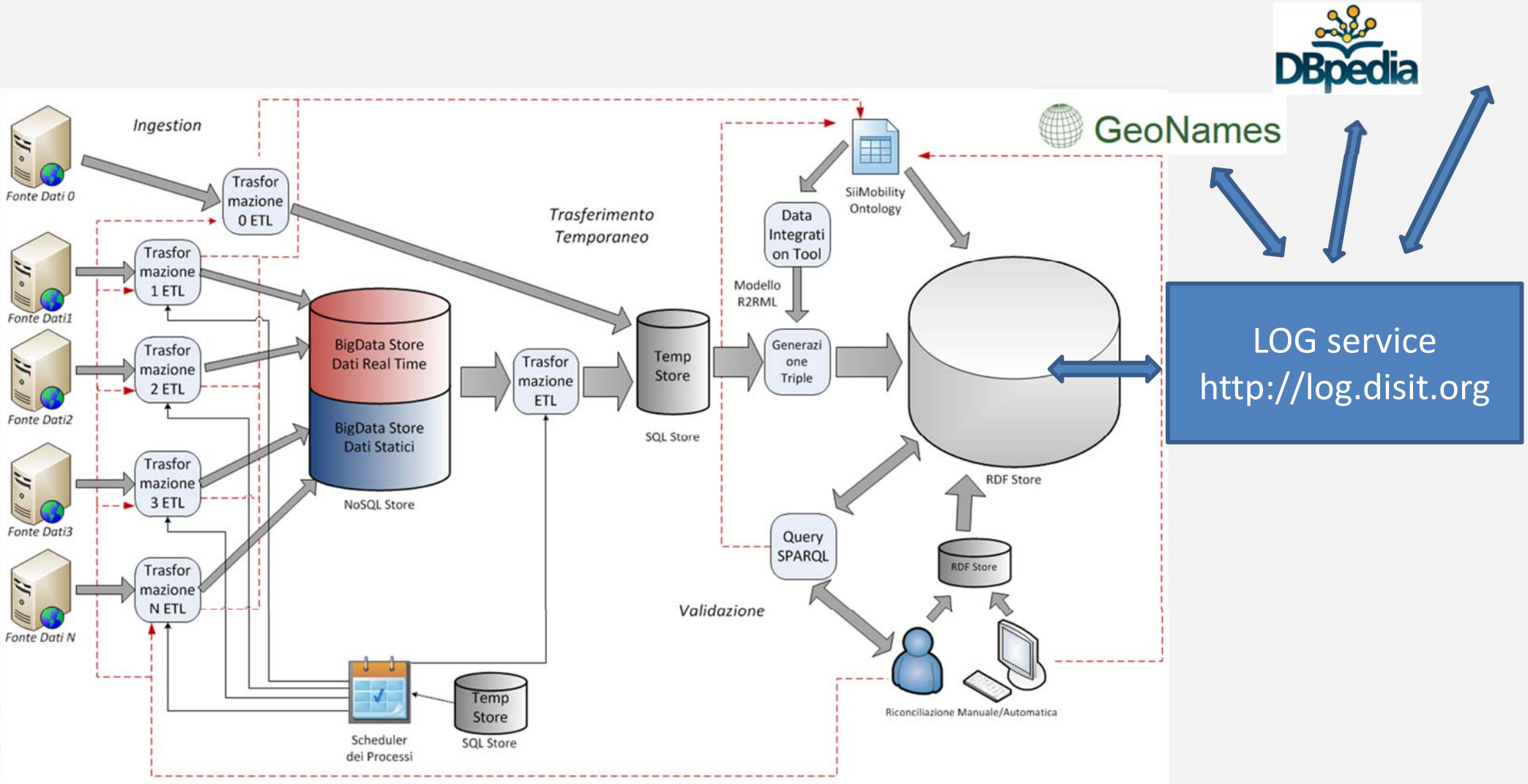
- Contiene le coordinate che

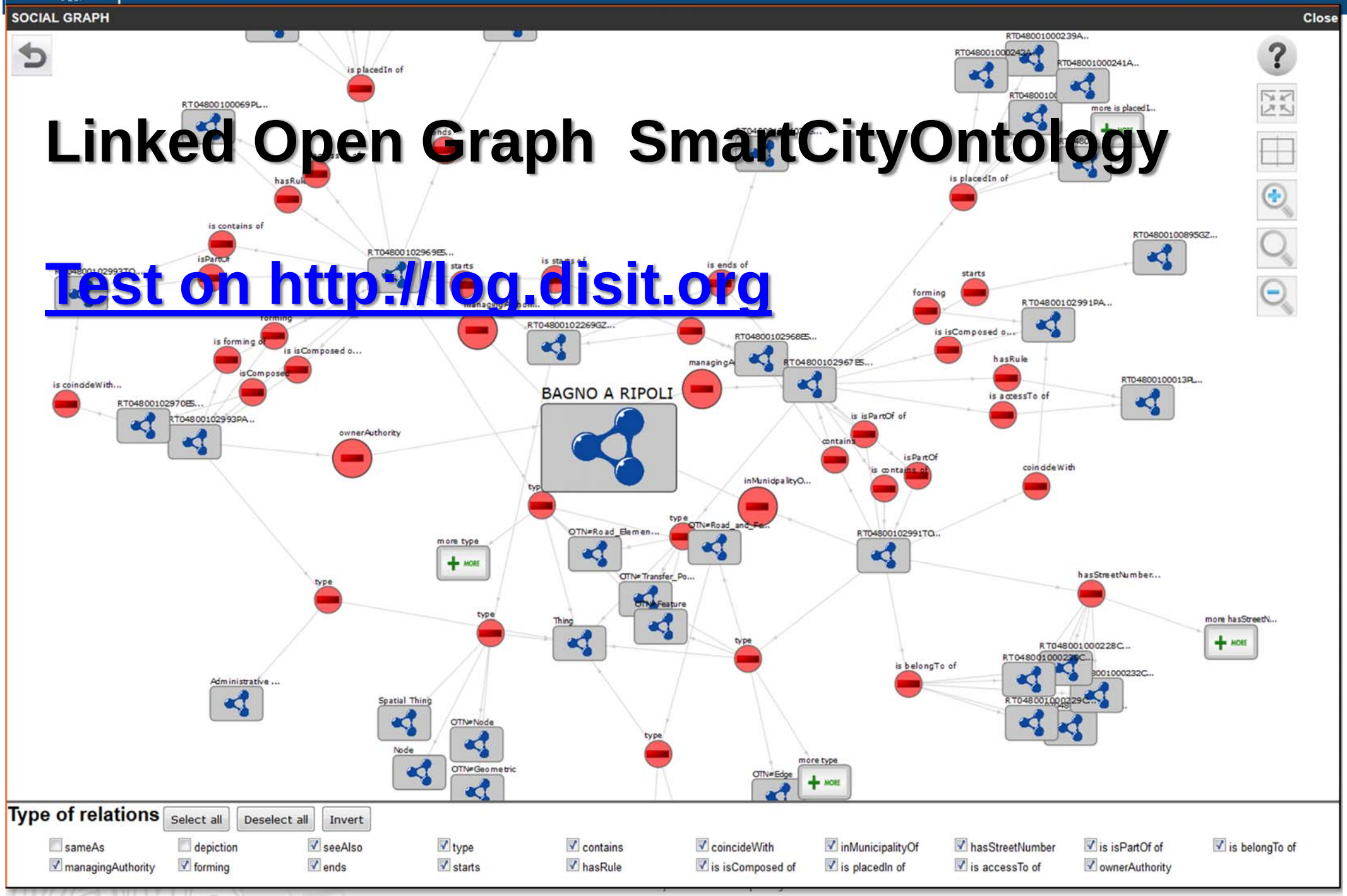
compongono il percorso coperto dalla linea tram

Architettura del sistema



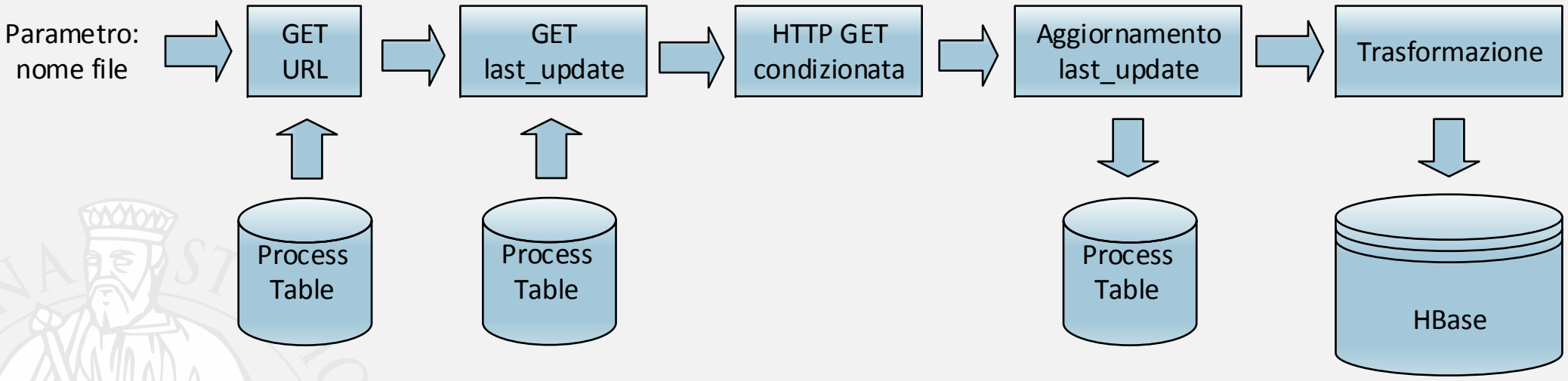
From Ingestion to LOG service



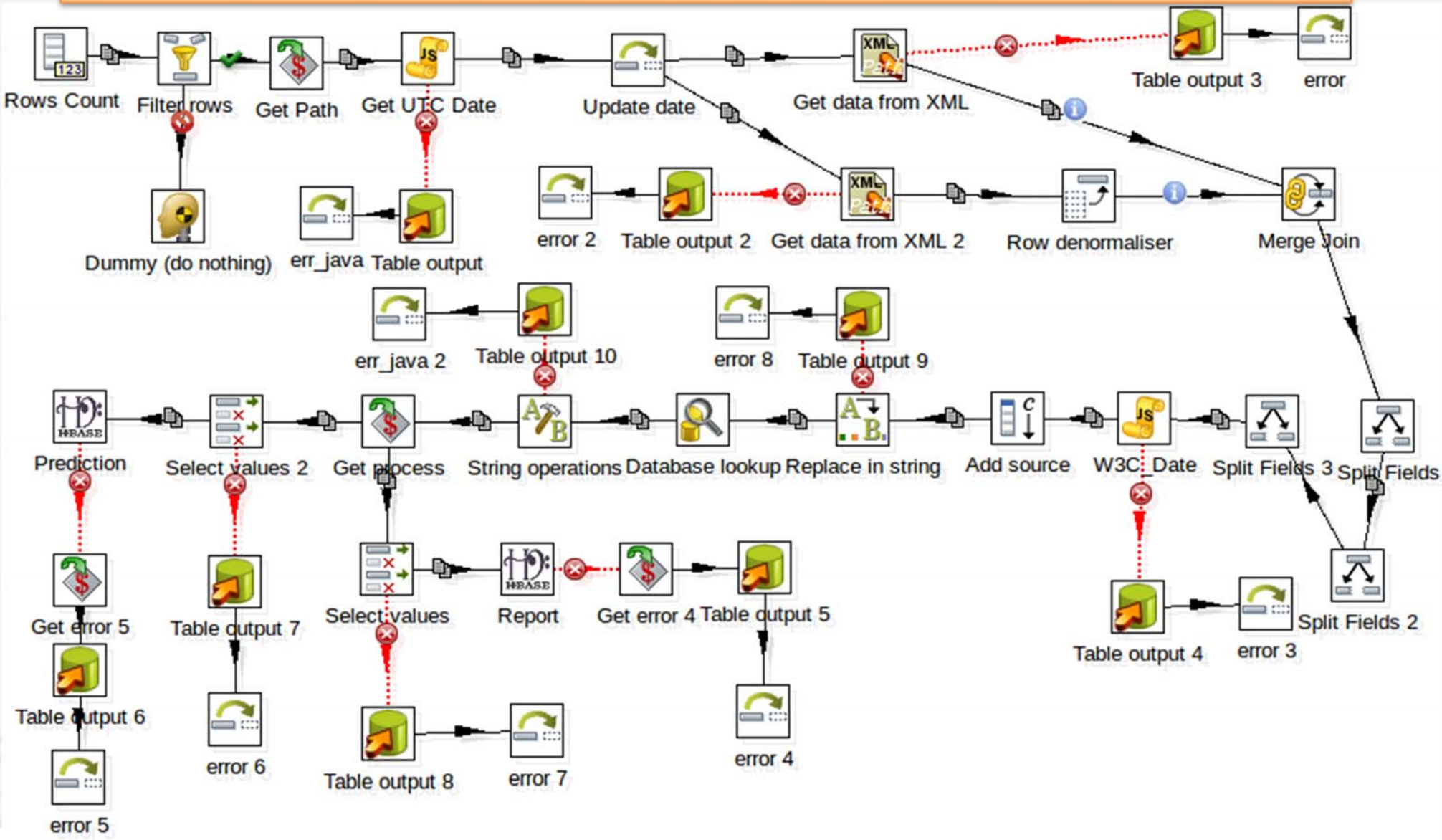

SOCIAL GRAPH
Close
Linked Open Graph SmartCityOntology
Test on http://log.disit.org
BAGNO A RIPOLI
is placedIn of
is contains of
isPartOf
starts
is ends of
forming
is forming of
is isComposed o...
isComposed
is coincideWith...
ownerAuthority
managingA...
is isPartOf of
contains
isPartOf
is contains of
coincideWith
inMunicipalityO...
hasRule
is accessTo of
more is placedI...
more type
type
Thing
Administrative ...
Spatial Thing
Node
OTN=Node
OTN=Geometric
OTN=Edge
OTN=Transfer_Po...
OTN=Feature
hasStreetNumber...
more hasStreetN...
is belongTo of
Type of relations
Select all
Deselect all
Invert
sameAs
depiction
seeAlso
type
contains
coincideWith
inMunicipalityOf
hasStreetNumber
is isPartOf of
is belongTo of
managingAuthority
forming
ends
starts
hasRule
is isComposed of
is placedIn of
is accessTo of
ownerAuthority

Ingestion, un esempio

- Eseguita da Pentaho Kettle (software ETL)
  1) Si passa il nome del file da processare
  2) Viene recuperato l'URL della risorsa
  3) Download del file (solo se è stato aggiornano dall'ultima volta)
  4) Aggiornamento della data di download
  5) Trasformazione risorsa: da XML a forma tabulare, eliminazione caratteri speciali, creazione chiavi, informazioni aggiuntive …
  6) Memorizzazione su HBase

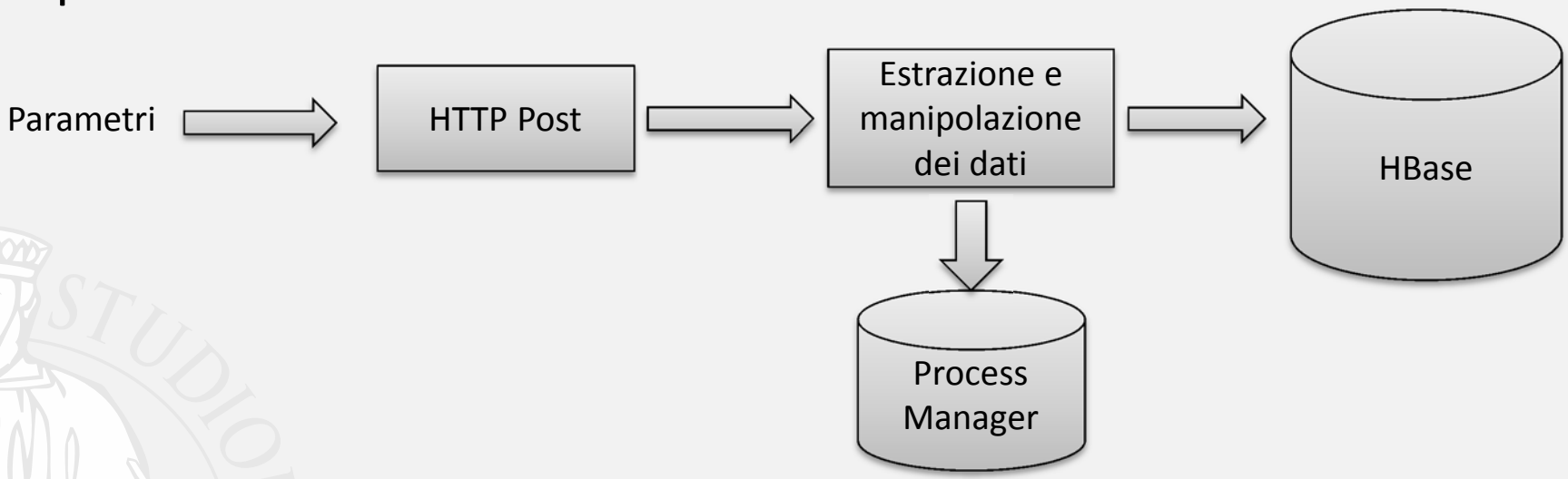


Ingestion (un esempio)

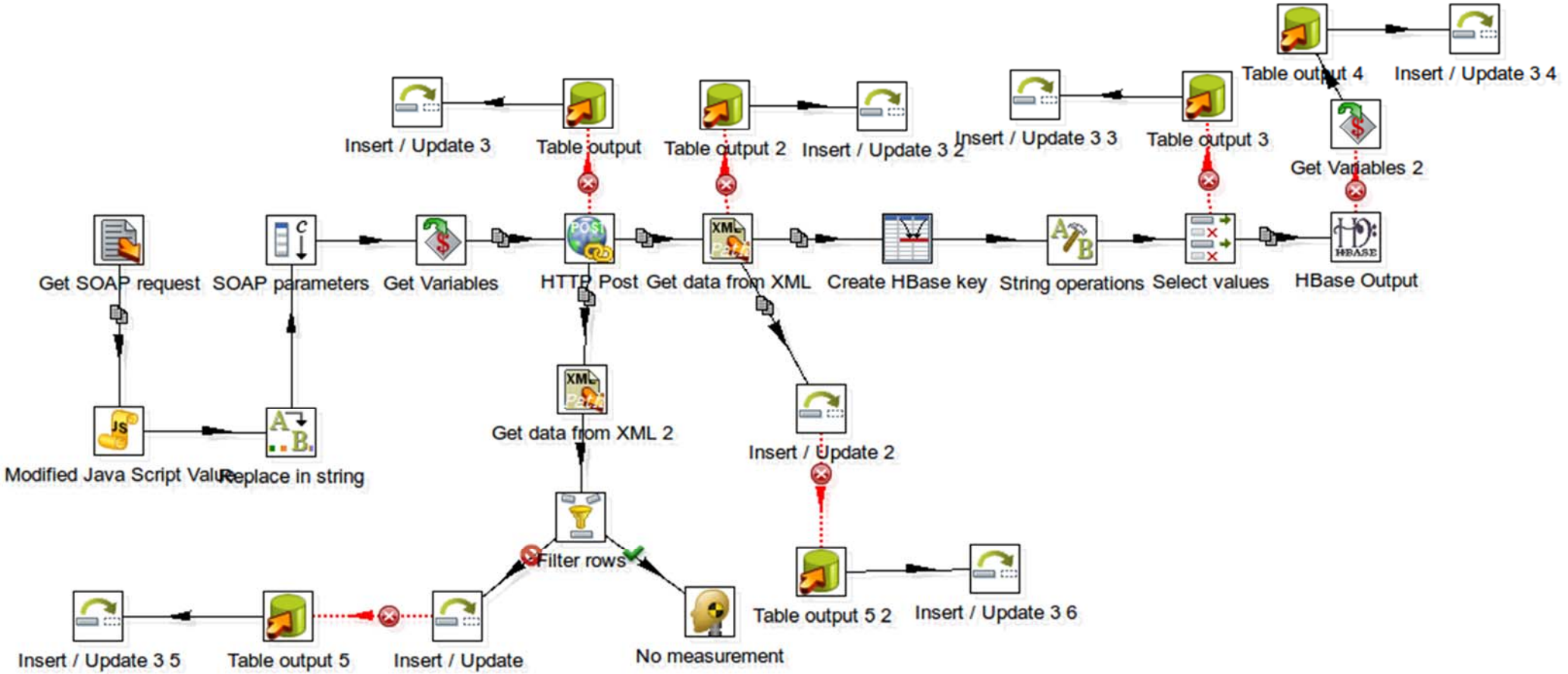


Fase A (ingestion)

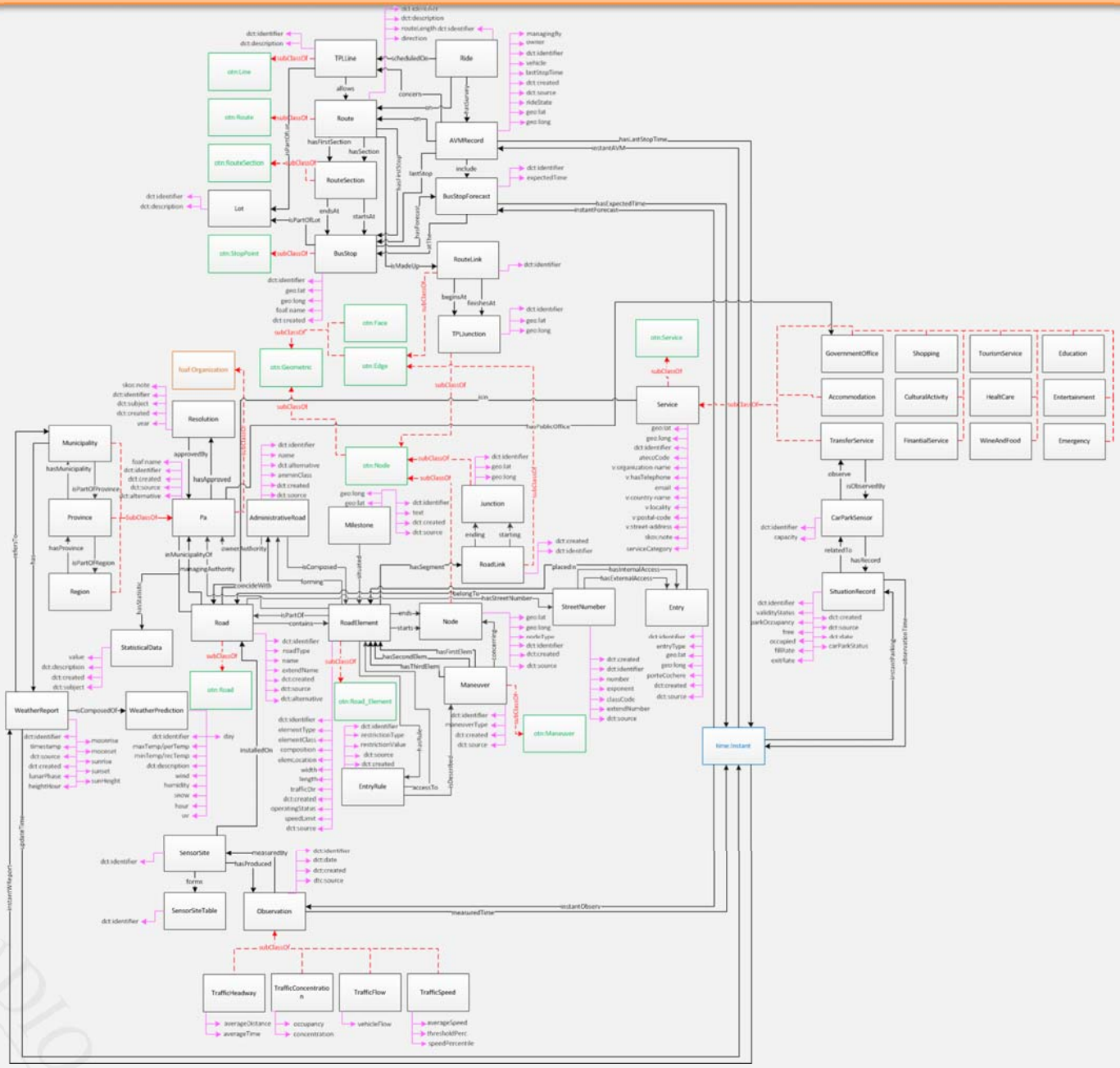
- Eseguita con Pentaho Kettle (software ETL)
- In fase di lancio, alla trasformazione vengono passati i parametri necessari
- Il web service viene invocato con un HTTP Post
- Dal risultato, in formato XML, vengono estratti con XPath i campi utili
- I campi vengono manipolati a seconda delle necessità dettate dal tipo di dato
- Viene aggiornata la tabella MySQL dei processi con il timestamp relativo all'ultima versione dei dati
- La procedura termina con la memorizzazione dei dati su HBase



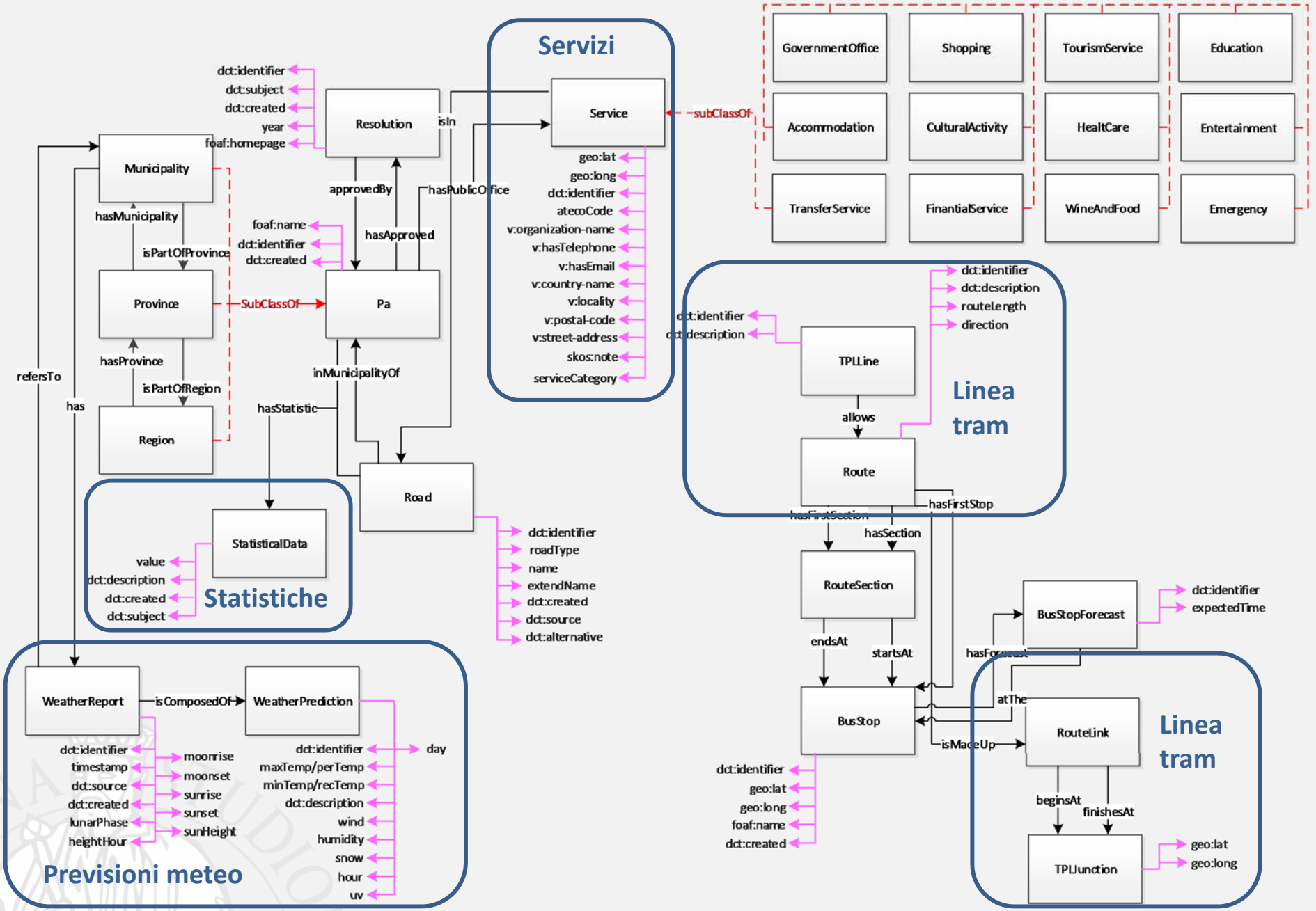
Fase A (ingestion)



DISIT Ontologia Smart City

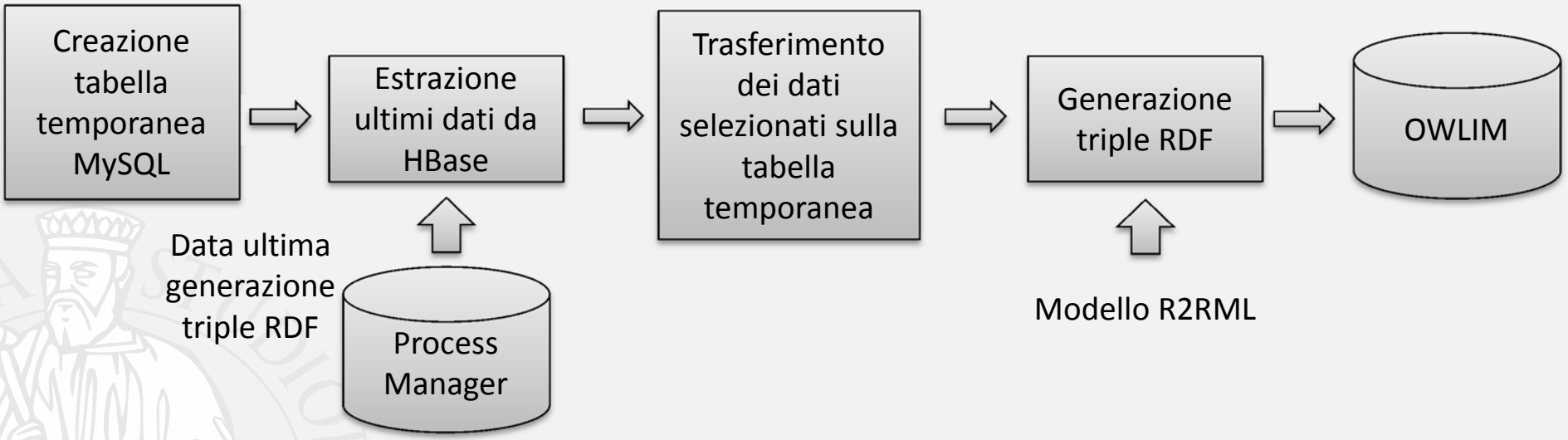


Ontologia (parte di)



Fase B

- Anche questa fase viene gestita da Pentaho Kettle
- Da HBase vengono selezionate le righe con data di inserimento maggiore rispetto a quella dell'ultima generazione di triple RDF
- Questi dati vengono scritti su una tabella temporanea MySQL
- Viene invocata la funzione di Karma Data Integration che trasforma dati in forma tabulare nella forma di triple RDF, utilizzando un modello R2RML
- Le triple generate vengono caricate su OWLIM



Riconciliazione

- Consiste nell'assicurarsi che i dati siano conformi all'ontologia e ai dati inseriti nello stesso database da altre fonti
- Avviene sia in fase di ingestion con la manipolazione dei dati (es: eventuale conversione date in formato W3C dateTime) sia successivamente con l'aggiunta di nuove triple RDF
- Per quanto riguarda i dati di **sensori** e **parcheggi**, di cui non viene fornita la geolocalizzazione, è stato necessario individuare il codice del toponimo corrispondente a ciascun indirizzo fornito dal MIIC.
- I dati in tempo reale AVM non hanno richiesto riconciliazione in quanto geolocalizzati con coordinate GPS

ES. Riconciliazione

- A volte risulta necessario aggiungere informazioni ai dati in modo da garantire il collegamento tra i vari elementi dell'ontologia

- Alle informazioni relative ai dati statistici e previsioni meteo si aggiungono, prima della generazione delle triple, le chiavi delle entità del grafo stradale

- Nei servizi la riconciliazione avviene dopo la generazione delle triple

Riconciliazione servizi

- Nel caso della riconciliazione dei servizi si deve fare in modo di mantenere le informazioni relative agli indirizzi scritti in maniera differente rispetto a quelli sul grafo stradale

- Collegamento tra gli oggetti di tipo *Road* e le entità di tipo *Service*
  - Creazione di un nuovo oggetto *Road* (contenente il nuovo indirizzo)
  - *Legame di uguaglianza tra il nuovo oggetto e l'elemento Road corrispondente (owl:sameAS)*
  - *Relazione SiiMobility:isIn* tra il nuovo oggetto e l'elemento *Service*

ES. Riconciliazione servizi

Service_Key	Road_Key	R_address	extendName	Cod_ toponimo
BELLAVISTA-Largo_F.lli_Alinari_15	048017LargoF.lliAlinari	Largo F.lli Alinari	LARGO FRATELLI ALINARI	RT04801701866TO
CASA_DEL_LAGO-Lungarno__A._Vespucci_58	048017LungarnoA.Vespucci	Lungarno A. Vespucci	LUNGARNO AMERIGO VESPUCCI	RT04801701874TO
COSMOPOLITAN-Via_F._Baracca187	048017ViaF.Baracca	Via F. Baracca	VIA FRANCESCO BARACCA	RT04801702987TO
SAN_PAOLO_IMI-VIA_DE'_VECCHIETTI_22/R	048017VIADE'VECCHIETTI	VIA DE' VECCHIETTI	VIA DEI VECCHIETTI	RT04801702383TO
CREDITO_ARTIGIANO-VIA_DE'_BONI_1	048017VIADE'BONI	VIA DE‘ BONI	VIA DEI BONI	RT04801702326TO
Auditorium_al_Duomo-Via_De'_Cerretani_54/r	048017ViaDe'Cerretani	Via De' Cerretani	VIA DEI CERRETANI	RT04801702317TO

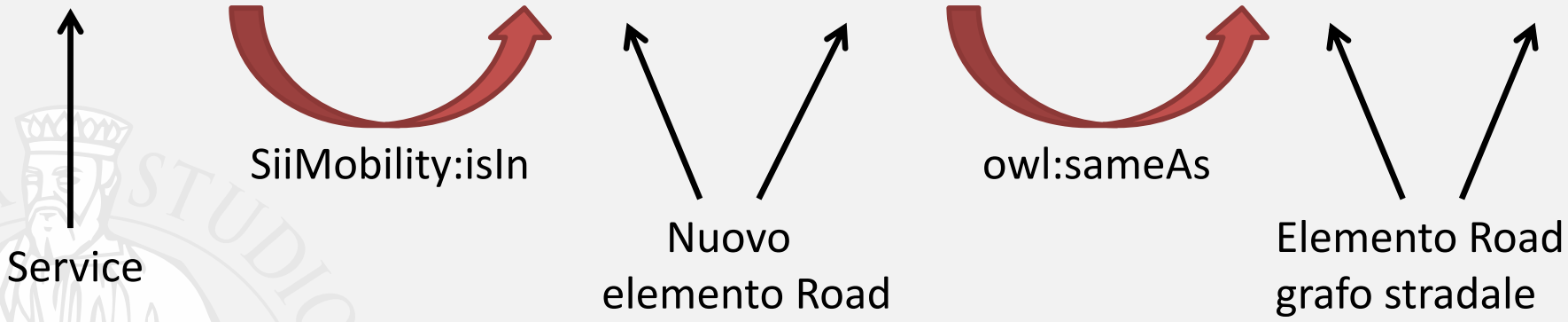
Service

Nuovo elemento Road

Elemento Road grafo stradale

Riconciliazione servizi

Service_Key	Road_Key	R_address	extendName	Cod_ toponimo
BELLAVISTA-Largo_F.lli_Alinari_15	048017LargoF.lliAlinari	Largo F.lli Alinari	LARGO FRATELLI ALINARI	RT04801701866TO
CASA_DEL_LAGO-Lungarno__A._Vespucci_58	048017LungarnoA.Vespucci	Lungarno A. Vespucci	LUNGARNO AMERIGO VESPUCCI	RT04801701874TO
COSMOPOLITAN-Via_F._Baracca187	048017ViaF.Baracca	Via F. Baracca	VIA FRANCESCO BARACCA	RT04801702987TO
SAN_PAOLO_IMI-VIA_DE'_VECCHIETTI_22/R	048017VIADE'VECCHIETTI	VIA DE' VECCHIETTI	VIA DEI VECCHIETTI	RT04801702383TO
CREDITO_ARTIGIANO-VIA_DE'_BONI_1	048017VIADE'BONI	VIA DE' BONI	VIA DEI BONI	RT04801702326TO
Auditorium_al_Duomo-Via_De'_Cerretani_54/r	048017ViaDe'Cerretani	Via De' Cerretani	VIA DEI CERRETANI	RT04801702317TO



Validazione e Verifica

- Query in linguaggio SPARQL
- Query di verifica del collegamento tra gli oggetti creati e quelli del grafo stradale
  - Si seleziona un elemento del grafo stradale e si richiedono tutti gli elementi ad esso collegati
  - Nel caso di grandi quantità di dati risulta necessario contare il numero di elementi collegati

Oggetto	Totale	Riconciliabili (7 comuni di Firenze)	Riconciliati
Previsioni meteo	286	7	7
Statistiche del comune	115	115	115
Uffici Pubblici	752	176	176
Servizi	28560	3559	3502
Statistiche sulle vie di Firenze	7987	7987	7987

Validazione e Verifica

❑ Ricerca delle vie percorse dalla linea tram

- Si seleziona un punto della linea tram
- Si sceglie un raggio di ricerca
- Tra i punti contenuti nell'area di ricerca si seleziona il più vicino
- Dal punto si risale al nome della via

Junction	Distanza (m)	Latitudine	Longitudine	Via
90	38	43.777796427323068	11.238888207612815	VIA ELIO GABBUGGIANI
250	22.3	43.774607511166131	11.225652786995651	VIA DEL SANSOVINO
290	39.9	43.773227993956844	11.214513710276817	VIA LUDOVICO CIGOLI
275	9.8	43.774052940009867	11.220394913130223	VIALE FRANCESCO TALENTI
340	12.6	43.767169341542193	11.208852373835150	VIALE PIETRO NENNI

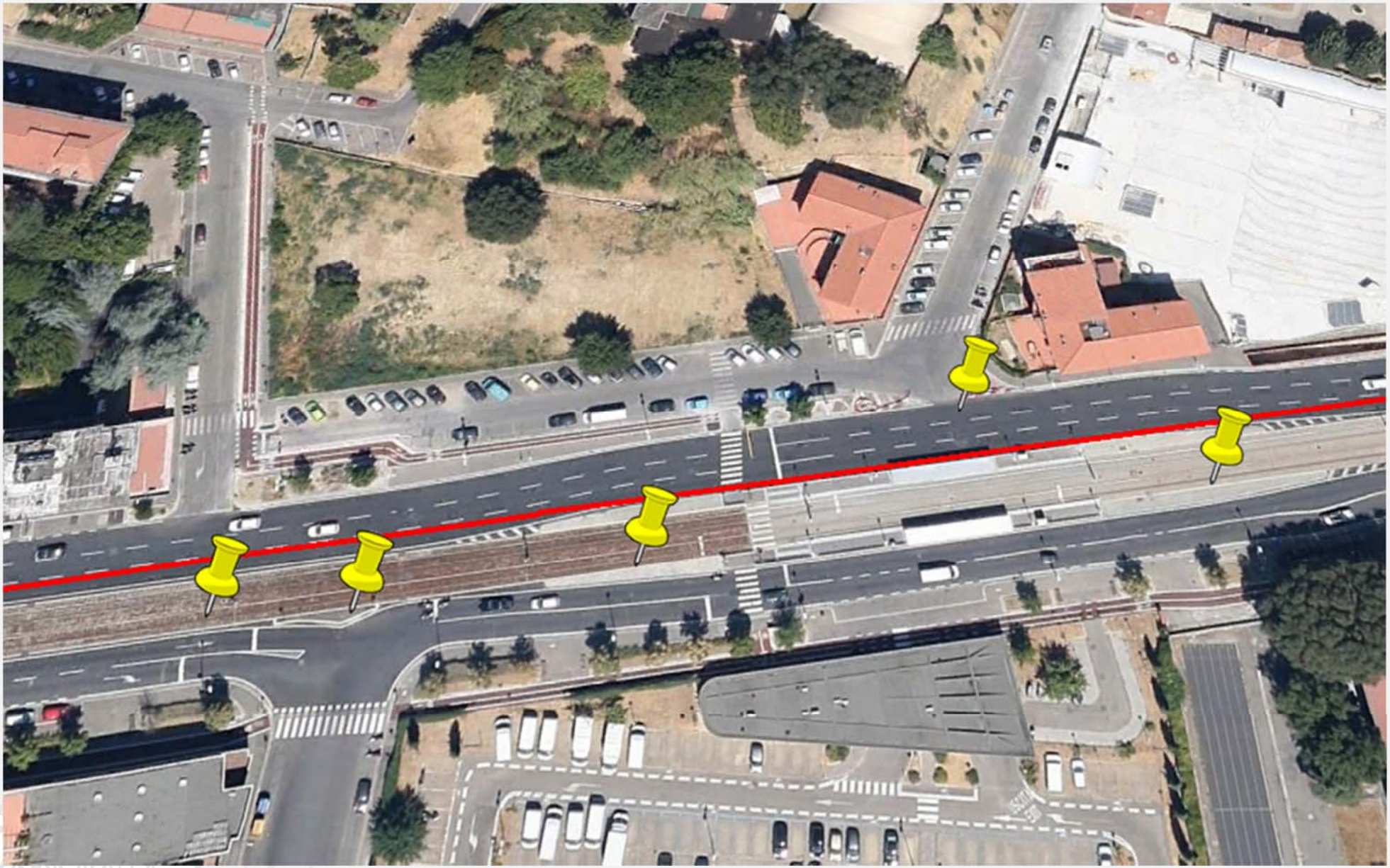
Validazione e Verifica

❑ Ricerca delle vie percorse dalla linea tram

- Si seleziona un punto della linea tram
- Si sceglie un raggio di ricerca
- Tra i punti contenuti nell'area di ricerca si seleziona il più vicino
- Dal punto si risale al nome della via

Junction	Distanza (m)	Latitudine	Longitudine	Via
90	38	43.777796427323068	11.238888207612815	VIA ELIO GABBUGGIANI
250	22.3	43.774607511166131	11.225652786995651	VIA DEL SANSOVINO
290	39.9	43.773227993956844	11.214513710276817	VIA LUDOVICO CIGOLI
275	9.8	43.774052940009867	11.220394913130223	VIALE FRANCESCO TALENTI
340	12.6	43.767169341542193	11.208852373835150	VIALE PIETRO NENNI

Validazione e Verifica



Validazione e Verifica

considerazioni

- I dati risultano essere ben collegati permettendo di effettuare query che coinvolgono diverse aree dell'ontologia
- Potrebbero essere utilizzati strumenti che aiutano ad analizzare e riconciliare i dati
- Già adesso il database RDF creato sembra collocarsi in un ambito Big Data

Dati	Elementi	Triple
Linea tram	570	5132
Statistiche	8100	59315
Servizi	28560	269121
Previsioni meteo	3,5 milioni l'anno	46 milioni l'anno

Verifiche e validazioni

- Le query vengono effettuate con il linguaggio SPARQL
- Sono state eseguite query per verificare i risultati della riconciliazione e query per mettere alla prova le capacità del database semantico

- **Verifica della riconciliazione**:
  - E' stato verificato che ogni parcheggio/sensore fosse correttamente collegato al grafo stradale
  - Il numero di istanze collegate deve coincidere col numero di parcheggi/sensori

Interrogazione del database semantico

- **Posizione delle fermate:**
  - Le fermate vengono geolocalizzate con coordinate GPS, con questa query viene trovata l'elemento stradale più vicino a una fermata, per poter poi risalire al nome della strada in cui è posizionata

Stop ID	Stopname	Roadname	Distance
FM0405	GRAMSCI 02	VIALE ANTONIO GRAMSCI	6.5m
FM0515	MARCONI 01	VIA GUGLIELMO MARCONI	5.7m
FN0465	D'ANNUNZIO 09	VIA GABRIELE D'ANNUNZIO	9.7m

Considerazioni

- Ci possono essere notevoli margini di miglioramento da parte del MIIC, principalmente sarebbero da migliorare i tempi di risposta del web service relativo ai dati dei sensori e terminare lo sviluppo del web service relativo ai dati dei dispositivi AVM
- Supponendo un funzionamento a pieno regime dei vari web services, si può dire che si lavora nell'ambito dei **big data**

Tipo di dato	Rilevazioni al giorno	Righe di HBase al giorno	Byte al giorno	Byte al mese
Sensori	18.000	18.000	9 MB	270 MB
Parcheggi	7.200	7.200	21 MB	630 MB
AVM	45.000	500.000	300 MB	9 GB
Totale	70.200	525.200	330 MB	10 GB

Data Analytics and Data Reasoning

- DISIT personnel is capable to integrate and define algorithms grounded on: statistical analysis, data mining, learning and knowledge reasoning.
- The typical goals are for: data reconciliation, data quality assessment and correction, data connection and inference, deduction, prediction, pattern detection, critical condition detection, discovering un expected correlations.