Semantic Traffic Diagnosis with STAR-CITY: Architecture and Lessons Learned from Deployment in Dublin, Bologna, Miami and Rio[*]

Freddy Lécué[1], Robert Tucker[1], Simone Tallevi-Diotallevi[1], Rahul Nair[1], Yiannis Gkoufas[1], Giuseppe Liguori[2], Mauro Borioni[2], Alexandre Rademaker[3], and Luciano Barbosa[3]

[1] IBM Dublin Research Centre, Ireland
[2] SRM - Reti e Mobilita, Bologna, Italy
[3] IBM Rio Research Centre, Brazil

Abstract. IBM **STAR-CITY** is a system supporting **S**emantic road **T**raffic **A**nalytics and **R**easoning for **CITY**. The system has ben designed (i) to provide insight on historical and real-time traffic conditions, and (ii) to support efficient urban planning by integrating (human and machine-based) sensor data using variety of formats, velocities and volumes. Initially deployed and experimented in Dublin City (Ireland), the system and its architecture have been strongly limited by its flexibility and scalability to other cities. This paper describes its limitations and presents the "*any-city*" architecture of STAR-CITY together with its semantic configuration for flexible and scalable deployment in any city. This paper also strongly focuses on lessons learnt from its deployment and experimentation in Dublin (Ireland), Bologna (Italy), Miami (USA) and Rio (Brazil).

1 Introduction

Entering 2014, the transportation system has matured in all major cities in the world; it only expands its infrastructure by a fraction of a percentage each year [1]. However, as projections indicate that more than half the world's population will be living in cities by 2030, congestion will continue to grow at an alarming rate, adversely impacting our quality of life and increasing the potential for accidents, long delays and other indirect consequences such as bus bunching. These are expected to escalate, calling for IT professionals to increase the functionalities, scalability, integration and productivity of existing transportation systems through the use of operational improvements.

There are several traffic analysis tools available, and some open, for use; however, they rarely encompass mechanisms for handling data heterogeneity, variety and integration. Therefore very few traffic systems are easily really portable from one city to another one. Most of the existing modern traffic systems[4] such as US TrafficView [2], TrafficInfo, French Sytadin or Italian 5T mainly focus on monitoring traffic status in cities using pre-determined and dedicated sensors (e.g., loop indiction detectors), all exposing numerical data. Others, more citizen-centric such as the traffic layer of Google Maps or [3], provide real-time traffic conditions and estimation but do not deliver insight to interpret historical and real-time traffic conditions. For instance the diagnosis

[*] The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement ID 318201 (SIMPLI-CITY).

[4] Traffic Systems: trafficview.org, trafficinfo.lacity.org, www.sytadin.fr, www.5t.torino.it/5t

of traffic condition [4] or the problem of explaining traffic condition is not addressed by state-of-the-art traffic systems. Basic in-depth but semantics-less state-of-the-art analytics are employed, limiting also large scale real-time data interpretation and integration. Thus, context-aware computing together with reusability of the underlying data and flexible deployment of traffic systems are limited. The reasoning functionalities are also very limited and reduced to basic analytics such as traffic monitoring or prediction.

STAR-CITY[5,6] (**S**emantic **T**raffic **A**nalytics and **R**easoning for **CITY**) [5], as a daily-used system which integrates heterogeneous data in terms of format variety (structured and unstructured data), velocity (static and dynamic data) and volume (large amount of historical data), has been mainly designed to provide such insights on historical and real-time traffic conditions. STAR-CITY completely relies on the W3C semantic Web stack e.g., OWL 2 (Web Ontology Language) and RDF (Resource Description Framework) for representing semantics of information and delivering inference outcomes. The strength of STAR-CITY lies in the ability of the system to perform various types of semantic inferences i.e., spatio-temporal analysis, diagnosis, exploration and prediction of traffic condition and congestion (cf. [5] for an high level presentation). These inferences are all elaborated through a combination of various types of reasoning i.e., (i) semantic based i.e., distributed ontology classification-based subsumption [6], (ii) rules-based i.e., pattern association [7], (iii) machine learning-based i.e., entities search [8] and (iv) sensor dynamic-based i.e., correlation [7].

Initially deployed and experimented in Dublin City (Ireland), the system, its architecture and its semantic-related components have shown limitations regarding their flexibility and scalability to other cities. This paper describes their scenarios and their limitations. We also present the "*any-city*" architecture of STAR-CITY together with its semantic configuration for flexible and scalable deployment in any city. The paper also strongly focuses on lessons learnt from the deployment and experimentation of the new architecture in Dublin (Ireland), Bologna (Italy), Miami (USA) and Rio (Brazil), which is completely novel with respect to past presented work [4] (STAR-CITY diagnosis in Dublin), [9] (STAR-CITY prediction in Dublin) and [5] (STAR-CITY in Dublin). To the best of our knowledge there is no single traffic system which (i) supports advanced traffic analysis functionalities as STAR-CITY does, and (ii) scales up to major cities.

The paper is organized as follows: Section 2 presents the contexts and scenarios associated to Bologna, Miami, Rio and their main differentiators with Dublin. Section 3 describes the new flexible system architecture and configuration for "*any city*". Section 4 reports some experimental results regarding scalability, flexibility and semantic expressivity. Section 5 reports on lessons learned from deploying STAR-CITY in major cities. Section 6 draws some conclusions and talks about possible future directions.

2 Diagnosing Anomalies in Dublin, Bologna, Miami and Rio

As highlighted in Section 1 the STAR-CITY system has been designed for analyzing, diagnosing, exploring and predicting traffic condition in cities. We focus on the diagnosis-based reasoning scenarios of Bologna, Miami and Rio as they are the most

[5] Video (.avi, .mov, m4v format) available: http://goo.gl/TuwNyL

[6] Live system: http://dublinked.ie/sandbox/star-city/

representative and exposed in terms of semantic Web technologies. In particular we differentiate the latter three innovative in-use scenarios with the one from Dublin, which has already been implemented, tested and experimented [4]. Table 1 synthesizes the main important details of the data sets we have considered for this reasoning task.

Source Type	Data Source	Description	City			
			Dublin (Ireland)	Bologna (Italy)	Miami (USA)	Rio (Brazil)
Traffic Anomaly	Journey travel times across the city	Traffic Department's TRIPS system[a]	CSV format (47 routes, 732 sensors) 0.1 GB per day[b]	✗ (not available)		
	Dublin Bus Dynamics	Vehicle activity (GPS location, line number, delay, stop flag)	✗ (not used)	SIRI: XML format[c] (596 buses, 80KB per update 11GB per day[d])	CSV format (893 buses, 225 KB per update 43 GB per day[e])	CSV format (1, 349 buses, 181 KB per update 14 GB per day[f])
Traffic Diagnosis	Social-Media Related Feeds	Reputable sources of road traffic conditions in Dublin City	"Tweet" format - Accessed through Twitter streaming API[g]			
			Approx. 150 tweets per day[h] (approx. 0.001 GB)	✗ (not available)	Approx. 500 tweets per day[i] (approx. 0.003 GB)	✗ (not available)
	Road Works and Maintenance		PDF format (approx. 0.003 GB per day[j])	XML format (approx. 0.001 GB per day[k])	HTML format (approx. 0.001 GB per day[l])	✗ (not available)
	Social events e.g., music event, political event	Planned events with small attendance	XML format - Accessed once a day through Eventbrite[m]APIs			
			Approx. 85 events per day (0.001 GB)	Approx. 35 events per day (0.001 GB)	Approx. 285 events per day (0.005 GB)	Approx. 232 events per day (0.01 GB)
		Planned events with large attendance	XML format - Accessed once a day through Eventful[m]APIs			
			Approx. 180 events per day (0.05 GB)	Approx. 110 events per day (0.04 GB)	Approx. 425 events per day (0.1 GB)	Approx. 310 events per day (0.08 GB)
	Bus Passenger Loading / Unloading (information related to number of passenger getting in / out)		✗ (not available)	✗ (not available)	CSV format (approx. 0.8 GB per day[e])	CSV format (approx. 0.1 GB per day[e])

[a] Travel-time Reporting Integrated Performance System - http://www.advantechdesign.com.au/trips
[b] http://dublinked.ie/datastore/datasets/dataset-215.php (live)
[c] Service Interface for Real Time Information - http://siri.org.uk
[d] http://82.187.83.50/GoogleServlet/ElaboratedDataPublication (live)
[e] Private Data - No Open data
[f] http://data.rio.rj.gov.br/dataset/gps-de-onibus/resource/cfeb367c-c1c3-4fa7-b742-65c2c99d8d90 (live)
[g] https://sitestream.twitter.com/1.1/site.json?follow=ID
[h] https://twitter.com/LiveDrive - https://twitter.com/aaroadwatch - https://twitter.com/GardaTraffic
[i] https://twitter.com/fl511_southeast
[j] http://www.dublincity.ie/RoadsandTraffic/ScheduledDisruptions/Documents/TrafficNews.pdf
[k] http://82.187.83.50/TMC_DATEX/
[l] http://www.fl511.com/events.aspx
[m] https://www.eventbrite.com/api - http://api.eventful.com

Table 1. (Raw) Data Sources for STAR-CITY in Dublin, Bologna, Miami and Rio.

We report major in-use challenges for each scenario where concrete solutions are presented (Section 3) and experimentation conducted (Section 4) for validation.

2.1 Diagnosing Traffic Congestion in Dublin City (*Reminder* of [4])

• ***Description***: The diagnosis task in Dublin consists in explaining why the road traffic is congested. Anomalies are captured by the Dublin journey travel time data set in Table 1 (cf. traffic anomaly row). There are a number of specific circumstances which cause or aggravate congestion. However capturing an accurate explanation of the reasons of congestion is a challenging problem. Traffic accidents, road works and social events (e.g., music, political events) are considered as potential sources of explanation in the Dublin context (cf. traffic diagnosis related rows).

• ***Motivation***: Traffic congestion has a number of negative effects, which strongly affects cities, their citizens and operators. For instance it reduces economic health because of the (i) non-productive activity of people stuck in their vehicles, (ii) wasted fuel, among others. Capturing the explanation of traffic congestion will support the city and transportation operators to act upon changing scenarios in real-time. For instance, given accurate explanations of congestion, the city traffic manager could be pro-active by (i) taking corrective actions on incoming traffic flow by changing the traffic strategies of close traffic lights, (ii) alerting the appropriate emergency services, (iii) re-routing traffic, (iv) better planning events in the city and more importantly (iv) informing its citizen in real-time

• ***Challenge***: Diagnosing traffic condition is a challenging research problem of interest for the semantic Web community because (i) relevant data sets (e.g., road works, social events, incidents), (ii) their correlation (e.g., road works and social events connected to the same city area) and (iii) historical traffic conditions (e.g., road works and congestion in Canal street on May 9th, 2014) are not fully integrated, linked and jointly exploited. Recent progress in the area [4] demonstrated the applicability of semantic Web technologies in solving this challenge.

2.2 Diagnosing Bus Congestion in Bologna

• ***Description***: The diagnosis task in Bologna consists in explaining why buses are congested. Contrary to the Dublin scenario, bus data is considered, providing more sparse data (because of the moving bus-related sensors) and a different data format i.e., SIRI XML instead of CSV. In addition the amount of data used for diagnosis (cf. traffic diagnosis row) is not as significative as in the Dublin scenario in terms of (i) size and (ii) number of data sets e.g., no report of traffic incident in Bologna. Finally the road works are exposed in Italian and digitalized in a different format.

• ***Motivation***: cf. Motivation of Section 2.1 with a focus on bus congestion in Bologna.

• ***Challenges and STAR-CITY Limitations***: Conceptually, diagnosing bus congestion relies in a similar reasoning task of the one described in Section 2.1. However from an in-use perspective diagnosing bus congestion in Bologna requires to address the following technical challenges:

C_1 Traffic anomalies are identified and represented differently.
How to capture a semantic, core representation of anomalies in any city?

C_2 The sources of diagnosis and their size are not similar e.g., social media is missing.
How accuracy of diagnosis is impacted by its sources? Are they still representative?

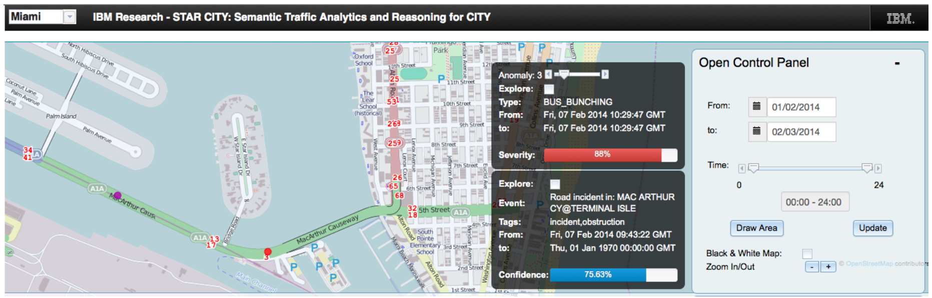
C_3 The sources of diagnosis are heterogeneous from one city to another one.
How to configure STAR-CITY in a way that is scalable and flexible to any city?

C_4 Data exposed in Bologna is real-time information but with a very low throughput.
Could the architecture of STAR-CITY be decoupled from its streaming components?

C_5 The schema of some data sources is in Italian cf. road works in Table 1.
How to make use of cross languages data sources?

2.3 Diagnosing Bus Bunching in Miami

• ***Description***: The diagnosis task in Miami consists in explaining why buses bunched. Fig.1 illustrates STAR-CITY in Miami. In public transport, bus bunching refers to a group of two or more buses, which were scheduled to be evenly spaced running along the same route, instead running in the same location at the same time. This occurs when at least one of the vehicles is unable to keep to its schedule and therefore ends up in the same location as one or more other vehicles of the same route at the same time. Contrary to Dublin but similarly to Bologna scenario, bus data is considered but with a much higher throughput, which may raise some questions regarding scalability of STAR-CITY in Miami. Contrary to Dublin and Bologna scenarios, much more data sources (i.e., passengers-related data) with larger size are considered. Again the format of data slightly changed across cities.



Miami
IBM Research - STAR CITY: Semantic Traffic Analytics and Reasoning for CITY

Anomaly: 3
Explore: □
Type: BUS_BUNCHING
From: Fri, 07 Feb 2014 10:29:47 GMT
to: Fri, 07 Feb 2014 10:29:47 GMT
Severity: 88%

Explore: □
Event: Road incident in: MAC ARTHUR CY@TERMINAL ISLE
Tags: incident,obstruction
From: Fri, 07 Feb 2014 09:43:22 GMT
to: Thu, 01 Jan 1970 00:00:00 GMT
Confidence: 75.63%

Open Control Panel
From: 01/02/2014
to: 02/03/2014
Time: 0 24
00:00 - 24:00
Draw Area
Update
Black & White Map: □
Zoom In/Out - +

Fig. 1. STAR-CITY In Miami. (color).

• ***Motivation***: The end result can be unreliable service and longer effective wait times for some passengers on routes that had nominally shorter scheduled intervals. Another unfortunate result can be overcrowded vehicles followed closely by near-empty ones.

• ***Challenges and STAR-CITY Limitations***: In addition to challenges C_1, C_2 and C_3 in Section 2.2 which are also valid in this context, diagnosing bus bunching in Miami requires to address the following technical challenge:

C_6 The number of diagnosis sources is larger e.g., bus passenger loading set is added. *How accuracy of diagnosis is impacted by new external sources? (dual to (b)).*

2.4 Diagnosing Low On-Time Performance of Buses in Rio

• ***Description***: The diagnosis task in Rio consists in explaining the low on-time performance of buses i.e., buses which are heavily delayed. The reasons can range from traffic incidents, accidents, bus bunching, detour, or unrealistic scheduling. Contrary to the Dublin and Miami scenarios the amount of data sets of potential use for diagnosis is very low i.e., only events and information about passengers loading are available. In addition the schema of the latter data set is different and in Portuguese.

• ***Motivation***: Such problems can result in unreliable bus services for Rio, which could turn in complex problems such as bus bunching, and even more critical problems such as emphasized in motivation of Section 2.1.

• ***Challenges and STAR-CITY Limitations***: In addition to challenges C_1, C_2, C_3, C_5 and C_6 in Sections 2.2 and 2.3 which are also valid in this context, diagnosing bus delays in Rio requires to address the following technical challenges:

C_7 The historic of information is 480 days while it is more than 3 years for other cities. *How accurate is the diagnosis in a context of limited historical information?*

This section described the problems which have not been foreseen by the initial architecture of STAR-CITY, but which have strong impacts and limitations for flexible and scalable deployment in Bologna, Miami and Rio. All challenges C_1, C_3-C_5 are problems where semantic web technologies have been strongly considered in the innovative and deployed architecture of STAR-CITY in Bologna, Rio and Miami (cf. Section 3) while challenges C_2, C_6 and C_7 are related to data characteristics (availability, relevance, accuracy) and their fit-for-purpose (cf. Section 4).

3 Flexible System Architecture and Semantic Configuration

The high-level architecture of STAR-CITY (both Dublin and Bologna, Miami, Rio versions) in Fig.2 consists of four main components: (i) semantic application configuration, (ii) semantic application server, (iii) data layer and (iv) user interface. In this section we explain how we adapted the initial version of STAR-CITY (running for Dublin) and its underlying technologies (i) to address the aforementioned challenges C_1, C_3-C_5 of Section 2, and then (i) to be flexible for deployment in other major cities in the world.

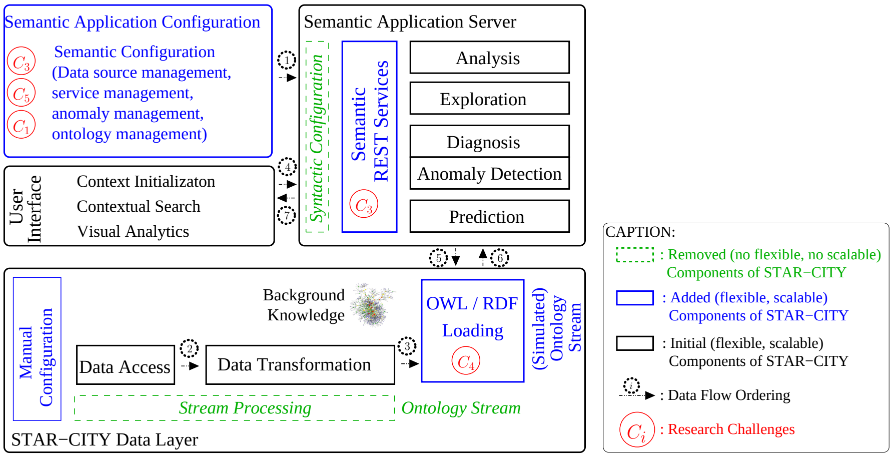


Fig. 2. High-Level System Architecture of STAR-CITY with References to (a) **Challenges** C_i of Section 2, and (b) **Initial**, **Removed**, **Added** Components of Dublin STAR-CITY to Support Flexible Deployment in Bologna, Miami and Rio. (color).

3.1 Semantic Application Configuration

The *semantic application management and configuration* component is the main component of STAR-CITY which enables flexible and scalable deployment of the system to other cities. Initially deployed and experimented in Dublin city, STAR-CITY did not address the challenges C_1, C_3 and C_5.

• ***Challenge*** C_1 ***"Anomaly Identification"***: The identification of anomalies in the initial version of STAR-CITY is pre-determined by some very simple fixed encoded rules, for instance (1) encoding the rule "*if travel time between sensorID*$_{203}$ *and sensorID*$_2$ *is less than* 183 *seconds then trigger diagnosis service*".

$$\begin{aligned} TriggerDiagnosis(s_1, s_2, time) \leftarrow\ & Sensor(s_1) \wedge Sensor(s_2) \wedge travelTime(s_1, s_2, time, value) \\ & \wedge\ equalTo(s_1, 203) \wedge equalTo(s_2, 2) \\ & \wedge\ lessThan(value, 183)\ \textit{\% with ``183'' is the min. threshold} \end{aligned} \quad (1)$$

Obviously such an approach is not scalable to other domains and even other cities. Indeed, one would need to redefine the rules at every new deployment phase from scratch. We address this problem by following [10], which provides semantics for capturing anomalies at semantic level. The approach consists in supervising the end-user in annotating values ranges of sensors with predefined concepts "*Anomaly*" and "*Normal*" from our domain ontology $\mathcal{O}_{\mathcal{D}}$. This ontology, represented as a RDF-S taxonomy, is simply used for defining the domain e.g., Dublin Travel Time and its anomalies. Further domains can be easily added e.g., Bologna, Miami and Rio Bus domain. The appropriate rules are then encoded semantically using SWRL rules[7], which connect the logical rules to the domain of application. Following this semantic extension of STAR-CITY, end-user can easily extend $\mathcal{O}_{\mathcal{D}}$, and then encode any anomaly identification rules.

```
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix dbpr: <http://dbpedia.org/resource/> .
@prefix addr: <http://schemas.talis.com/2005/address/schema#> .
@prefix rdfcal: <http://www.w3.org/2002/12/cal/icaltzd#> .
@prefix ibmVoc: <http://www.ibm.com/smartercities/cityfabric/voc#> .
@prefix busVoc: <http://www.ibm.com/SCTC/ontology/BusOntology.owl#> .
@prefix xmls: <http://www.w3.org/2001/XMLSchema#> .

<!-- Spatial Representation -->
<http://starcity.traffic.bus.miami.anomaly/venues/AltonRoad_10>
    a geo:SpatialThing ; dbpr:Country_Code <dbpr#ISO_3166-1:US> ;
    addr:countryName "USA" ; addr:localityName "Miami" ; addr:streetAddress "Alton Road" ;
    rdfcal:summary "Bus Bunching Anomaly" ;
    geo:lat "25.788371^^<xmls#float>" ; geo:long "-80.141280^^<xmls#float>" .

<!-- Temporal Representation and Type -->
<http://starcity.traffic.bus.miami.anomaly/event/Anomaly_1398032000_Bus113>
    a rdfcal:Vevent ; a ibmVoc:Anomaly ; a ibmVoc:Traffic ;
    ibmVoc:eventTag "flow" , "speed", "bunching", "delay", "road", "traffic" ;
    ibmVoc:hasEventCategory rdfcal:Vevent , ibmVoc:Anomaly , ibmVoc:BusBunching ;
    rdfcal:tzname "GMT" ; rdfcal:created "2014-04-20T20:01:20^^<xmls#dateTime>" ;
    rdfcal:dtstart "2014-04-20T20:01:20^^<xmls#dateTime>" ;
    rdfcal:dtend "2014-04-20T20:13:20^^<xmls#dateTime>" ;
    rdfcal:summary "Bus Bunching Anomaly" ;
    ibmVoc:hasSensingBUS <busVoc#Bus113> ; ibmVoc:hasSeverity "STOPPEDFLOW" ;
    geo:location <http://starcity.traffic.bus.miami.anomaly/venues/AltonRoad_10> .
```

Fig. 3. Example of a Bus Bunching Anomaly Representation in Miami (rdf/s prefixes omitted).

• ***Challenge*** C_1 ***"Anomaly Representation" (Fig.3)***: The initial representation of anomalies did not require any semantics has only one type of anomaly was diagnosed in the Dublin scenario. In larger cities, traffic anomalies could be of different types, which need to be captured. The new representation of anomalies in STAR-CITY follows a strict and simplistic (on purpose) representation of anomalies i.e., spatial, temporal representations, types and associated key tags e.g., Miami bus 113 bunching in Fig.3. Such a semantic representation is specifically important in the context of bus-related diagnosis since different types of bus anomalies may occur in one city e.g., delay, congestion, bunching. Capturing and representing their types is very important to (i) understand how anomalies and their types are correlated to their diagnoses, (ii) easily search among anomalies which are captured by different systems e.g., bus congestion by a TRIPS system, bus delays by a bus operator related system (cf. Dublin case where bus delay and travel time could be provided by two different systems).

• ***Challenge*** C_3 ***"Semantic Inter-City Configuration"***: The semantic inter-city configuration challenge C_3 is complimentary addressed by the semantic application config-

[7] http://www.w3.org/Submission/SWRL/

uration (i.e., configuration-layer in this section), server (services-side in Section 3.2) and data layer (data-side Section 3.3) in STAR-CITY. In the initial version of STAR-CITY for Dublin, each and every dimension of our data sets has been represented in OWL / RDF, reaching to a very detailed contextual information but also to (i) a very tight model which is not flexible to other cities, and (ii) a time-consuming mapping process (i.e., mapping from raw to RDF data cf. Section 3.3). The migration of STAR-CITY from one city to another one requires major customization and many steps of configuration. For instance the traffic impact of a road work event is defined in Dublin and Miami, not Bologna; its area of work (e.g., secondary, pavement) is defined in Miami and Bologna, not in Dublin. Similarly traffic accidents (through social media) are captured in the Dublin and Miami scenarios, but not for Bologna and Rio. Since the diagnosis is highly coupled to the level and categories of representation of events, it is very important that the inputs of diagnosis (i.e., traffic diagnosis row of Table 1) are pre-configurable. To this end we let the (admin) users define the relevant raw data and associated concepts to be considered for diagnosis. For instance, the diagnosis application of Bologna and Miami could be defined as in Fig.4. In such a configuration, the diagnosis reasoning can be configured with respect to its inputs (e.g., input of diagnosis for diagnosing bus congestion in Bologna), their types (e.g., RoadWork defined in the ibmVoc ontology) and raw data sources (e.g., BolognaRoadWorkComplete) and respective properties (e.g., latitude, longitude, area of work).

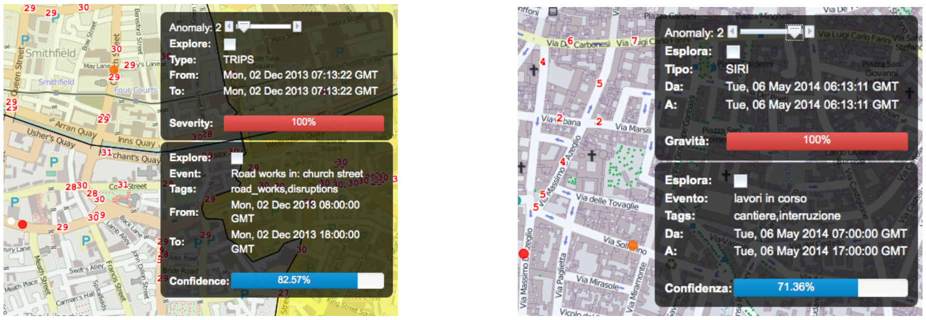
```
@prefix ibmVoc: <http://www.ibm.com/smartercities/cityfabric/voc#> .

<!-- Configuration Settings for Bus Congestion Diagnosis in Bologna -->
<http://starcity.traffic.bus.bologna/reasoning/diagnosis>
   <!-- Configuration of inputs to be considered for diagnosis reasoning -->
  <Class ibmVoc:Input> <Type ibmVoc:RoadWork> <Source "BolognaRoadWorkComplete">
       <Property geo:lat> <Property geo:long> <!-- Spatial Constraints -->
       <Property rdfcal:dtstart> <Property rdfcal:dtend> <!-- Temporal Constraints -->
       <Property ibmVoc:description> <Property ibmVoc:areaOfWork> <!-- RoadWork Features -->
  <Class ibmVoc:Input> <Type ibmVoc:MajorEvent> <Source "Eventful"> <!-- Property omitted -->
  <Class ibmVoc:Input> <Type ibmVoc:MinorEvent> <Source "Eventbrite"> <!-- Property omitted -->

<!-- Configuration Settings for Bus Bunching Diagnosis in Miami -->
<http://starcity.traffic.bus.miami/reasoning/diagnosis>
  <Class ibmVoc:Input> <Type ibmVoc:RoadWork> <Source "BolognaRoadWorkComplete">
       <Property geo:lat> <Property geo:long> <!-- Spatial Constraints -->
       <Property ibmVoc:description> <Property ibmVoc:areaOfWork> <!-- RoadWork Features -->
       <Property rdfcal:dtstart> <Property rdfcal:dtend> <!-- Temporal Constraints -->
       <Property ibmVoc:impact>
  <Class ibmVoc:Input> <Type ibmVoc:MajorEvent> <Source "Eventful"> <!-- Property omitted -->
  <Class ibmVoc:Input> <Type ibmVoc:MinorEvent> <Source "Eventbrite"> <!-- Property omitted -->
  <Class ibmVoc:Input> <Type ibmVoc:Incident> <Source "Twitter"> <!-- Property omitted -->
  <Class ibmVoc:Input> <Type ibmVoc:BusLoading> <Source "BusTransit"> <!-- Property omitted -->
```

Fig. 4. Semantic Configuration for Bologna and Miami Diagnosis Reasoning.

The new configuration settings of STAR-CITY, defined through the IBM Rational family of software configuration management solutions and extended with semantics, is flexible, easy to be exported to any city. Instead of directly interacting with the REST APIs (cf. Fig.2), the semantic configuration is used to automatically adapt the APIs with the appropriate settings. The city-wide customization is then driven by (the semantics of) the vocabulary used for defining the inputs, their types, sources, properties.

• ***Challenge*** C_5 ***"Multi-Lingual System"***: STAR-CITY has been designed for running with english vocabularies such as IBM ibmVoc. Such vocabularies, which strongly drive the reasoning engine, do not offer multi-lingual features and very few connections to open vocabularies. This strongly limits the entry of non-english speakers to the STAR-

CITY configuration, which turns to be the case for Bologna (Italian language) and Rio (Portuguese language) administrators. Therefore the interpretation and customization of inputs (e.g., configuration in Fig.4), results (e.g., diagnosis schema-related such as their types cf. Fig.5) are rather difficult, and even impossible in some cases. We address this problem by simply manually adding extra links (i.e., Linked Open Data resources [11]) to all concepts of our IBM vocabulary.



Anomaly: 2
Explore: □
Type: TRIPS
From: Mon, 02 Dec 2013 07:13:22 GMT
To: Mon, 02 Dec 2013 07:13:22 GMT
Severity: 100%

Explore: □
Event: Road works in: church street
Tags: road_works,disruptions
From: Mon, 02 Dec 2013 08:00:00 GMT
To: Mon, 02 Dec 2013 18:00:00 GMT
Confidence: 82.57%

Anomaly: 2
Esplora: □
Tipo: SIRI
Da: Tue, 06 May 2014 06:13:11 GMT
A: Tue, 06 May 2014 06:13:11 GMT
Gravità: 100%

Esplora: □
Evento: lavori in corso
Tags: cantiere,interruzione
Da: Tue, 06 May 2014 07:00:00 GMT
A: Tue, 06 May 2014 17:00:00 GMT
Confidenza: 71.36%

(a) Diagnosis in Dublin (English)

(b) Diagnosis in Bologna (Italian)

Fig. 5. Semantics-driven Multi-Lingual STAR-CITY (color). Diagnosis results are automatically provided in preferred language by using language-related links of DBpedia. (Hyperlinks are provided for describing specific terms e.g., cantiere - construction - not displayed for sake of readability).

By adding LOD links to our vocabulary we also give the possibility for non expert users to get extra and detailed information related to non self-explanatory events such as construction, obstruction, drainage (by simply follow new hyperlinks in STAR-CITY). Fig.6 illustrates a simple extension of our vocabulary, where associated Italian[8] and Portuguese[9] transcriptions of Traffic_collision are used in the appropriate context.

```
<http://www.ibm.com/smartercities/cityfabric/voc#TrafficIncident>
   a http://www.ibm.com/smartercities/cityfabric/voc#Event
   owl:sameAs <http://dbpedia.org/resource/Traffic_collision"> <!-- [A] -->
   <!-- owl:sameAs <http://it.dbpedia.org/resource/Incidente_stradale> through ref. to [A] -->
   <!-- owl:sameAs http://pt.dbpedia.org/resource/Acidente_rodovirio> through ref. to [A] -->
```

Fig. 6. Sample of a Simple LOD Extension of IBM STAR-CITY Vocabulary.

3.2 Semantic Application Server

• ***Challenge*** C_3 ***"Semantic Inter-City Configuration"***: As reported in Fig.2 and Fig.7, web-facing services use a set of SA-REST services[10]. These services are implemented on a custom application running on IBM WebSphere Application Server. REST-related technologies have been considered in STAR-CITY because of its lightweight protocol. Extended with a semantic layer, SA-REST was the most appropriate solution to accommodate our semantic configuration (cf. Challenge C_3 in Section 3.1). In details the semantic configuration is combined with the skeleton of each STAR-CITY APIs to

[8] Italian http://it.dbpedia.org/resource/Incidente_stradale for English Traffic Collision.

[9] Portuguese http://pt.dbpedia.org/resource/Acidente_rodovirio for English Traffic Collision.

[10] http://www.w3.org/Submission/SA-REST/

provide customized SA-REST analysis, exploration, diagnosis and prediction services e.g., diagnosing bus bunching in Miami with road works and accidents, or both. The description of the low-level implementation of the services is described in the APIs registry (e.g., how technically diagnosis reasoning is interacting with input data sources) while the semantic and high-level representation of the expected services are described in the semantic configuration (e.g., Figure 4 for configuring STAR-CITY Bologna and Miami). Such an architecture, which discharges the manual and syntactic configuration of services, ensures flexible deployment of customized STAR-CITY functionalities in the context of any city.

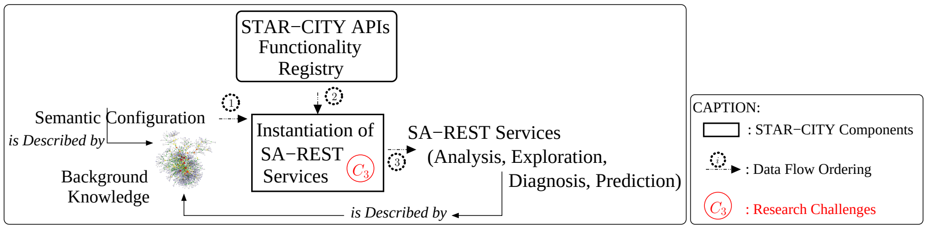


Fig. 7. Semantic Instantiation and Implementation of STAR-CITY SA-REST Services. (color).

3.3 Data Layer

Contrary to the semantic application server, a step of manual configuration is required in the data layer. It is mainly in charge of defining the data access points (e.g., URL of TRIPS data[b] in Table 1), protocols (e.g., HTTP for TRIPS), frequency (e.g., every minute for TRIPS) and various basic raw data parsing (e.g., adding timestamp to data collected for TRIPS). The Perl programming language, its standard modules together with CRON jobs are used for this purpose. We also manually define the mapping procedure from raw data source to semantic representation. The mapping procedure, completely described in Section 3.2 of [9], consists of a set of mapping files which describes how raw data is transformed in semantic representations associated to our domain ontology. Basic XSLT (for XML) and custom tabular transformation procedures (for CSV) are applied.

• ***Challenge*** C_4 ***"Semantic Stream Agnostic Architecture"***: Initially designed in a streaming infrastructure, the data access and transformation of STAR-CITY is now stream-agnostic. Ontology streams are not generated anymore from the data layer. The main reasons of this architecture shift are: (i) low throughput of STAR-CITY-related sensors in our city test cases, (ii) cost of streaming platform deployment, (iii) cost of configuration, and (iv) weak flexibility (regarding the on-the-fly integration of new data). Instead data transformation and aggregation is performed independently in a traditional manner (i.e., using pre-defined java routines and Perl scripts). The output of the transformation is a semantic and temporal representation. Therefore, conceptually, the output is similar to the initial version i.e., OWL statements are stored in jena TDB, where some temporal indexes have been added.

4 Experimental Results

This section focuses on challenges C_2, C_6 and C_7 by comparing and analyzing the scalability and accuracy of the reasoning component of STAR-CITY in Dublin, Bologna, Miami and Rio. In particular we aim at (i) analyzing how our approach reacts to the size (C_2), number (C_6) and historic (C_7) of data sources (cf. Sections 4.2 and 4.3) within our city context (cf. Section 4.1), and (ii) studying the impact of semantic expressivity by adjusting the underlying ontologies (cf. Section 4.4). Requested by traffic controllers, scalability and accuracy of the system have been extensively tested to validate the relevance, usefulness and (agreed) deployment of STAR-CITY. The experiments have been conducted on a server of 6 Intel(R) Xeon(R) X5650, 3.46GHz cores, and 6GB RAM.

4.1 Context

Live data from Dublin, Bologna, Miami and Rio (Table 1) are ingested and transformed in OWL/RDF (Table 2) following the principles of the STAR-CITY data layer (Fig.2 and Section 3.2 of [9]) and using different static background knowledge (Table 3), are used for experimentation. The highest expressivity of the ontologies is OWL EL.

Real Time, Live Data	City	Frequency of Update (s)	Raw Update Size (KB)	Semantic Update		Semantic Conversion Computation Time (s)
				Size (KB)	#RDF Triples	
[a] Journey Times	Dublin	60	20.2	6, 102	63, 000	0.61
[b] Bus	Bologna	120	31.8	1, 166	4, 550	0.295
	Miami	40	66.8	1, 766	11, 000	0.415
	Rio	60	96.8	2, 366	16, 145	0.595
[c] Incident	Dublin	600	0.2	1.0	7	0.002
	Miami	180	0.2	1.0	9	0.002
[d] Road Works	Dublin	once a week	146.6	77.9	820	3.988
	Bologna	once a day	78.9	133.2	1, 100	0.988
	Miami	3600	102.6	103.6	912	1.388
[e] City Events	Dublin	once a day	240.7	297	612	1.018
	Bologna		111.2	149	450	0.434
	Miami		637.2	789	1, 190	1.876
	Rio		585.3	650	950	1.633
[f] Bus Loading	Miami	40	833	2, 500	4, 500	0.390
	Rio	60	69.7	650	1, 230	0.147

Table 2. Details of Real-time Live Data in No Particular Order (average figures).

The objective is to diagnose traffic anomalies in the different test cities i.e., traffic congestion in Dublin, bus congestion in Bologna, bus bunching in Miami, low on-time performance of buses in Rio. The evaluation is achieved on a different data sets combinations since our test cities have access to different data sets. From the most to the least complete case we have: [b,c,d,e,f] for Miami, [a,c,d,e] for Dublin, and [b,e,f] for Rio,[b,d,e] for Bologna (cf. Table 2 for data set {a,b,c,d,e,f} reference). Specifically we evaluate the impact of the data sets combination on scalability and accuracy.

Ontology	Size (KB)	#Concepts	#Object Properties	#Data Properties	#Individuals	Imported Ontologies	Data Sets Covered
IBM Travel Time	4, 194	41	49	22	1, 429	Time Geo	[a]
IBM SIRI-BUS [4]	41.9	21	17	18	-		[b]
LODE[a](initial)	12	14	16	-	-		[e]
(extended)	56	87	68	31	-	Time, Geo	[c-f]
W3C Time[b]	25.2	12	24	17	14	-	[a-f]
W3C Geo[c]	7.8	2	4	-	-	-	[a-f]
DBpedia	Only a subset is used for annotation i.e., 28 concepts, 9 data properties						[c-e]

[a] http://linkedevents.org/ontology/2010-10-07/rdfxml/
[b] http://www.w3.org/TR/owl-time/
[c] http://www.w3.org/2003/01/geo/

Table 3. Static Background Knowledge for Semantic Encoding.

4.2 Scalability Experimentation and Results

Fig.8 reports the scalability of our diagnosis reasoning and core components (i.e., data transformation, OWL / RDF loading in Jena TDB, anomaly detection) of STAR-CITY by comparing their computation time in different cities and contexts. Similarly to data transformation and OWL / RDF loading, the anomaly detection and diagnosis reasoning have been performed over one day of traffic.

• ***Challenges*** C_2***,*** C_6***"Impact of Data Sources (and their combination)"***: The number and size of data sets have strong negative impact on the overall STAR-CITY. Indeed the more data sets the more overhead on transformation, loading, and reasoning. For instance STAR-CITY performs better in Bologna (data sets [b,d,e]) than in Miami (data sets [b,c,d,e,f]), although the latter results remain scalable.

• ***Challenges*** C_7 ***"Impact of Historic Data"***: As expected the computation performance (of one day) of raw data transformation is not impacted by the size of historical information (cf. secondary vertical x axis) while the computation of the OWL / RDF loading slightly increases accordingly. The latter is caused by the overhead of RDF triples loading on the TDB store, which requires some non negligible time times for re-indexing e.g. 100 minutes of indexing over one complete day of RDF storage in Rio. More interestingly the more historical information the more computation time, specifically for diagnosis reasoning e.g., a factor of 5.3 from an historic of 10 days to 480 days in Miami. This is caused by the intensive event similarity search over historical events performed by the diagnosis [4].

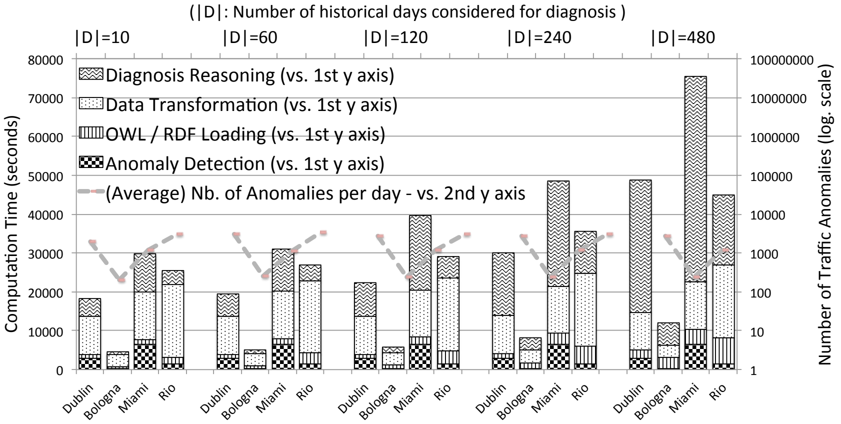


Fig. 8. Scalability of STAR-CITY in Dublin, Bologna, Miami and Rio.

4.3 Accuracy Experimentation and Results

Fig.9 reports the impact of historical information (challenges C_2, C_6) and size and number of data sets (challenge C_7) on accuracy of diagnosis results in Dublin, Bologna, Miami and Rio. The accuracy has been evaluated by comparing our explanation results against those estimated by transportation experts (used as ground truth) in their respective cities. A basis of one complete day of experimentation has been used i.e., $2,800$, 240, 1190 and $3,100$ traffic anomalies for respectively Dublin, Bologna, Miami and Rio. Fig.9 reports the average accuracy of diagnosis results.

• ***Challenges** C_2, C_6**"Impact of Data Sources (and their combination)"***: The more data sources the more accurate the diagnosis results. For instance the accuracy of diagnosis is the highest in the context of Miami (with the largest number of datasets i.e., [b,c,d,e,f]) while the accuracy is the lowest for Bologna (with the smallest number of datasets i.e., [c,d,e]) for all historical configurations. Interestingly, we learned that the *bus passenger loading* dataset has a stronger positive impact on diagnosis accuracy than the *traffic incident* dataset in all historical configurations $|D| = 10$, 60, 120, 240 and 480 cf. Bologna context vs. Miami context.

• ***Challenges** C_7 **"Impact of Historic Data"***: Reducing the number of historical events decreases accuracy of diagnosis. The more similar historical events the higher the probability to catch accurate diagnosis. For instance the accuracy of diagnosis results is improved by a factor of 1.5 by multiplying the number of historical days by a factor 8 (from 60 to 480 days).

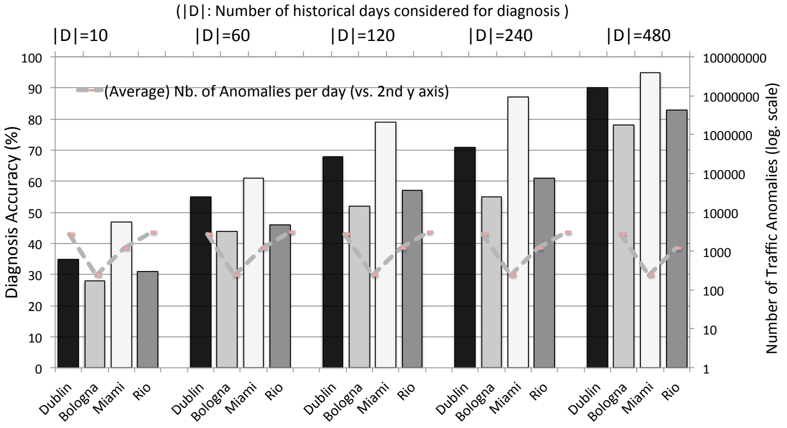


Fig. 9. Accuracy of STAR-CITY Diagnosis in Dublin, Bologna, Miami and Rio.

4.4 Expressivity Experimentation and Results

We slightly adjust the context (Section 4.1) by modifying the expressivity of the underlying ontologies (Table 3). Initially in OWL EL, we removed existential constructs of the representation to capture knowledge in RDF/S. We also extend the latter knowledge to capture the OWL RL dialect. Finally we consider OWL $\mathcal{SROIQ}(\mathcal{D})$ by adding extra artificial constraints to the initial model. The number of historical days $|D|$ considered for diagnosis is fixed to 480.

• **Expressivity vs. Scalability**: Fig.10 reports the scalability of STAR-CITY using different levels of representation. Unsurprisingly the RDF/S configuration is the most scalable while the $\mathcal{SROIQ}(\mathcal{D})$ is the most time consuming in all contexts. The diagnosis reasoning is the most impacted components i.e., (on average) $+750\%$ from RDF/S to $\mathcal{SROIQ}(\mathcal{D})$. The computation time of anomaly detection ($+410\%$) is also altered while the OWL / RDF loading ($+1.5\%$) and data transformation ($+1.1\%$) are less impacted. The diagnosis reasoning is based on consistency checking and semantic matching functionalities[11], which are constrained by the expressivity of the model. Similarly the more expressive the model the more time consuming is the anomaly detection, following results from [10]. These claims are demonstrated by results of Fig.10.

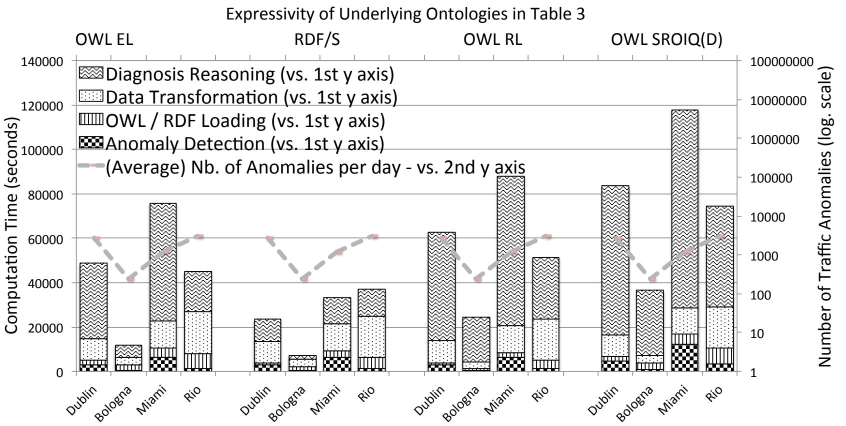


Fig. 10. Expressivity vs. Scalability of STAR-CITY in Dublin, Bologna, Miami and Rio.

• **Expressivity vs. Accuracy**: Fig.11 reports the accuracy of STAR-CITY using different levels of representation. Interestingly the RDF/S version of STAR-CITY is over performed by the OWL EL (+186%), OWL RL (+174%) and OWL $\mathcal{SROIQ}(\mathcal{D})$ (+190%) versions. By reducing the expressivity of the model (i.e., RDF/S) we tend to light and loose the semantic representation of events in Table 3, which in turn largely reduces the accuracy of the semantic matching functions (crucial during the diagnosis phase). In other words downgrading the model to RDF/S largely impacts the accuracy of diagnosis since all discriminating elements of the events cannot be considered by the matching procedure, which ends up with a large portion of similar (and more critically non-discriminable) events. Upgrading the models to OWL EL, RL or $\mathcal{SROIQ}(\mathcal{D})$ adds extra semantic features to events which can be used for semantic matching and comparison, hence a better semantic events discrimination and diagnosis accuracy.

The OWL EL, RL and $\mathcal{SROIQ}(\mathcal{D})$ configurations reach roughly similar accuracy results, although the OWL $\mathcal{SROIQ}(\mathcal{D})$ version is slightly better than the OWL EL (+0.97%) and RL (+0.91%) versions. The differences are not significative since the OWL RL and $\mathcal{SROIQ}(\mathcal{D})$ versions do not differentiate events descriptions much further than OWL EL. They actually simply support a refinement of the matching scores.

5 Lessons Learned

Deploying STAR-CITY and its semantics-aware architecture in more than one city raised new challenges C_1-C_7 which we addressed in the new version of the deployed system. The universal anomaly identification, representation (C_1) and configuration (C_3) were the most critical challenges from a flexible, scalable deployment inter-city. We extensively use semantic technologies for addressing these issues i.e., (i) semantic model for C_1, (ii) semantic configuration and SA-REST services for C_3. Even so

[11] Diagnosis reasoning is achieved by semantically comparing events and their characteristics over time.

some manual tasks are required to be achieved e.g., identification of anomalies ranges, definition of OWL / RDF mapping process (for data transformation). The OWL / RDF (concept) linking (alignment) process has also been performed manually to address C_5, but only once. However the latter needs to be replicated for each new data source and mapping presented to STAR-CITY. The automation of this process is a complex task as it required to align descriptions from very expressive vocabularies with concepts from unexpressive models such as DBpedia.

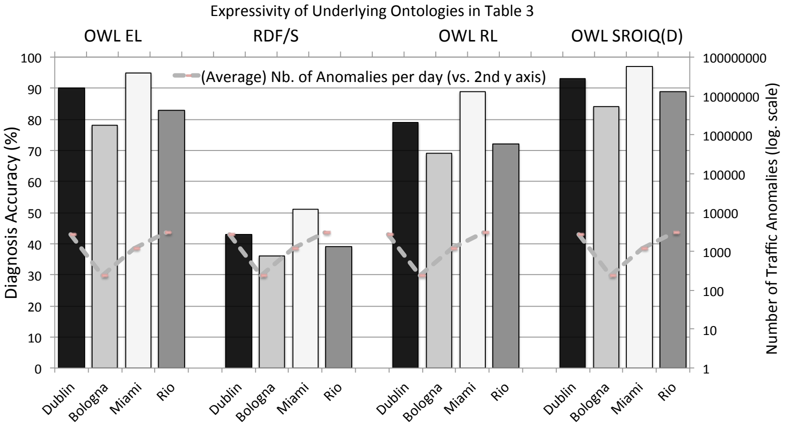


Fig. 11. Expressivity vs. Accuracy of STAR-CITY in Dublin, Bologna, Miami and Rio.

The semantic stream conversion was not beneficial to the overall architecture as it adds overhead on costs, deployment, configuration, systems interactions. Since the throughput of sensors in the four cities was considerably low we shifted the semantic transformation to a more traditional architecture. Shifting architectures did not impact the performance of the system (experimentation not reported in this paper). Even if higher throughput sensors could be an issue, we did not face it in our city contexts.

As experimented in Fig.10, expressive representation models means scalability issues. Even if the accuracy of the reasoning results is correlated to the expressivity of the semantic model, we noted differences in scale and impact cf. OWL EL vs. RDF/S configurations in Fig.11, cf. OWL EL vs. $\mathcal{SROIQ}(\mathcal{D})$ configurations in Fig.11. Therefore defining the appropriate level of representation is not a trivial task, and need to be driven by the application needs while ensuring scalable and accurate processing.

Data from sensors evolve over time. We considered a subset of the W3C Time ontology to represent the starting date/time and a simple temporal extension of TDB. However more complex time feature could have been used for compacting semantic information e.g., temporal intervals. We did not address this problem but a complex temporal-aware representation mode would support more complex reasoning e.g., over time intervals. STAR-CITY uses basic methods to evaluate loose temporal similarity. However research challenges, already tackled by [12], would need to be considered for more accurate temporal joints.

From a pure STAR-CITY perspective, reducing the number of historical events (together with the number and size of sources for diagnosis) increases scalability, but also decreases accuracy. Therefore the more source the better for STAR-CITY. However the scalability of the ingestion, transformation and loading of semantic representation is strongly altered by these dimensions (cf. indexing issues raised by challenge C_7 in Section 4.2). The latter raises requirements towards scalable (big) semantic data structure.

Applying STAR-CITY to other cities raise also challenges regarding the well-known problem of data interpretation in general. Before adding any semantics, we were facing the problem of making sense of schema-less data, specifically when data was described in Italian and Portuguese. For instance most content of *bus passengers loading* data set is not really necessary and does not need semantic transformation.

6 Conclusion

IBM **STAR-CITY** is a system supporting **S**emantic (road) **T**raffic **A**nalytics and **R**easoning for **CITY**. Initially deployed and experimented in Dublin City (Ireland), the system, its architecture and its semantic-related components have shown limitations regarding their flexibility and scalability to other cities. This paper, focusing on the diagnosis reasoning component of STAR-CITY, described (i) its semantics-related limitations in the context of Bologna (Italy), Miami (USA), Rio (Brazil), and (ii) the innovative "*any-city*" architecture of STAR-CITY together with its semantic configuration for flexible and scalable deployment in any city. The paper also reported experimentations of STAR-CITY in Bologna, Miami and Rio, which have validated the architecture, design and specifications of new deployed system.

As emphasized in Section 5 the challenges related to automated semantic data linking and loading are immediate in-use problems to be addressed, while the issues related to temporal compact representation are longer-term challenges.

References

1. Alexiadis, V., Jeannotte, K., Chandra, A.: Traffic analysis toolbox volume i: Traffic analysis tools primer. Technical report (2004)
2. Nadeem, T., Dashtinezhad, S., Liao, C., Iftode, L.: Trafficview: traffic data dissemination using car-to-car communication. ACM SIGMOBILE Mobile Computing and Communications Review **8**(3) (2004) 6–19
3. Valle, E.D., Celino, I., Dell'Aglio, D., Grothmann, R., Steinke, F., Tresp, V.: Semantic traffic-aware routing using the larkc platform. IEEE Internet Computing **15**(6) (2011) 15–23
4. Lécué, F., Schumann, A., Sbodio, M.L.: Applying semantic web technologies for diagnosing road traffic congestions. In: International Semantic Web Conference (2). (2012) 114–130
5. Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M.L., Tommasi, P.: Star-city: semantic traffic analytics and reasoning for city. In: IUI. (2014) 179–188
6. Mutharaju, R.: Very large scale owl reasoning through distributed computation. In: International Semantic Web Conference (2). (2012) 407–414
7. Lécué, F., Pan, J.Z.: Predicting knowledge in an ontology stream. In: IJCAI. (2013)
8. Bicer, V., Tran, T., Abecker, A., Nedkov, R.: Koios: Utilizing semantic search for easy-access and visualization of structured environmental data. In: International Semantic Web Conference (2). (2011) 1–16
9. Lecue, F., Tucker, R., Bicer, V., Tommasi, P., Tallevi-Diotallevi, S., Sbodio, M.L.: Predicting severity of road traffic congestion using semantic web technologies. In: ESWC. (2014) 611–627
10. Lécué, F.: Towards scalable exploration of diagnoses in an ontology stream. In: AAAI. (2014)
11. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
12. Lutz, C.: Interval-based temporal reasoning with general tboxes. In: IJCAI. (2001) 89–96