Semantic Cities
AAAI Technical Report WS-12-13

Capturing the Pulse of Cities: Opportunity and Research Challenges for Robust Stream Data Reasoning

Freddy Lécué and **Spyros Kotoulas** and **Pól Mac Aonghusa**
IBM Research, Smarter Cities Technology Centre
Damastown Industrial Estate, Dublin, Ireland
{(firstname.lastname)@ie.ibm.com}

Abstract

In a Smarter City, available resources are harnessed safely, sustainably and efficiently to achieve positive, measurable economic and societal outcomes. Data and information from people, systems and things is the single most scalable resource available to city stakeholders but difficult to publish, organize, discover and consume, especially in a real-time context. Enabling city information as a utility, through a robust (expressive, dynamic, scalable) and (critically) a sustainable technology and socially synergistic ecosystem, could drive significant benefits and opportunities. In the context of stream data (as real-time, gigantic, noisy and private data), this paper targets research issues we identify as important to harness the fused information resources of cities, Citizens and Stakeholders to reach the concept of Smarter Cities.

Introduction

As defined by (Valle et al. 2009), stream reasoning is a new multidisciplinary approach, merging synergies from Artificial Intelligence (Machine Learning, Semantic Web), Database, Data Mining and Distributed Systems communities. Integrating the above disciplines can provide the abstractions, foundations, methods, and tools required to reason on data streams in a scalable way, thus providing a way to answer questions on real time events. What is happening in the city? What are the causes of their events (i.e, Why)? Could we predict their impact on the city e.g., its traffic? Are these events correlated? These are some of the questions we (as citizens and authority representatives) may ask and want explanations. Could we capture the (spatial and temporal) pulse of cities? Could we make the city run better, faster, cheaper? These are more general questions for which we envision stream data processing and reasoning as a potential approach to provide insights. The purpose of this position paper is to identify key research challenges to embrace the full potential of stream data processing, specially in the context of Smarter Cities.

The remainder of this paper is organized as follows: Section 2 presents a scenario motivating the relevance of stream data processing and reasoning. Section 3 presents research challenges we identified as important to address robust stream data reasoning in the Smarter City context. Finally Section 4 draws some conclusions.

A Motivating Scenario: Dublin City Traffic

Even if traffic jams can be easily detected, visualized and analyzed through stream data and optimization mechanisms using existing data mining (Luo et al. 2005), stream processing (Biem et al. 2010), machine learning approaches (Babu and Widom 2001; Liu, Pu, and Tang 2000), explaining their causes, predicting their impact and recommending alternative solutions are more complex and challenging problems, mainly due to the lack of information interpretation.

What could be the cause of a highway traffic jam? Is it broken traffic light, an accident, a brief stall, a temporarily overcrowded highway entrance or exit? The latter are potential causes of unexpected events which could happen in a city. Unfortunately, it is not always straightforward to obtain clear and descriptive explanations on reasons for unlikely events, especially in real-time situations. Understanding potential causes is important for informing interested parties, for instance, car drivers and public authorities, in real time. This is important not only for providing explanations to drivers who are sitting in bumper-to-bumper traffic, but also for ensuring that public authorities will take decisions and appropriate actions (e.g., rerouting or changing traffic light strategy in case of an accident or a broken traffic light) in time, especially in case of emergency.

How do weather forecasts could impact traffic conditions? Shall we expect delays or re-routing? Such questions remain open because of the mis-(or non) integration of data, information and knowledge from different domains. However their answers are important requirements for cities to make immediate and future decision on infrastructures, for instance. Predicting the impact of unexpected traffic conditions on (connected) roads, citizens, pollution, cities in more general term is also a challenging problem that most of cities are facing nowadays. Explaining causes of unexpected events and predicting their effects, which is in a sense improving the urban dynamics, are also challenging parts of the objectives of the Traffic scenario. Stream data from sensors or any other real-time feeds is obviously the basis we need to start with in order to extract information about real-time events and model knowledge of the city domain. Given this extent of knowledge domain, recommending alternative

Data Source	Description	Format Type	Temporal Frequency (s)	Historic (mm/yyyy)	Size Estimation per day (GBytes)	Data Provider
Dublin Bus	Vehicle activity (GPS location, line number, delay, stop flag)	SIRI: XML-based[a]	20	11/2010	4-6	(Private)
Dublin Traffic Flow Measurement along 24 Traffic Intersection IDs	Traffic Light Strategy	XML	30	01/2011	0.055	DCC
	Strategic Intersection Sensing		900		0.022	
CCTV Monitoring	Real-Time monitoring of Dublin City	Stream Video	Real Time	No (privacy reasons)	10^4	
Wunderground for Dublin	Real-time weather information	CSV	[5, 600] (depending on stations)	01/1996	[0.050, 1.5] (depending on stations)	(Public) Wunder-ground[b]
Road Weather Condition (54 stations)		CSV	600	11/2010	0.1	(Public) NRA[c]
Road Works and Maintenance		CSV	3600	11/2010	0.01	(Public) Dublinked[d]
Events in Dublin	Events with small attendance	XML	Not considered	11/2011	0.001	(Public) Eventbrite[e]
	Events with large attendance			11/2011	0.05	(Public) Eventful[f]
DBPedia	Structured facts extracted from wikipedia	RDF	No	No	3.5×10^6 concepts	(Public) DBPedia[g]
Dublin City Roads (listing of type, junctions, GPS coordinate)		RDF	No	No	0.1	(Public) Linked-geodata[h]

[a] SIRI (Service Interface for Real Time Information) is a standard for exchanging real-time information about public transport services and vehicles - http://siri.org.uk

[b] http://www.wunderground.com/weather/api/

[c] NRA - National Roads Authority http://www.nratraffic.ie/weather

[d] http://dublinked.ie/

[e] https://www.eventbrite.com/api

[f] http://api.eventful.com

[g] http://dbpedia.org/

[h] http://linkedgeodata.org

Table 1: (Incomplete) Overview of Traffic Scenario Data sets (Dublin City Dependant).

and complete solutions (e.g., by analyzing social media) is part of the requirements to reach the concept of sustainable cities that we envision in this scenario. The suggested solution would be as generic as possible not only to be applied to other cities, but also to be applicable to other domains such as water, energy or supply chain management, which also need explanation and prediction of impact of different events.

Capturing the pulse of the city in temporal and spatial perspective requires (1) capturing, filtering, analyzing, diagnosing massive amounts of data (i.e., raw data organized in unstructured format such as video streams for visualization of traffic hotspots for a given time of the day or more structured data such as location-based descriptions of moving entities – see Table 1[1]) by applying innovative data mining (Luo et al. 2005) and diagnosis approaches (Sampath et al. 1996; Lécué 2012), (2) explaining and predicting events by representing semantics of stream data, information and reasoning on the underlying continuous knowledge, and (3) recommending alternative solutions by analyzing and interpreting results of reasoning processes in way that social behaviors are considered.

Research Challenges

The chart in Fig.1 positions existing approaches towards data processing in relation to three dimensions: (1) knowledge expressivity i.e., how descriptive is the logics, (2) querying and reasoning i.e., how elaborated is the inference model, and (iii) data dynamicity i.e., how fast and how big data could be transferred.

The latter three challenges have been addressed in recent research EU projects such as KnowledgeWeb[2], LarKC[3], LOD2[4], PlanetData[5] among others, and also tackled by different approaches such as (Barbieri et al. 2010a; Ren and

[1] A large part of data is provided by DCC (Dublin City Council) through dublinked.ie agreement (http://dublinked.ie/), and hosted at IBM.

[2] http://knowledgeweb.semanticweb.org/

[3] http://www.larkc.eu/

[4] http://lod2.eu/Welcome.html

[5] http://www.planet-data.eu/

Pan 2011), but studied mainly as separated research challenges. Therefore, no method is able to support complex reasoning with expressive knowledge under large amount of data (as we consider in the motivating scenario section and more generally in our Smarter Cities dedicated projects) as envisioned by (Valle et al. 2009). Even more complex, scalability (through high performance computing-based distribution of knowledge management and reasoning processing), data uncertainty and privacy are rarely considered in the context of stream data. For instance, storing and querying rapidly changing information from high frequency sensors requires highly-performant and distributed systems. Processing data from city-aware data such as vehicles traffic has much more privacy issues than open social media feeds. In addition data collected from sensors tends to be incomplete, noisy, and unreliable. Robust stream data reasoning aims at addressing all these different aspects.

Reasoning on Stream Data and their Descriptions

Classical deductive reasoning tasks have been developed over the last years to decide on subsumption, classification, consistency and instance checking. In the same time, other so-called *non-standard* (Horrocks 2002) or *constructive* (Colucci et al. 2010) reasoning tasks have been proposed in the DLs (Description Logics) literature to address new issues related to knowledge-based domains, especially in retrieval scenarios, ontology design and maintenance and automated negotiation. Concept abduction (Noia, Sciascio, and Donini 2007), approximation (Stuckenschmidt 2007; Brandt, Kusters, and Turhan 2002), contraction (Colucci et al. 2004), covering (Benatallah et al. 2002), difference (Teege 1994), explanation (McGuinness and Borgida 1995), least common subsumer (Cohen, Borgida, and Hirsh 1992), most specific concepts (Baader 2003), similarity (Borgida, Walsh, and Hirsh 2005) can be cited among others. What are the reasoning approaches we could absorb, integrate and extend so we can achieve complex goals such as identifying, explaining and predicting events, and more specially unexpected situations? Are new methods of inference required?

Managing large amounts of data, maintaining an up-to-date view, and deriving knowledge on the fly are important parts of stream data reasoning. In such a context, how to integrate these processes so that materialization of knowledge and reasoning is optimized?

Knowledge Representation and Expressivity

A large number of different DLs-based reasoners have been proposed e.g., CEL (Baader, Lutz, and Suntisrivaraporn 2006), Fact++ (Horrocks 1998), HermiT (Motik, Grau, and Sattler 2008), Mamas (Noia, Sciascio, and Donini 2007), SHER (Dolby et al. 2009) (on top of Pellet (Sirin et al. 2007))), or Racer (Haarslev and Möller 2001) among others[6]. Some of them differ from each other from the type of reasoning they provide (see reasoning on stream data section) while others differ from the expressivity they support. What is most appropriate expressivity required to model data streams and their descriptions, so reasoning is accurate and scalable? How to integrate data streams with different underlying expressivity?

Even if some approaches present lightweight ways of describing data streams (Bolles, Grawunder, and Jacobi 2008), (Rodriguez et al. 2009; Barbieri et al. 2009), is it the most appropriate "semantic" model? What kind of representation model data streams required to be precessed so we can derive useful knowledge?

Data Dynamicity

Continuously collected data streams are generated by dynamic processes through sensors, actuators or even social media feeds. Since data changes over time, even drastically in some cases, catching the knowledge of a dynamic environment and infer new facts in real time are not straightforward tasks to achieve. Towards these issues, different models such as Aurora (Carney et al. 2002; Liu, Pu, and Tang 2000), OpenCQ (Liu, Pu, and Tang 2000), Stream (Babu and Widom 2001), Stream Mill (Luo et al. 2005), TelegraphCQ (Chandrasekaran et al. 2003) have been introduced. From a high level perspective, they extend basic database model to support the continuous aspect of stream data.

From an heterogenous integration perspective, Streaming SPARQL (Bolles, Grawunder, and Jacobi 2008), Time annotated SPARQL (Rodriguez et al. 2009) or C-SPARQL (Barbieri et al. 2009) are potential approaches extending SPARQL, which is a syntactically-SQL-like language for querying RDF graphs, to manage RDF-based data streams. As a complementary work, authors of (Ren and Pan 2011) present an ontology stream management system to deal with relatively large volumes of data and updates efficiently. Contrary to pure Terminological Box-based, reasoning is processed on an evolving knowledge, which is materialized at query time. However, stream data is important not only for its current values but also for past values produced. In order to support this, the history of the stream must be archived and stream reasoning systems must support history queries. Due to scalability and performance reasons, the latter approaches do not keep track of past derived facts (issued at querying and reasoning time), and they cannot support correlation (e.g., logic implication) of facts on a time basis, thus limiting explanation of facts and potential prediction of their impact on future knowledge. How to manage large number of views of knowledge? How to maintain flexible and scalable evolution of open knowledge? How to track, link, explain and predict their evolution?

Performance and Scalability

Deployment of stream data reasoning in large scale applications is a must to be successful. In particular integration of data streams from heterogeneous sources is one important aspect to consider, emphasizing issues related to distributed knowledge management (Bonifacio et al. 2002), querying and reasoning (Serafini and Tamilin 2005). In recent years, a

[6] A more complete list is maintained here: http://www.cs.man.ac.uk/ sattler/reasoners.html.

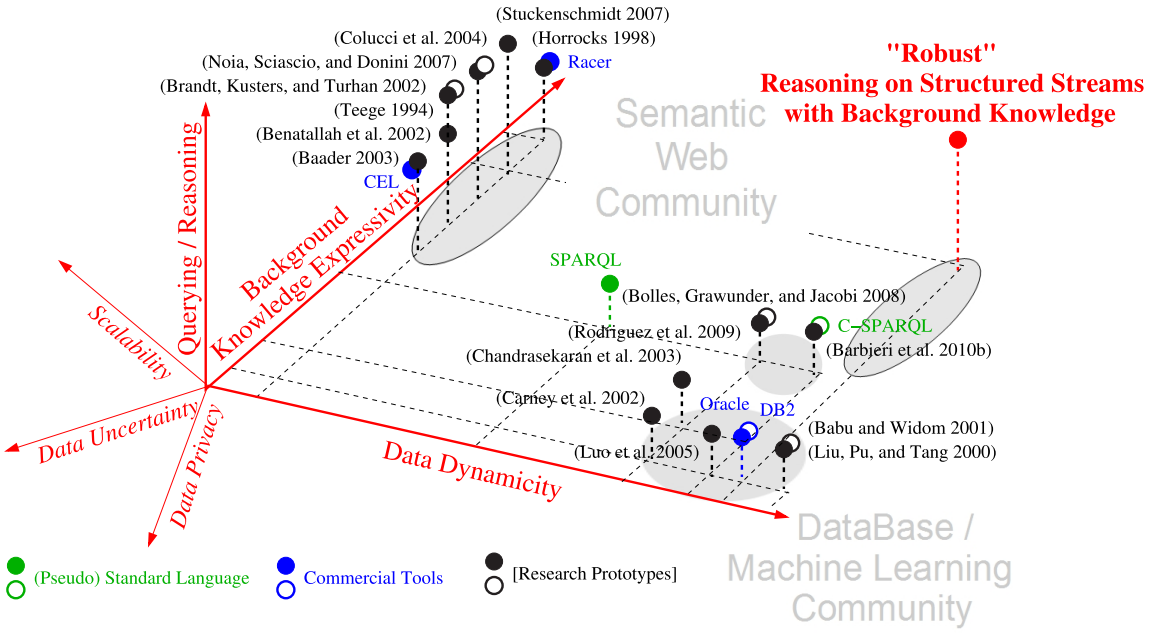

Figure 1: Classification of (some) Data Processing & Reasoning Approaches.

series of parallel reasoning approaches have pushed the scalability limit to larger and larger amounts of data (Kotoulas, Oren, and van Harmelen 2010), (Urbani et al. 2010), (Goodman and Mizell 2010), mainly by using clusters and highly scalable frameworks like Mapreduce (Dean and Ghemawat 2004). However, this has been done at the expense of data dynamicity (currently, the most scalable reasoner can not retract facts (Urbani et al. 2010), let alone deal with dynamic data), expressivity (most scalable approaches are limited to rule-based logics), and privacy (all data needs to be gathered in a cluster). How can be utilize powerful hardware while respecting privacy constraints? What is the optimal trade-off between scalability and expressivity? How should data be distributed, guaranteeing both optimal performance, freshness and privacy? What machinery is required to process city information?

New analytic solutions (not reported in Fig.1) such as Teradata[7], column stores (Stonebraker et al. 2005) and nosql / graph / array data engines such as Cassandra (Lakshman and Malik 2009), neo4j[8], sciDB[9] are important approaches that need to be deeply investigated to address the issues of performance and scalability.

Data Uncertainty

One problem which arises is that data streams published by distributed sources may have missing or incorrect data values, e.g. due to a network failure, wrong calibration. Therefore, data generated from sensors, actuators or feeds from social media could be noisy and incomplete (Barbieri et al. 2010b). These issues have an impact not only on derived information and knowledge, but also on the accuracy of querying and reasoning. How to deal with incomplete or wrong information where relevant data is missing? What about gaps between stored history of streams? How to represent the knowledge gap? Could we fix them using inference models?

A significant number of recent works have tackled data uncertainty, e.g., fuzzy DL (Straccia 1998), Trio (Widom 2005), mauvedb (Deshpande and Madden 2006) among other, and are relevant and potential approaches that would need to be investigated further.

Data Privacy

Public data is data that is supposed to be not subject to valid privacy, security or privilege limitations. However, it is common that public authorities misevaluate these limitations, leading to privacy issues related to the exposed data sets e.g., it could be straightforward to derive personal information in real-time by joining different sources of information. Privacy is then an important dimension to consider in an open data environment. As more and more data is released as open by governments or third parties, more sensitive information could be derived. How to control access to such information and the underlying data? How to anonymize streaming data for privacy protection? More specifically how to continuously facilitate anonymity on data streams?

Conclusion

In this position paper, we exposed some research challenges in the area of linked, open, and stream data (and more generally Semantic Web) that need to be addressed to reach the objective of Smarter Cities. In particular, we illustrated existing stream data reasoning approaches and their limitations

[7]http://www.teradata.com/
[8]http://neo4j.org/
[9]http://www.scidb.org/

along data description expressivity, reasoning, dynamicity, and partially along scalability, uncertainty and privacy. We do not consider robust stream data reasoning as a new research challenge, but rather as an important and integrated research challenge.

References

Baader, F.; Lutz, C.; and Suntisrivaraporn, B. 2006. Cel - a polynomial-time reasoner for life science ontologies. In *IJCAR*, 287–291.

Baader, F. 2003. Least common subsumers and most specific concepts in a description logic with existential restrictions and terminological cycles. In *IJCAI*, 319–324.

Babu, S., and Widom, J. 2001. Continuous queries over data streams. *SIGMOD Record* 30(3):109–120.

Barbieri, D. F.; Braga, D.; Ceri, S.; Valle, E. D.; and Grossniklaus, M. 2009. C-sparql: Sparql for continuous querying. In *WWW*, 1061–1062.

Barbieri, D. F.; Braga, D.; Ceri, S.; Valle, E. D.; and Grossniklaus, M. 2010a. Incremental reasoning on streams and rich background knowledge. In *ESWC (1)*, 1–15.

Barbieri, D. F.; Braga, D.; Ceri, S.; Valle, E. D.; Huang, Y.; Tresp, V.; Rettinger, A.; and Wermser, H. 2010b. Deductive and inductive stream reasoning for semantic social media analytics. *IEEE Intelligent Systems* 25(6):32–41.

Benatallah, B.; Hacid, M.; Leger, A.; Rey, C.; and Toumani, F. 2002. On automating web services discovery. *VLDB Journal* 1–26.

Biem, A.; Bouillet, E.; Feng, H.; Ranganathan, A.; Riabov, A.; Verscheure, O.; Koutsopoulos, H. N.; and Moran, C. 2010. Ibm infosphere streams for scalable, real-time, intelligent transportation services. In *SIGMOD*, 1093–1104.

Bolles, A.; Grawunder, M.; and Jacobi, J. 2008. Streaming sparql - extending sparql to process data streams. In *ESWC*, 448–462.

Bonifacio, M.; Bouquet, P.; Mameli, G.; and Nori, M. 2002. Kex: A peer-to-peer solution for distributed knowledge management. In *PAKM*, 490–500.

Borgida, A.; Walsh, T.; and Hirsh, H. 2005. Towards measuring similarity in description logics. In *Description Logics*.

Brandt, S.; Kusters, R.; and Turhan, A. 2002. Approximation and difference in description logics. In *KR*, 203–214.

Carney, D.; Çetintemel, U.; Cherniack, M.; Convey, C.; Lee, S.; Seidman, G.; Stonebraker, M.; Tatbul, N.; and Zdonik, S. B. 2002. Monitoring streams - a new class of data management applications. In *VLDB*, 215–226.

Chandrasekaran, S.; Cooper, O.; Deshpande, A.; Franklin, M. J.; Hellerstein, J. M.; Hong, W.; Krishnamurthy, S.; Madden, S.; Reiss, F.; and Shah, M. A. 2003. Telegraphcq: Continuous dataflow processing. In *SIGMOD Conference*, 668.

Cohen, W. W.; Borgida, A.; and Hirsh, H. 1992. Computing least common subsumers in description logics. In *AAAI*, 754–760.

Colucci, S.; Noia, T. D.; Sciascio, E. D.; Donini, F. M.; and Mongiello, M. 2004. A uniform tableaux-based method for concept abduction and contraction in description logics. In *ECAI*, 975–976.

Colucci, S.; Noia, T. D.; Sciascio, E. D.; Donini, F. M.; and Ragone, A. 2010. A unified framework for non-standard reasoning services in description logics. In *ECAI*, 479–484.

Dean, J., and Ghemawat, S. 2004. Mapreduce: Simplified data processing on large clusters. In *Operating Systems Design and Implementation*, 137–147.

Deshpande, A., and Madden, S. 2006. Mauvedb: supporting model-based user views in database systems. In *SIGMOD Conference*, 73–84.

Dolby, J.; Fokoue, A.; Kalyanpur, A.; Schonberg, E.; and Srinivas, K. 2009. Scalable highly expressive reasoner (sher). *J. Web Sem.* 7(4):357–361.

Goodman, E. L., and Mizell, D. 2010. Scalable in-memory rdfs closure on billions of triples.

Haarslev, V., and Möller, R. 2001. Description of the racer system and its applications. In *Description Logics*.

Horrocks, I. 1998. Using an expressive description logic: Fact or fiction? In *KR*, 636–649.

Horrocks, I. 2002. Reasoning with expressive description logics: Theory and practice. In *CADE*, 1–15.

Kotoulas, S.; Oren, E.; and van Harmelen, F. 2010. Mind the data skew: distributed inferencing by speeddating in elastic regions. In *WWW*, 531–540.

Lakshman, A., and Malik, P. 2009. Cassandra: structured storage system on a p2p network. In *PODC*, 5.

Lécué, F. 2012. Diagnosing changes in an ontology stream: A dl reasoning approach. In *AAAI*, (to appear).

Liu, L.; Pu, C.; and Tang, W. 2000. Correction to "continual queries for internet scale event-driven information delivery". *IEEE Trans. Knowl. Data Eng.* 12(5):861.

Luo, C.; Thakkar, H.; Wang, H.; and Zaniolo, C. 2005. A native extension of sql for mining data streams. In *SIGMOD Conference*, 873–875.

McGuinness, D. L., and Borgida, A. 1995. Explaining subsumption in description logics. In *IJCAI (1)*, 816–821.

Motik, B.; Grau, B. C.; and Sattler, U. 2008. Structured objects in owl: representation and reasoning. In *WWW*, 555–564.

Noia, T. D.; Sciascio, E. D.; and Donini, F. M. 2007. Semantic matchmaking as non-monotonic reasoning: A description logic approach. *J. Artif. Intell. Res. (JAIR)* 29:269–307.

Ren, Y., and Pan, J. Z. 2011. Optimising ontology stream reasoning with truth maintenance system. In *CIKM*, 831–836.

Rodriguez, A.; McGrath, R. E.; Liu, Y.; and Myers, J. D. 2009. Semantic management of streaming data. In *International Workshop on Semantic Sensor Networks at the International Semantic Web Conference*.

Sampath, M.; Sengupta, R.; Lafortune, S.; Sinnamohideen, K.; and Teneketzis, D. 1996. Failure diagnosis using dis-

crete event models. *IEEE Transactions on Control Systems Technology* 4(2):105–124.

Serafini, L., and Tamilin, A. 2005. Drago: Distributed reasoning architecture for the semantic web. In *ESWC*, 361–376.

Sirin, E.; Parsia, B.; Grau, B. C.; Kalyanpur, A.; and Katz, Y. 2007. Pellet: A practical owl-dl reasoner. *J. Web Sem.* 5(2):51–53.

Stonebraker, M.; Abadi, D. J.; Batkin, A.; Chen, X.; Cherniack, M.; Ferreira, M.; Lau, E.; Lin, A.; Madden, S.; O'Neil, E. J.; O'Neil, P. E.; Rasin, A.; Tran, N.; and Zdonik, S. B. 2005. C-store: A column-oriented dbms. In *VLDB*, 553–564.

Straccia, U. 1998. A fuzzy description logic. In *AAAI/IAAI*, 594–599.

Stuckenschmidt, H. 2007. Partial matchmaking using approximate subsumption. In *AAAI*, 1459–1464.

Teege, G. 1994. Making the difference: A subtraction operation for description logics. In *KR*, 540–550.

Urbani, J.; Kotoulas, S.; Maassen, J.; van Harmelen, F.; and Bal, H. 2010. Owl reasoning with webpie: calculating the closure of 100 billion triples. In *Proceedings of the Seventh European Semantic Web Conference*, LNCS. Springer.

Valle, E. D.; Ceri, S.; van Harmelen, F.; and Fensel, D. 2009. It's a streaming world! reasoning upon rapidly changing information. *IEEE Intelligent Systems* 24(6):83–89.

Widom, J. 2005. Trio: A system for integrated management of data, accuracy, and lineage. In *CIDR*, 262–276.