Eight Fallacies when Querying the Web of Data

Jürgen Umbrich [†], Claudio Gutierrez [‡], Aidan Hogan [†], Marcel Karnstedt [†], Josiane Xavier Parreira [†]

[†] *Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland*
[‡] *DCC, Universidad de Chile, Santiago, Chile*

Abstract— **The Web of Data refers to the universal database constituted by interlinked data sources on the Web. This global system is creating a new way of publishing and consuming data on the Web. A number of assumption that were valid in bounded, controlled, closed worlds of data are now being challenged.**

In this paper, following the seminal ideas presented in 1994 by Peter Deutsch and later completed by James Gosling, known as "The Eight Fallacies of Distributed Computing" [1], we present a set of fallacies for the area of the Web of Data.

I. Preamble

In this paper, we discuss eight fallacies that arise when processing queries over the Web of Data. We primarily focus on the Semantic Web and Linked Data movement for our fallacies, however, we believe that these arguments can be applied in general for a wider Web of Data. We aim to provide a checklist of essential challenges and common fallacies that we feel should be discussed by the communities concerned with querying the Web of Data.

II. The fallacies

1) **Data services/endpoints are reliable**
The first fallacy relates to the assumption that systems on the Web of Data are reliable and offer high availability. However, in reality, online data-hosting and query-execution platforms are provided as-is and without QoS guarantees and often suffer from downtimes, high loads, etc., affecting their reliability.[1] Systems that rely on external services must be robust in the face of downtimes for those services. In particular, the overall reliability of systems such as federated SPARQL engines that rely on multiple external endpoints is a product of the availability of those underlying engines: for example, a system relying on five independent SPARQL endpoints each with an availability of 80% uptime can expect all endpoints to be up only 32.7% of the time. While this is a well-known problem in P2P systems and is usually considered when querying Web sources live, many proposals for querying multiple SPARQL endpoints fall short in respecting this issue: for example, federated SPARQL engines have yet to consider the type of fault-tolerance or replication strategies that are now commonplace in P2P systems.

2) **Consumer behavior can be anticipated**
Many performance factors of a system can be optimised if the expected number and behavior of users is known or can be predicted. However, the openness of the Web allows anybody to connect and interact with data and services at any time. This makes it impossible to know the type of user, the volume of users, their behaviour and their requirements in advance. This is different from scenarios investigated in the area of data management and information retrieval where the profile, needs and volume of users can be planned for in advance. Given the diversity of consumers and publishers, their requirements and offerings, we cannot further rely on traditional statistics for modelling user behaviour without supporting a certain level of openness and flexibility, far beyond the levels supported today. To adapt a quote from Abraham Lincoln: *query engines for the Web of Data can satisfy some of the consumers all of the time and all of the consumers some of the time, but they cannot satisfy all of the consumers all of the time.*

3) **Publishers are infallible and play no role**
The open Web infrastructure allows anybody to access and publish data about anything at anytime in arbitrary locations without specifying the provenance of information. While this should be of no surprise to anyone reading this paper, the consequences of this fact are often ignored for the Web of Data. The Semantic Web standards have been noticeably quiet on the subject of the trustworthiness of data available on the Web: RDF, RDFS and OWL have struggled with notions of provenance, defaulting to the assumption that data are infallible. Although later standards such as SPARQL allow for assigning a coarse level of provenance (through named graphs), provenance means nothing without trust. Verification of results—which should be an essential ingredient for a Web query engine—thus becomes impossible: how can results then be trusted? As the level of automated processing increases: be it processing of joins across sources, the application of formal reasoning, etc., the problem of verification is compounded further and minor data quality issues snowball throughout the process. Towards solving this problem, it would seem that closer relationships between publishers, service-providers and users should be established, opening channels for a feedback loop through which issues of verification and data-quality can be tackled. The discipline of data management has focused primarily on consuming data (querying, retrieving, exchanging, etc.). In a sustainable infrastructure, publishers should also be first-class citizens.

4) **You can know what's out there**
Web data is dynamic. Publishers are autonomous and innumerable. Sources come and go. Although the effi-

[1] See `http://labs.mondeca.com/sparqlEndpointsStatus/`

ciency offered by materialised approaches makes them appealing, the assumption that such approaches can have a consistent view of the Web of Data is too simplistic, particularly in light of developments on, e.g., the Internet of Things. Furthermore, for the Web of Data, the connectivity of sources is directly dictated by the URIs embedded in the data (unlike, say, P2P systems). Thus, any query approaches relying on locally replicated knowledge to answer queries or to find data-sources from the Web cannot claim to reflect what's out there on the Web of Data at that time.

5) **UNIVERSAL COST MODELS CAN BE MAINTAINED**
Classical database cost models are based on (often) predictable factors such as the selectivity estimates, number of sources, available bandwidth and latency, local processing costs, etc. While such cost models could be naïvely applied for the Web of Data, they would need to be constantly updated to reflect the ever-changing nature of the open Web, where selectivity estimates are dynamic and are difficult to globally maintain; the number of sources, bandwidth and latency costs are constantly fluctuating and can vary widely across geolocations and hardware specifications (e.g., servers, mobiles, sensors); and where local processing costs depend on unknown load and input data. Such volatile factors suggest that expectations for the benefit of cost models should be lowered and balanced with the high cost of maintenance. Again, as per Abraham Lincoln's core message, cost models can be tailored for specific needs, or generalised in a lossy manner for global needs.

6) **QUERY EXECUTION IS ALWAYS DETERMINISTIC**
A core assumption in classical (distributed) database systems is that the results for a query are always deterministic: in other words, for a fixed query and a fixed set of data, the result should not vary. This doesn't always hold for the Web of Data. For example, ask the same query twice to a public (black-box) SPARQL endpoint on the Web today and you can sometimes get different results, even if the underlying data hasn't changed. This is due to a number of factors: public endpoints often implement hidden policies on execution times, fixed limit sizes, handling of multiple requests, etc., in such a manner that the consumer does not know if (and how) the results are complete. Coupled with a lack of default ordering in SPARQL results, this can lead to different subsets of results being returned for repeated runs. Like availability, non-determinism can snowball when considering multiple SPARQL endpoints. Also, publishers may make different data available through different media: content in SPARQL endpoints and dereferenceable documents may not correspond. This problem is not unique to the Web of Data, and is encountered for mediator and P2P systems. However, while such systems only have to deal with the availability of sources, non-determinism is a much deeper and still uncharted factor for the Web of Data.

7) **STANDARDS = INTEROPERABILITY**
Despite the efforts of various standardisation bodies—most prominently the W3C—the beast cannot be tamed: the Web is a wild creature that cannot be domesticated by the bridle of standards. In particular, the provision of standards does not imply that they will be uniformly followed or that they will be sufficient to universally enable interoperability across the Web. For example, many proposals for querying the Web of Data rely on the widespread adoption of Linked Data principles, particularly related to the dereferenceability and interlinkage of data. However, on the current Web of Data, such guidelines are only partially adhered to. Similarly, although SPARQL has been standardised for four years, its underlying semantics is defined for querying closed datasets and not the open Web. Even as standards expand to foster further interoperability, the cumulative costs of implementation grow prohibitively, ultimately affecting compliance across the Web. Thus, proposals for querying the Web of Data should not depend on the complete or even near-complete compliance with existing standards and guidelines.

8) **ONE SYSTEM CAN ACE THEM ALL**
In analogy to the CAP theorem [2], which is well-accepted in decentralised systems, we conclude with the ACE (Alignment, Coverage, Efficiency) theorem based on the following aspirations when querying the Web of Data:

Alignment: How well aligned results are with current data on the Web.

Coverage: How much coverage of the Web of Data the results exhibit.

Efficiency: How efficiently the query can be run.

Based on the fallacies and issues presented above, we argue that any approach for querying the Web of Data can at most guarantee two of the three ACE aspects.

III. CONCLUSION

We are currently witness to the dawn of a novel Web of Data for which classical query approaches are showing their strain. The eight fallacies we present should not be considered comprehensive, static or a critique of current directions. Instead, based on our own humble experiences as Semantic Web researchers and practitioners, we highlight these eight fallacies in order to raise awareness in a timely fashion of what we see as the fundamental challenges to come for querying the Web of Data. (Please forgive us for thinking aloud.)

Acknowledgments This research has been supported by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), Project Fondecyt 1110287 and European Commission under contract number FP720117287661 (GAMBAS).

REFERENCES

[1] A. Rotem-Gal-Oz, "Fallacies of distributed computing explained," Tech. Rep., 2012.

[2] S. Gilbert and N. Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services," *SIGACT News*, 2002.