From: **Petra Selmer** petra.selmer@neo4j.com
Subject: [Moderator Action] Petra Selmer: Position Statement for W3C Workshop on Web Standardization for Graph Data
Date: 11 January 2019 at 12:20
To: group-data-ws-pc@w3.org

Good day

Please find below my position statement.

Background

I obtained my PhD in Computer Science from Birkbeck, University of London, in 2016. My supervisors were Prof. Alexandra Poulovassilis and Prof. Peter T. Wood. My research investigated the flexible querying of graph-structured data: the approximation (through edit operations) and relaxation (induced by rules in an accompanying ontology) of conjunctive regular path queries, with a focus on proving the correctness of the underlying constructs and evaluation, as well showing that these were tractable. The motivation underpinning this research -- predicated on irregular, ever-evolving and heterogeneous data, with which a user may not be fully familiar -- was to provide additional opportunities to users to retrieve results that are of relevance to them by, for example, returning potentially the correct answers (in cases where the original query was incorrectly formulated) or additional related answers, enabling sophisticated data discovery and insights to be made.

I joined Neo4j in 2014 as part of the team designing and optimizing Neo4j's Cypher query engine. Since 2018, I have been a member of the Query Languages Standards and Research group at Neo4j, undertaking research into graph query languages and language standards, with the aim of evolving and standardizing property graph querying, and also to contribute materially to research involving Cypher, and to foster collaboration with academic institutions (such as the ongoing partnership between Neo4j and the team led by Prof. Leonid Libkin at the University of Edinburgh). I also support the openCypher project at www.opencypher.org,

Since 1997 and up to the point where I joined Neo4j (in 2014), I worked as a consultant and developer in a variety of different domains and roles, such as architecting and developing a rules engine to validate data in the clinical domain, and developing a system to manage heterogeneous data in a neurosciences research programme, among others.

Suggested Topic: A comprehensive survey of existing graph query languages
---

A few months ago, and preparatory to the commencement of planning for GQL, interested parties* -- drawn from industry (Neo4j, Oracle, Redis Labs and TigerGraph), the community (a noted data modelling expert and published technical author), and academia (the University of Talca in Chile) -- formed an informal working group called the "Existing Languages Working Group".

We have worked in an incremental fashion on systematically identifying, surveying, analysing and comparing graph query language features, drawn from the existing query languages: Cypher; PGQL; GSQL; SQL (in particular, the property graph extensions envisaged for SQL 2020); and G-CORE.

The work is envisioned to comprise a catalogue of: the groups of features; to which extent (if at all) these are supported in each language; exemplar syntax; supplementary artifacts to aid in the understanding of the underlying semantics; grammar constructs; and any additional details of interest.

The idea is to have this landscape of existing query languages to hand in order to inform the design and development of GQL by virtue of a well-informed work plan and helping to lead to a more robust outcome; i.e. this would help us to have clear and meaningful discussions on scope and priorities, and will facilitate clear and unambiguous design choices. Moreover, this will help us to identify areas of consolidation, innovation and opportunities for language interoperation in GQL (for example, with SPARQL).

We have used a phased approach, and began by examining the basic feature groups, such as basic querying and operations, and will move onto composable querying and complex patterns in future phases.

We have amassed a large amount of data so far, and would very much like to report back to others about our results (previously, our short updates on this work have garnered much interest) and, just as importantly, we would wholeheartedly welcome feedback and comments from the expert audience. For instance, what are the artifacts we should produce from this work? Are there other aspects of such an analysis the audience would find useful? Is there something we have perhaps not addressed as fully as possible? Are there surprises in the data -- and so on and so forth.

*The ELWG (Existing Languages Working Group) comprises:

- Renzo Angles (Universidad de Talca (Chile))
- Alin Deutsch (TigerGraph)
- Thomas Frisendal (Independent data modelling expert and author)
- Victor Lee (TigerGraph)
- Roi Lipman (Redis Labs)
- Petra Selmer (Neo4j)
- Oskar van Rest (Oracle)
- Mingxi Wu (TigerGraph)

Additional Topics

In addition to the topic above, which I would suggest being led by one or more ELWG members present, there are a number of other topics which I think would be useful and very relevant. I lay these out below.

1. Path queries and pattern matching
I think the whole area of path querying is crucial to graphs. It would be useful to discuss this, and somehow collate all the different approaches and valuable work that has been undertaken in this area over the past few decades. As an adjunct to this, path semantics such as homomorphic vs. isomorphic pattern matching would be very interesting too: are there cases where in reality, one form is far more useful to another...etc?

2. Graph schema
The issue of graph schema is very fundamental and, like path querying, another strand where new ground is being broken (compared to the relational database world). There has been much prior research in this area, and also there is ongoing work: it would therefore be useful to discuss this. It may also be an idea to look more closely at the work done in implemented systems (such as the Cypher for Apache Spark project).

3. GQL

I think a report and ensuing discussion on the status and envisaged work plan for GQL would be extremely useful.

Links

ELWG:

https://docs.google.com/presentation/d/1aL_6jyLjWR-qVgiVWCx_mrwgBNGrJu3x5MoLiPMMeDM/edit?usp=sharing

https://www.dropbox.com/sh/0hglt7tvqqkqxti/AABS0eCDirBgjjHn5vXF0m3Na?dl=0&preview=zoom_0.mp4

Cypher:

https://arxiv.org/pdf/1802.09984.pdf

https://s3.amazonaws.com/artifacts.opencypher.org/openCypher9.pdf

PGQL:

http://pgql-lang.org/spec/1.1/

https://event.cwi.nl/grades/2016/07-VanRest.pdf

GSQL:

https://docs.tigergraph.com/dev/gsql-ref

https://cdn2.hubspot.net/hubfs/4114546/IntegrationQuery%20PrimitivesGSQL.pdf

G-CORE: https://arxiv.org/pdf/1712.01550.pdf

SQL Property Graph Extensions: https://s3.amazonaws.com/artifacts.opencypher.org/website/ocig3/oCIG3+-+DM32.2+ad+hoc+on+SQL+extensions+for+property+graphs+July+update+20170727.pdf

GQL: https://www.gqlstandards.org/

Dr. Petra Selmer
Query Languages Standards and Research

Neo4j
8th Floor, Friars Bridge Court
41-45 Blackfriars Road
London SE1 8NZ