Extended Property Paths: Writing More SPARQL Queries in a Succinct Way

Valeria Fionda¹, Giuseppe Pirrò² and Mariano P. Consens³

¹ DEmaCS, University of Calabria, Italy
² WeST, University of Koblenz-Landau, Germany
³ MIE, University of Toronto, Canada

Abstract
We introduce Extended Property Paths (EPPs), a significant enhancement of SPARQL property paths. EPPs allow to capture in a succinct way a larger class of navigational queries than property paths. We present the syntax and formal semantics of EPPs and introduce two different evaluation strategies. The first is based on an algorithm implemented in a custom query processor. The second strategy leverages a translation algorithm of EPPs into SPARQL queries that can be executed on existing SPARQL processors. We compare the two evaluation strategies on real data to highlight their pros and cons.

1 Introduction
Motivated by the spread of graph-like data, the area of graph databases has received renewed attention. SPARQL, the W3C standard query language for data in RDF, has recently been enhanced with property paths (Harris and Seaborne 2013) to support graph navigation capabilities. The regular-expression-like syntax of property paths enables to write SPARQL navigational queries in a more succinct way and extend the matching of triple patterns to arbitrary length paths.

However, property paths offer limited navigation capabilities; queries like “Find my exclusive friends”, that is, my friends that are not friend of any of my friends, cannot be expressed. To enhance the expressive power of property paths and enable to write more navigational queries in a succinct way, we introduce Extended Property Paths (EPPs). In designing EPPs we identified a core of new features and investigated how to make available such features; indeed, one can devise a custom query processor like in (Alkhateeb, Baget, and Euzenat 2009; Fionda and Pirrò 2013) and/or leverage existing SPARQL processors. The goal of this paper is to discuss the design and implementation of EPPs.

Related Work. Graph query languages have been deeply studied (Wood 2012). It emerged that for certain classes of languages, like Conjunctive Regular Path Queries (CR-PQs) (Barceló et al. 2012; Alkhateeb, Baget, and Euzenat 2009), the evaluation problem can become too expensive, making the class unattractive for practical purposes. Hence, restricted classes of languages like acyclic CRPQs and Nested Regular Expressions (NREs) (Pérez, Arenas, and Gutierrez 2010) have been proposed. NREs are at the core of navigational languages for RDF such as nSPARQL (Pérez, Arenas, and Gutierrez 2010) and NautiLOD (Fionda, Gutierrez, and Pirrò 2012). Languages like TriAL (Libkin, Reutter, and Vrgoč 2013), TriQ (Arenas, Gottlob, and Pieris 2014) and NEMODEQ (Rudolph and Krötzsch 2013) are grounded on (extensions of) Datalog. Other expressive languages like SPARQLieR (Kochut and Janik 2007) and extended CRPQs (Barceló et al. 2012) focus on discovering paths.

In terms of expressiveness, Barceló et al. (Barceló, Pérez, and Reutter 2012) showed that NREs cannot express queries involving, for instance, path conjunction. SPARQL property paths are even less expressive than NREs because of the lack of nesting and tests also within (arbitrary length) paths. TriAL can capture queries not captured by NREs and nSPARQL. However, NREs/nSPARQL queries can be evaluated in linear time. Arenas et al. (Arenas, Gottlob, and Pieris 2014) studied the relationships between TriQ, TriAL and NEMODEQ in terms of Datalog programs (Cali et al. 2010). Fletcher et al. (Fletcher et al. 2011) compared the expressiveness of navigational languages by considering different sets of features. We compare EPPs with closely-related languages in Section 3.3.

EPPs extend the expressive power of SPARQL property paths and NRE-based languages with new features such as path conjunction, path negation and powerful types of tests. These features are partially available in the W3C standard language XPath 2.0 used to query tree-like data (Berglund et al. 2010). From a concrete point of view, with EPPs we faced two main challenges: (i) how to make EPPs a conservative extension of SPARQL property paths; (ii) how to make EPPs readily available in existing SPARQL processors.

1.1 EPPs by Example
We now give an overview of EPPs via a concrete example. The syntax and semantics of EPPs will be discussed in detail.
in Section 2.

**Example 1 (Extended Property Paths).** By considering the graph in Fig. 1, find pairs of nodes (n₁, n₂) simultaneously connected via two paths: one labeled as p₅ and the other labeled as p₁₀; moreover, from n₁ an edge labeled as p₆ that leads to nodes having value < 6 must not exists.

NREs-based languages (and PPs) cannot express such request due to the lack of path conjunction/negation. With EPPs it can be expressed as:

\[ ?x ((p₅ & p₁₀) \sim (p₅ & TP(, p₆ & T (, o < 6))) ) \sim ?y \]

The path conjunction & enables to check that both paths p₅ and p₁₀ connect n₁ with n₂, while path negation \(\sim\) enables to discard from the set of nodes satisfying the first path those that also satisfy the second one. Note also that SPARQL property paths and NREs-based languages lack tests like TP, to check the existence of a path (via p₀) to nodes having value < 6.

The evaluation of the EPP expression with the algorithm described in Section 3 starts from all the bindings of the variable ?x; we discuss the case ?x \(\rightarrow\) n₀. From n₀, the evaluation of the first path (p₅ & p₁₀) enables to reach n₃. The second path (negated via \(\sim\)) is evaluated again from n₀ and consists in the logical AND (via & & k) of two tests. The first test checks for the existence of an edge p₅, which leads to n₃. The second test (i.e., TP) includes: (i) the position in the RDF triple from which the test starts; (ii) a path. The position symbol \(\omega\) means that the test starts from the object of the previous navigational step, that is, the object of the triple (n₀, p₅, n₃). From n₃, another logical AND of two tests is evaluated. The first one checks the existence of an edge p₆ and enables to reach the node 6. The second AND, which starts from the object of the previous step (i.e., 6), checks that the value is < 6; in this case the test fails and n₃ is not included in the results of the negated path. Conversely, n₄ satisfies the test, it is included in the results of the negated path and thus, it is not a valid binding for ?y when considering ?x \(\rightarrow\) n₀.

Overall, for ?x \(\rightarrow\) n₀ we get ?y \(\rightarrow\) n₃ meaning that from n₀ the navigation of the graph according to the expression enabled to reach n₃.

We have also devised a translation procedure (see Section 3.2) of EPPs into SPARQL. The translation of Example 1 is shown in Fig. 2; here, \(\omega_0, 0, 0 \sim 1, 0, 1, 0\) is a variable automatically generated by the translation. The advantage of the translation is that the query in Fig. 2 can be executed on existing SPARQL processors. From the syntactic point of view, the advantage of using EPPs to write navigational queries instead of writing them directly into pure SPARQL is that the same request can be expressed more succinctly and without the need to deal with intermediate variables.

**SELECT DISTINCT ?x ?y WHERE { { ?x p₅ ?y. ?x p₁₀ ?y. } MINUS { ?x p₅ ?y. FILTER EXISTS { ?y p₆ ?o_0_0_1_0_1_0. FILTER( ?o_0_0_1_0_1_0 < 6 ) } } }

**Figure 2:** Translation into SPARQL of Example 1.

### 1.2 Contributions and Organization

The contributions of this paper are the following: (i) EPPs a language more expressive than NREs-based languages and SPARQL property paths; (ii) a concise syntax and a formal semantics for EPPs; (iii) an algorithm for the evaluation of EPPs and an implementation in a custom query processor; (iv) a formalization and implementation of a translation procedure of EPPs into SPARQL; (v) a comparison between the custom query processor and Jena ARQ (for EPPs translated into SPARQL).

EPPs can be seen as a language per se, which goes beyond NREs-based languages and SPARQL property paths. Moreover, EPPs serve also the purpose of capturing a larger fragment of pure SPARQL navigational queries in a succinct way. To the best of our knowledge, EPPs are the first proposal covering both aspects.

The remainder of the paper is organized as follows. Section 2 presents EPPs; syntax and semantics. Section 3 introduces an evaluation algorithm for EPPs and the translation of EPPs into SPARQL. The implementation and evaluation of EPPs are discussed in Section 4. We conclude in Section 5.

## 2 Extended Property Paths

Let \(U\) (URIs), \(B\) (blank nodes), \(L\) (literals), and \(V\) (variables, starting with the ‘‘?’’ symbol) be four pairwise disjoint, countably infinite sets. An **RDF triple** is a tuple of the form \((s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)\). An RDF graph \(G\) is a set of triples. We indicate by \(nodes(G) \subseteq U \cup B \cup L\) the set of URIs, blank nodes and literals that appear as subject or object of some triples in \(G\). Moreover, \(T \subseteq U \cup B \cup L\) is the set of terms of \(G\).

### 2.1 Extended Property Paths Syntax

The syntax of EPPs is reported in Table 1.

**Commonalities with SPARQL property paths (PPs)**

The first line of the syntax covers the syntax of PPs (Harris and Seaborne 2013) (section 9.1); moreover, (Negated) PropertySets like ‘‘\(\{u_1, \ldots, u_n\}’’ that are
a combination of forward/reverse predicates in a (negated) set are expressible in the EPPs syntax via the productions test : base → u. Note that EPPs use the symbol ‘|’ while PPs use ‘\).'.

New features introduced by Extended Property Paths

The following additional features are introduced: path conjunction (epp1&epp2), path negation (epp1~epp2) and different types of tests (test) within a path, also by specifying the starting and ending positions (pos). EPPs enable to test from each of the subject, predicate and object positions in RDF triples, mapped in the syntax to .s, .p and .o, respectively. Positions do not need to be always specified; by default a test starts from the subject (.s) and ends on the object (.o) of a triple.

Tests (test) can be of different types and can be combined by using the logical operators AND (\&\&), OR (|) and NOT (!). A test can be a simple check for the existence of a URI or a nested EPP, i.e., TP(pos,epp), which corresponds to the evaluation of epp starting from a position pos (of the last triple traversed) and whose evaluation returns true if, and only if, there exists at least one node that can be reached via epp. A base test (production base) can be of type T, which is a SPARQL boolean expression; here, EExp (not reported here for sake of space) extends the production [110] in the SPARQL grammar2 where BuiltInCall3 is substituted with EBuiltInCall, which enables to use in EPPs tests available in SPARQL as built-in conditions also augmented with positions (pos). Built-in conditions are constructed using elements of the set U ∪ L and constants, logical connectives (\&, \&, |, \neg), (in)equality symbol(s) (\,<,\,\leq,\,\geq), unary (e.g., isURI,) and binary (e.g., STRSTARTS) functions.

2.2 Extended Property Paths Semantics

The semantics (shown in Table 2) for the interpretation of an EPP expression epp on a graph \( G \) uses two functions: (i) \( E[\text{epp}]^G \) defined as a binary relation \( (u,v) \) such that \( u \) and \( v \) are nodes in \( G \) and \( v \) is reachable from \( u \) via a path in \( G \) satisfying epp; and (ii) \( E[T[\text{test}]]^G \) defined as a boolean function, which evaluates true if, and only if, the triple \( t \) satisfies the test test. The semantics also uses the position mapping function \( P_m \) defined as follows:

**Definition 2 (Position mapping function).** Let \( t=(x,y,z) \) be a triple pattern, \( \{x,y,z\} \subseteq T \cup U \). The position mapping function \( P_m(\text{pos},t) \) is defined as:

(i) \( P_m(\text{s},t) = x, (ii) P_m(\text{p},t) = y, (iii) P_m(\text{o},t) = z. \)

\( P_m \) selects one among the subject, predicate and object of a triple on the basis of the value of pos. Consider the triple \( \langle u_1, p_1, u_2 \rangle \) and the test \( T(p=p_1) \); this instantiates \( P_m \) as \( P_m(\text{p}, \langle u_1, p_1, u_2 \rangle) = p_1 \), which checks \( p_1 = p_1 \) returning true while testing \( T(\text{o}=u_3) \) gives false.

3 Algorithms and Complexity

This section presents two strategies for the evaluation of EPPs expressions. The first via an ad-hoc algorithm; the second one via a translation into SPARQL queries that can be executed on existing processors.
The W3C spec. (Harris and Seaborne 2013) informally mentions the fact that non-recursive property paths can be evaluated via a translation into equivalent SPARQL algebraic forms. However, no formal proof of the correctness and completeness of such translation is provided. Recursive property paths are handled in the standard via auxiliary functions called ALP (Harris and Seaborne 2013). With EPPs
Consider a NREPP pattern $\mathcal{P}$. In the operational tree associated to $\mathcal{P}$, nodes are of two types: (i) operational nodes (labeled with $\sim$, $\&, /$, $|$) and (ii) test nodes (labeled with test, $\neg$, $\&, \TP$, $u$, $T$($EExp$)). Each node $n_i$ in the operational tree has a single parent and its children are an ordered set. Moreover, $n_i$ has an attribute $\id$ whose unique value is computed as the concatenation of the parent’s $\id$ and $n_i$’s position in the ordered set of children. The operational tree for Example 1 is shown in Fig. 4.

(ii) Propagation of variables and terms

The key point is the propagation of variable names and RDF terms (kept in nodes’ attributes). This is done by traversing the operational tree top-down and propagating attributes from each parent node to its children.

(iii) Application of the translation rules

This phase always starts by applying rule $R_{m}$ in Fig. 5 on the root of the operational tree. This generates the outermost part of the final rewritten query into the SPARQL syntax. The translation proceeds by applying rules at each node of the operational tree visited according to a pre-order depth-first traversal.

For instance, in Fig. 4, after the root, the node with $\id=0.0$ and labeled with $\sim$ is visited. This causes the triggering of rule $R5$, which enables to generate another (internal) chunk of the final SPARQL query. Node 0.0 is an operational node representing path negation $\sim$. As it can be noted, our translation procedure uses the SPARQL $\textit{MINUS}$ operator to reflect the semantics of EPPs dealing with path negation. The semantics of nodes representing EPPs test (e.g., $0.0_{o}1.0_{1}$) is reflected into SPARQL via the $\textit{FILTER}$ operator.

Correctness of the translation

In order to prove the correctness of the translation of NREPPs into SPARQL we have defined a SPARQL based semantics (not reported here for sake of space) where conjunction ($\&$) and negation ($\sim$) are translated into join ($\bowtie$) and difference ($\setminus$) of (multi)sets of solution mappings (Pérez, Arenas, and Gutierrez 2009). Let $\mathcal{P}^{\mathcal{G}}_{\text{EPPs}}$ denote the SPARQL-based semantics of EPPs, where $\mathcal{P}$ is an NREPP pattern, and $\mathcal{S}_{\mathcal{G}}^{W3C}$ denote the SPARQL W3C semantics (Harris and Seaborne 2013), where $\mathcal{S}$ is a SPARQL query. The following theorem shows the correctness of the translation.

Theorem 4 The translation algorithm $\mathcal{A}^{t}$ is correct and runs in polynomial time in the size of the expression to be translated. Moreover, for any RDF graph $\mathcal{G}$ given $\mathcal{P}$ it holds that $\mathcal{P}^{\mathcal{G}}_{\text{EPPs}}=\mathcal{S}_{\mathcal{G}}^{W3C}$. where $\mathcal{S}=\mathcal{A}^{t}(\mathcal{P})$ is the query produced by translating $\mathcal{P}$.

Proof: [SKETCH] $\mathcal{A}^{t}$ is polynomial as it requires one scan of the operational tree (whose size is polynomial in the size of the EPP expression). As for the correctness, it is enough to prove that the propagation of variable names and RDF terms in the operational tree is correct. This proof can be
done by structural induction on the depth of the operational tree. Finally, the semantic equivalence can be proved by associating to all types of EPP patterns the corresponding semantics as per SPARQL specification (Harris and Seaborne 2013).

**Complexity of the translated queries.** Consider a NREPP \( epp \) in a pattern \( P \) and its translation into SPARQL \( S_{epp} \). Clearly, the complexity of evaluating \( S_{epp} \) depends on the fragment of SPARQL used in the translation. In particular (see Fig. 5) we make usage of \( \text{SELECT, UNION, MINUS and FILTER} \) and we do not use \( \text{OPTIONAL} \). The complexity of this and other SPARQL fragments has been studied in (Pérez, Arenas, and Gutierrez 2009).

**Summary.** EPPs can be seen as a language per se and can be evaluated by implementing the algorithm in Fig. 3 in a custom processor, thus being independent from existing SPARQL processors. EPPs can also be seen as a conservative extension of SPARQL property paths and, thanks to the translation algorithm, can be evaluated on existing SPARQL processors. We have implemented both strategies; their pros and cons will be discussed in Section 4.

### 3.3 Comparison with Related Languages

As the goal of EPPs is to extend the expressive power of SPARQL property paths (PPs) and NREs-based languages, we compared such proposals with EPPs; Table 3 summarizes the (informal) comparison. We consider the following features of EPPs: path conjunction (\&), path negation (\(\sim\)), nesting (TP), tests over node values (test) and usage of positions (pos). As it can be observed, PPs are the least expressive language; they do not support any of the new features of EPPs. This motivated our choice of extending PPs as described in Section 2.1. As for NREs, they clearly support nesting but neither other types of tests (e.g., node equality) nor path conjunction/negation as discussed in (Barceló, Pérez, and Reutter 2012).

<table>
<thead>
<tr>
<th>Language</th>
<th>&amp;</th>
<th>(\sim)</th>
<th>TP</th>
<th>test</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPPs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PPs</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NREs</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>nSPARQL</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3: Comparison of EPPs with related languages.

As for nSPARQL (based on NREs), it supports path conjunction and negation only via the SPARQL algebra; it also supports nesting and positions. However, it does not allow to test (in)equalities of nodes reached with a nested expression. EPPs support logical combination of tests representing nesting and tests representing (in)equalities as well as safe-negation\(^6\). Note that neither PPs nor NREs nor nSPARQL can express Example 1. nSPARQL supports positions by transforming an RDF graph into another graph where navigational axes (similar to those defined in XPath) are made explicit. On the other hand, EPPs do not require any transformation. In the syntax of EPPs, positions enable to perform edge to node traversals allowing to reach the node representing the predicate of a triple whence it is possible to traverse e.g. the property hierarchy. Finally, we want to emphasize that EPPs are syntactically compatible with PPs and are the only language, which thanks to the \(\text{NREPPtoSPARQL} \) translation, can be used in existing SPARQL processors.

### 4 Implementation and Evaluation

We have implemented both a custom query processor for EPPs and the \(\text{NREPPtoSPARQL} \) translation\(^7\).

**Dataset and query set.** We used a crawl of the FOAF social network (\(~500MBs\)) obtained from the BTC2012\(^8\) by traversing from the URI of T. Berners-Lee (TBL) foaf:knows predicates up to distance 4. We call the resulting graph \(G^F\), which has \(\sim4M\) triples. We created 4 groups \(G_i, i \in \{1,\ldots,4\}\) of EPP expressions each with 3 queries; this gives a queryset \(Q\) of 12 queries. The experiments have been performed on an Intel i5 machine with 8GBs RAM. Results are the average of 5 runs after removing lowest and highest values.

**Experiment 1: Running time**

For each \(epp \in Q\) we generated the corresponding SPARQL query \(S_{epp}\) via the algorithm described in Section 3.2. We measured the execution time for each \(epp \in Q\) with our processor and the execution time for \(S_{epp}\) in Jena ARQ\(^9\). Fig. 6 shows the running times. The number of results ranges from \(\sim50\) to \(\sim8000\).

![Figure 6: EPPs-custom-processor vs. Jena ARQ.](image)

As it can be observed, for \(G_1\), which contains queries asking for friends of TBL up to distance 3, our evaluator performs better than ARQ at distance 1 and 2; at distance 3 times are comparable. The group \(G_2\) additionally considers a test based on nesting. Again, our evaluator performs better at distance 1 and 2; at distance 3 it shows a higher running time. In \(G_3\), which considers negation (e.g., exclusive friends at various distances) our approach performs consistently better. Finally, in \(G_4\) that includes conjunction (to ask for mutual friends at various distances) our approach performs better at distance 1 and 2 and obtains a higher running time at distance 3.

---

\(^6\)The first element must be positive path.

\(^7\)Available at \[http://extendedpps.wordpress.com\]

\(^8\)http://km.aifb.kit.edu/projects/btc-2012

\(^9\)http://jena.apache.org
The complexity of the algorithm described in Section 3 is lower than that of Jena ARQ, which implements SPARQL. However, our evaluation suggests that for real-world data and natural queries (e.g., mutual friends) working with the translation of EPPs and using existing processors is still useful. Hence, the advantage of our approach is that navigational queries can be written in a succinct way via the EPPs syntax and execute after their translation via NREPToSPARQL. Anecdotally, while the EPP asking for mutual friends at distance 3 contains \(\sim 200\) characters, the SPARQL query (obtained from the translation) contains \(\sim 700\) characters. Moreover, when writing the SPARQL query one has also to deal with a large number of variables that need to be consistently joined.

**Experiment 2: Translation overhead**

We compared the elapsed times of our NREPToSPARQL with the SPARQLtoALGEBRA translation of Jena ARQ. We used 28 queries generated in two steps. We started with 4 base expressions plus a fifth one combining them. Second, starting from them we generated increasingly longer expressions. We found that our NREPToSPARQL translation performs comparably (in the order of ms) to the existing SPARQLtoALGEBRA translation.

5 Conclusions and Future Work

We introduced Extended Property Paths (EPPs), a significant extension of SPARQL property paths. We have provided an algorithm for their evaluation and also a translation procedure into SPARQL.

EPPs bring some benefits: (i) users can leverage their syntax to write more expressive navigational queries in a succinct way; (ii) EPPs are immediately available into existing SPARQL processors. Our study opens some research questions such as investigating properties of the translation (e.g., minimality) and studying optimization techniques. Also containment of EPPs is an open question.

References


Harris, S., and Seaborne, A. 2013. SPARQL 1.1 Query Language, W3C.


