Querying for Meta Knowledge

Bernhard Schueler, Sergej Sizov, Steffen Staab
ISWeb — Information Systems and Semantic Web
University of Koblenz-Landau, Germany
{bernie,sizov,staab}@uni-koblenz.de

Duc Thanh Tran
Institute AIFB
University of Karlsruhe, Germany
dtr@aifb.uni-karlsruhe.de

Abstract

The Semantic Web is based on accessing and reusing RDF data from many different sources, which one may assign different levels of authority and credibility. Existing Semantic Web query languages, like SPARQL, have targeted the retrieval, combination and reuse of facts, but have so far ignored all aspects of meta knowledge, such as origins, authorship, recency or certainty of data, to name but a few.

In this paper, we present an original, generic, formalized and implemented approach for managing many dimensions of meta knowledge, like source, authorship, certainty and others. The approach re-uses existing RDF modeling possibilities in order to represent meta knowledge. Then, it extends SPARQL query processing in such a way that given a SPARQL query for data, one may request meta knowledge without modifying the query proper. Thus, our approach achieves highly flexible and automatically coordinated querying for data and meta knowledge, while completely separating the two areas of concern.

1. Introduction

Integrating and re-using Semantic Web data becomes more and more fruitful and worthwhile in order to answer questions and deliver results. Typically, engines like Swoogle provide points of access for RDF data, crawlers may fetch relevant RDF data, and query languages like SPARQL with their corresponding query engines allow for selecting and re-using data in the appropriate format. With the arrival of more and more data in the Semantic Web and more sophisticated processing through query and reasoning engines, one now, however, encounters challenging questions linked to meta knowledge about the data like:

- Where is this data from?
- Who provided the data?
- When was this data provided?
- Was the provider certain about the truth of this data?
- Was the data believed by others, too?

For instance, when querying the Semantic Web with the help of SPARQL for the affiliation of a person by the name of "James Hendler", one finds (at least) two answers, i.e. "University of Maryland" and "Rensselaer Polytechnic Institute". Obviously, without further indication as to where, by whom, when, etc. such information was given, it is impossible to decide which of the two affiliations is still valid.

The problem might be remedied in several ways. First, an ideosyncratic solution by the search engine, such as returning the corresponding RDF files or links to sources of knowledge extraction (say `http://www.cs.umd.edu/survey.pdf` and `http://www.rpi.edu/report.doc`), might help in this special case. However, an ideosyncratic solution may not be appropriate in a second case in which the 'when' was more relevant than the 'where' or in a third case where such a piece of information had to be aggregated from several resources. Second, the person or system requesting the meta knowledge might manually extend the SPARQL query formalizing the request for the affiliation in order to return the where, the who and the when. Such a modification will, however, be very tedious, as it will include a number of additional optional statements, and expressing it manually will be error prone. Also, it will not help in delivering meta knowledge that arises from joining several statements, e.g. meta knowledge about uncertainty that was based on several meta knowledge statements with different values of uncertainty.

Therefore, querying Semantic Web data requires a principled, generic approach to the treatment of meta knowledge that is able to adapt to many dimensions of meta knowledge and that is open to accommodate to new dimensions when the need arises. Such a principled, original framework is given in this paper. We start to explain our approach with a discussion of important design choices in section 2. We model meta knowledge in existing RDF structures by embedding a slightly more expressive language, which we call RDF$^+$, into RDF[1]. We define the abstract syntax of RDF$^+$, its semantics and its embedding in RDF in Section 4. In Section 5, we extend the SPARQL syntax and semantics to work on data and meta knowledge of RDF$^+$. The extension allows the user to extend a given conventional SPARQL query by a keyword for meta knowledge triggering the construction of meta knowledge by the query processor. Section 6 summarizes the overall use and processing of SPARQL queries with meta knowledge. Section 7 reports on initial graceful results for meta knowledge processing from a theoretic point of view and Section 8 provides pointers to

Copyright is held by the author/owner(s).
WWW2008, April 21–25, 2008, Beijing, China.
.

[1] This proposal is a completely revised and extended version of [17]. Major revisions include a novel formal model, discussion of the design space, complexity analysis, and prototype implementation.

the prototype implementation of the system.

2. Scenario

In our sample application scenario, we assume that the user utilizes knowledge which has been initially extracted from Web pages of Computer Science departments and stored in form of RDF triples in his personal "active space" [16], backed by a local RDF repository. Example 2.1 shows the relevant facts that may have been obtained from departments of different universities. For better readability, we use for our examples in this paper the RDF triple language Turtle [8] with Named Graphs [1, 4] in a simplified form that abstracts from default namespaces.

Example 2.1. *Extracted Knowledge and SPARQL query*

```
G1 { JamesHendler researchTopic SemanticWeb .
     JamesHendler affiliatedWith RensselaerPI }

G2 { JamesHendler researchTopic KnowledgeDiscovery .
     JamesHendler affiliatedWith UnivMaryland  .
     RudiStuder researchTopic SemanticWeb .
     RudiStuder affiliatedWith UnivKarlsruhe }
```

The extracted knowledge comes from different sources, at different timepoints, and with different degrees of extraction confidence. This information is also captured and stored into the same RDF repository as shown in example 2.2, using the notion of Named RDF Graphs [1, 4].

Example 2.2. *Associated meta knowledge*

```
G3 {
     G1 mk:source <www.rpi.edu/report.doc> .
     G1 mk:certainty "0.9" .
     G1 mk:timestamp "5/5/2007"
   }
G4 {
     G2 mk:source <www.cs.umd.edu/survey.pdf> .
     G2 mk:certainty "0.6" .
     G2 mk:timestamp "6/6/2001"
   }
```

In our scenario, the sample user aims to explore the knowledge and meta knowledge using the RDF query language SPARQL. We assume that he aims to find experts in the domain of Semantic Web and their affiliations. The corresponding SPARQL query is shown in example 2.3. In addition, the user wants to exploit meta knowledge from example 2.2 for obtaining results with best certainty and for analyzing contradictive answers (e.g. different affiliations for the same person "James Hendler" in example 2.1).

Example 2.3. *Extracted Knowledge and SPARQL query*

```
CONSTRUCT (?X worksAt ?Y)
FROM NAMED G1,G2
WHERE
{
    {
      {?X researchTopic  ?Y } AND
      {?X affiliatedWith ?Z }
    }
    FILTER  {?Y = SemanticWeb}
}
```

3. Design choices

This section summarizes and shortly motivates the design choices for our meta knowledge framework.

Reification. Establishing relationships between knowledge and meta knowledge requires appropriate reification mechanisms for supporting statements about statements. Our general objective is to execute queries for data proper (i.e. without meta knowledge) directly, without complex transformations. For compliance with existing applications that access the repository in a common way (e.g. using SPARQL queries), we do not modify existing user data. This requirement does not allow us to use mechanisms like RDF reification, which decompose existing triples and fully change the representational model. In our framework described in section 4, we adopt the notion of Named RDF Graphs for meta knowledge representation [1, 4].

Storage mechanisms. Following the overall philosophy of RDF, we do not separate meta knowledge from "normal" user knowledge in the repository. Following this paradigm, a user or developer has unlimited access to all contents of the triple store and can manipulate meta knowledge directly. In other words, the user can directly access meta knowledge (e.g. using suitable SPARQL queries). Beyond explicitly designed queries for meta knowledge access, in Section 5 we describe the extension of SPARQL that allows us to access meta knowledge about the result set automatically without user intervention.

Dimensions of Meta Knowledge. An important point for the application design is the definition of relevant meta knowledge properties and their suitable interpretation for arbitrary complex query patterns. In general, meta knowledge properties are application dependent and must be carefully chosen by the system administrator. In our sample scenario (sections 2 and 6) we discuss common and widely used properties, such as timestamp, source, and (un)certainty, and show ways of defining and utilizing them in our framework.

Syntax extensions. Seamless integrated access to meta knowledge requires corresponding extensions of existing querying mechanisms. These can be realized at different levels, e.g. query language (e.g. SPARQL) or application-specific interfaces (e.g. Sesame API). In Section 5 we describe our SPARQL extension for constructing query results with associated meta knowledge. It is system-independent and not related to some particular implementation of the RDF repository. Furthermore, it fully supports the existing SPARQL syntax and semantics. Compliance with existing established standards makes the integration with existing applications and interfaces substantially easier.

4. Syntax and Semantics for RDF with Meta-knowledge

In the course of representing and reasoning with meta knowledge we embed a language with meta-knowledge reasoning, i.e. RDF$^+$, in a language without such specific facilities, i.e. in RDF. This embedding implies that we may consider an RDF snippet in its literal sense *and* we may possibly interpret it as making a meta-knowledge statement. Embedding meta knowledge in RDF is not the most expressive means to deal with all needs of meta-knowledge processing, but it retains upward compatibility with existing usage of the language and corresponding tools and methods, which is a major concern for Semantic Web approaches.

Though we denote meta knowledge in RDF, we must distinguish the notation of RDF with only *implicit* notation of meta-knowledge, but no semantic consequences specifically due to this meta knowledge, from a formally extended model of RDF with *explicit* notation of meta-knowledge. The following definition of RDF$^+$ helps us to draw this line very clearly and concisely. The abstract syntax for this embedded language, RDF$^+$, is given in Section 4.1 and its semantics in Section 4.2. Eventually in this section, we show how to embed RDF$^+$ in RDF with named graphs.

4.1 An Abstract Syntax for RDF$^+$

The abstract syntax of RDF$^+$ is based on the same building blocks as RDF:

- U are Uniform Resource Identifiers (URIs).
- L are all RDF literals.
- $G \subseteq U$ is the set of graph names.
- $P \subseteq U$ is the set of properties.

In addition, we must be able to refer to statements directly without use of reification. For this purpose, we introduce statement identifiers:

- Θ is a set of statement identifiers, which is disjoint from U and L.

Now, we may define RDF$^+$ literal statements that are placed in named graphs and have, in addition to RDF, a globally unique statement identity.

Definition 4.1 (RDF$^+$ Literal Statements).
The set of all RDF$^+$ literal statements, $\mathfrak{S}$, is defined as quintuples by:
$\mathfrak{S} := \{(g, s, p, o, \theta)|g \in G, s \in U, p \in P, o \in U \cup L, \theta \in \Theta\}$.

Thereby, θ and (g, s, p, o) are keys such that there exists a bijection f_1 with $f_1(g, s, p, o) = \theta \wedge f_4(\theta) := f^{-1}(\theta) = (g, s, p, o)$. Moreover, we define the overloaded function f_5 to return the complete quintuple given either θ or (g, s, p, o), i.e. $f_5(\theta) := (g, s, p, o, \theta) =: f_5(g, s, p, o)$, when $f_1(g, s, p, o) = \theta$.

The reader may note that we assume that f_1 is fixed and given before any statement is defined. Furthermore, this definition of literal statements and the rest of this paper abstracts from RDF blank nodes in order to keep the formalization more concise. However, we do not see any principle problem in extending our treatments to blank notes, too.

The two statements of Graph G1 of Example 2.1 may now be represented in RDF$^+$ in the following way.

Example 4.1.
$\mathfrak{S} \supseteq K \supseteq \{$
(G1, JamesHendler, researchTopic, SemanticWeb, θ_1),
(G1, JamesHendler, affiliatedWith, RensselaerPI, θ_2) $\}$

Thereby, the exact form of statement identifiers in Θ is up to the implementation, as they are only used for internal processing.

Having represented the literal interpretation of RDF statements in RDF$^+$, we may now address the representation of selected RDF statements as RDF$^+$ meta knowledge. This is done using a structure of RDF$^+$ meta-knowledge statements, $\mathfrak{M}$, that is separate from the set of RDF$^+$ literal statements:

Definition 4.2 (RDF$^+$ Meta-knowledge Statements).
Let $\Pi \subseteq P$ be the set of meta-knowledge properties. Let $\Omega_\pi \subseteq U \cup L$, with $\pi \in \Pi$, be sets providing possible value ranges for the meta-knowledge properties $\pi \in \Pi$.

Then, the set of all RDF$^+$ meta-knowledge statements, $\mathfrak{M}$, is defined by: $\mathfrak{M} := \{(\theta, \pi, \omega)|\theta \in \Theta, \pi \in \Pi, \omega \in \Omega_\pi)\}$.

The following example partially demonstrates the target representation of the first two meta-knowledge statements of graph G3 from Example 2.2.

Example 4.2.
$\mathfrak{M} \supseteq M \supseteq \{$
(θ_1, mk:source, <www.rpi.edu/report.doc>),
(θ_1, mk:certainty, 0.9)$\}$

Together we may now define a RDF$^+$ theory.

Definition 4.3 (RDF$^+$ Theory).
A RDF$^+$ theory of literal statements and associated meta knowledge statements is a pair (K, M) referring to a set of literal statements $K \subseteq \mathfrak{S}$ and a set of meta-knowledge statements $M \subseteq \mathfrak{M}$.

A (partial) example for such a theory is given by the pair (K, M) with definitions for K and M as given in examples 4.1 and 4.2, respectively.

4.2 A Semantics for RDF$^+$

Though we now have an abstract syntax for representing literal statements like "JamesHendler researchTopic SemanticWeb" in G1 and meta-knowledge statements like "the source of the statement that James Hendler's research topic is Semantic Web is found in the document
<www.rpi.edu/report.doc>"
in G3, such an abstract syntax may remain remarkably ambiguous if it cannot be linked to a formal semantics. For instance, given two meta-knowledge statements:
(θ_1, mk:source, <www.rpi.edu/draftReport.doc>) and
(θ_1, mk:source,<www.rpi.edu/finalReport.doc>)
for the same literal statement identified by θ_1, the question may arise whether this means a disjunction, i.e. one of the two documents has provided the fact, or a conjunction, i.e. both documents have provided the fact, or a collective reading, i.e. the two documents together gave rise to the fact, or whether this situation constitutes invalid meta-knowledge.

In order to prevent such ambiguities we introduce a generic semantic framework for meta-knowledge in RDF$^+$. However, the framework must also be able to reproduce the literal interpretations found in RDF. For the latter purpose, we first define a 'standard' model for a RDF$^+$ theory.

Definition 4.4 (Standard Interpretation and Model).
A standard interpretation $I_s : \mathfrak{S} \rightarrow \{\top, \bot\}$ for a structure (K, M) assigns truth values to all statements[2] *in K.*

A standard interpretation is a standard model if and only if it makes all statements in K become true.

[2] Note that because f_1 is fixed there are no two tuples $(g, s, p, o, \theta_1), (g, s, p, o, \theta_2)$, where $\theta_1 \neq \theta_2$. This implies that the standard interpretation is independent of the identifiers θ_1, θ_2.

For instance, any standard model I_s for (K, M) in example 4.1 would include

(G1, JamesHendler, researchTopic, SemanticWeb, θ_1)

in its set of literal statements evaluating to $\top$.

In order to address the level of meta-knowledge we foresee an additional model layer that provides a different interpretation to each meta-knowledge property.

Definition 4.5 (Π-Interpretation and Model). *A* Π*-interpretation* $I_\pi : \mathfrak{S} \rightharpoonup \Omega_\pi$ *for a property* $\pi \in \Pi$ *is a partial function mapping statements into the allowed value range of* π*.*

A Π*-interpretation* I_π *is a* Π*-model for* (K, M) *if and only if for all meta-knowledge statements* $(\theta, \pi, \omega) \in M$ *where* $f_1(\theta) = (g, s, p, o)$ *the value of the interpretation coincides with* ω*, i.e.* $I_\pi((g, s, p, o, \theta)) = \omega$*.*

As an example, consider the source interpretation $I_{mk:source}$ of the literal statement (G1, JamesHendler, researchTopic, SemanticWeb, θ_1) from Examples 4.1 and 4.2. A model I would map this literal statement using $I_{mk:source}$ onto <www.rpi.edu/report.doc>.

The literal and the meta-knowledge interpretations may now be combined to define what an overall, unambiguous model is:

Definition 4.6 (Meta-knowledge Interpretation and Model). *A meta-knowledge interpretation* $\Im$ *is a set including a standard interpretation* I_s *and the* Π*-interpretations* I_π *for all meta-knowledge properties* $\pi \in \Pi$*.*

A meta-knowledge interpretation $\Im$ *is a model for a theory* (K, M) *if and only if all its interpretations* $I \in \Im$ *are standard models or* Π*-models for* (K, M)*.*

4.3 Mapping between RDF and RDF$^+$

The mapping between RDF and RDF$^+$ needs to be defined in two directions. First, one must be able to map from RDF as given in the examples from Section 2 to RDF$^+$. Second, one must be able to map from RDF$^+$to RDF. Because RDF$^+$is more fine-grained than RDF, the first direction will be easy, while the second will have to make compromises on the granularity of the representation.

4.3.1 *From RDF to RDF$^+$*

The examples of Section 2 reify groups of statements, i.e. the ones found in G1 and G2, in order to associate meta-knowledge, such as given in G3 and G4. In order to allow for an interpretation of the meta knowledge as defined in the preceding section, we map RDF into RDF$^+$. For all RDF statements, including statements in graphs G1 and G2 of Example 2.1, the mapping performed is close to an identity mapping. One only needs to add statement identifiers. The result for G1 in RDF$^+$ is:

Example 4.3.
$K \supseteq$ { (G1, JamesHendler, researchTopic, SemanticWeb, θ_1),
(G1, JamesHendler, affiliatedWith, RensselaerPI, θ_2) },
with
$\theta_1 := f_1$(G1, JamesHendler, researchTopic, SemanticWeb) and
$\theta_2 := f_1$(G1, JamesHendler, affiliatedWith, RensselaerPI)

The same mapping – close to the identity mapping – is performed for meta-knowledge statements like statements of graph G3, resulting in their representation as literal statements:

Example 4.4.
$K \supseteq$ {
(G3, G1, mk:source, <www.rpi.edu/report.doc> , θ_3),
(G3, G1, mk:certainty, "0.9", θ_4), ...}

Note that this step is necessary in order to achieve upward and – limited – downward compatibility between RDF$^+$ and RDF. Of course, the interpretation of statements, like the ones found in G3, also require an interpretation as meta knowledge. This is achieved by mapping RDF statements with designated properties from Π like 'mk:source' and 'mk:certainty' to the additional meta-knowledge layer:

Example 4.5.
$M \supseteq$ {
(θ_1, mk:certainty, "0.9"),
(θ_1, mk:source, <www.rpi.edu/report.doc>), ...}

The mapping of predicates and objects of these meta-knowledge statements from RDF to RDF$^+$ are obvious, they are mapped to itself. For the subjects, however, there arise modeling choices. For instance, if 'mk:certainty' were interpreted using probability theory, one might assign a distributive or a collective reading. In the distributive reading, each fact in G1 receives the probability value of 0.9 and, eventually, the distributive reading will assign a joint probability of close to 0 for a large number of n stochastically independent facts, i.e. the joint probability 0.9^n. In the collective reading, the collection of facts in G1 as a whole will receive the probability value 0.9. Therefore, the collective reading will assign an individual certainty close to 1 for each individual fact, when the number of facts is high and each fact is independent from the others, i.e. the individual probability would be $\sqrt[n]{0.9}$. *A priori*, none of the two (and more) modeling choices is better than the other, but they constitute different modeling targets.

The mapping from RDF to RDF$^+$ for the *distributive reading* of a meta property π is easy to achieve.

Definition 4.7 (Distributive Embedding). *Given an RDF statement ":G {S P O}" and an RDF meta-knowledge statement ":H {:G* π ω*}", a distributive embedding of RDF$^+$ in RDF adds the meta-knowledge statement* $\{(\theta, \pi, \omega) | \theta = f_1(G, s, p, o) \wedge f_5(\theta) \in K\}$ *to* M*.*

This means that such a meta-knowledge statement is applied individually to all statements in the graph to which it refers in RDF, as indicated in the example above.

In order to assign a joint probability of 0.9, one must assign this value to the conjunction of statements, i.e. $f_5(\theta_1) \wedge f_5(\theta_2)$ must be assigned probability value 0.9. An interpretation for this will be given in Section 5.

4.3.2 *From RDF$^+$ to RDF*

The serialization of RDF$^+$ data in the knowledge base K is straightforward. Each quintuple (g, s, p, o, θ) is realized as a corresponding triple in a named graph and the tuple identifier θ is discarded.

Example 4.6.
(G5, JamesHendler, researchTopic, SemanticWeb, θ)
is mapped to
G5 {JamesHendler researchTopic SemanticWeb }

For meta-knowledge statements the situation is more challenging, because literal statements with different statement identifiers may belong to only one named graph. Their corresponding meta-knowledge statements may differ, but the realization of the meta-knowledge statements in RDF does not allow for retaining these fine-grained distinctions – unless one chooses to change the modeling approach drastically, e.g. by assigning each literal statement to a named graph of its own, which seems undesirable (cf. discussion in Section 3).

We have preferred to pursue a more conventional modeling strategy for RDF with named graphs. Therefore, we weaken the association between meta knowledge statements and their corresponding literal statements when mapping to RDF. I.e. we group sets of meta-knowledge property values into one complex value.

Definition 4.8 (Generating Grouped Meta-knowledge). *Given a RDF*$^+$ *theory* (K, M), *RDF meta-knowledge is generated by grouping RDF*$^+$ *meta-knowledge statements as follows:*
Add one RDF statement hashGraph(g)$\{g\ \pi\ \omega'\}$ *for each*

$$\omega' := group_\pi(\bigvee_{\exists\theta,\omega: f_5(\theta)\in K, (\theta,\pi,\omega)\in M} \omega)$$

where group is a meta-property specific grouping function and hashGraph *is a function mapping existing graph names onto graph names suitable for associating meta knowledge.*

The suitability of hashGraph may be application specific. A general strategy may map graph names g to graph names prefixed by <http://metaknowledge.semanticweb.org> in a deterministic manner.

In the following example the grouping of meta-knowledge values is illustrated.

Example 4.7.

```
K:={
(G5, JamesHendler, researchTopic, SemanticWeb, θ1),
(G5, JamesHendler, affiliatedWith, UnivMaryland, θ2) },
M:={
(θ1, mk:source, <www.rpi.edu/report.doc>),
(θ2, mk:source, <www.cs.umd.edu/survey.pdf>) }
                          is mapped to
:G5 {
JamesHendler researchTopic SemanticWeb .
JamesHendler affiliatedWith UnivMaryland }
:G6 {
G5 mk:source {<www.rpi.edu/report.doc>,
<www.cs.umd.edu/survey.pdf>} }
```

In Example 4.7, the resulting grouped value is simply denoted as a set of source documents, {<report.doc>, <survey.pdf>}. The property specific grouping function must provide the mechanism to represent such a grouped value in an appropriate RDF data structure, e.g. as a rdf:bag or as an appropriate datatype.

5. SPARQL for RDF and Meta-Knowledge

In this section we first introduce a small extension to standard SPARQL syntax [15] and then define how SPARQL can be applied to an RDF$^+$ knowledge base. The objective of our considerations is the derivation of meta-knowledge about query results.

5.1 SPARQL Syntax Revisited

When using SPARQL to query RDF$^+$ we propose only two modifications to obtain meta knowledge. First, we introduce one additional expression "WITH META *MetaList*". This expression includes the named graphs specified in *MetaList* for treatment as meta knowledge. This statement is optional. When it is present the SPARQL processor may digest the RDF$^+$ meta-knowledge statements derivable from the RDF named graphs appearing in the *MetaList*. The SPARQL processor will then use this meta knowledge to compute and output all the meta knowledge statements deriveable by successful matches of RDF$^+$ literal statements with the WHERE pattern.

In order to determine which literal statements should be considered we introduce a second modification. We do not process FROM expressions with our meta knowledge framework, but only FROM NAMED. The reason is that FROM g expressions replicate all RDF triples of g into the default triple space of the query. Thereby, they remove the links between the RDF statements of g and possible meta knowledge. Hence, FROM expressions are not relevant for our treatment of meta-knowledge, but of course they may still be processed using the standard SPARQL semantics.

Thus, SPARQL queries on RDF$^+$ have one of the two following overall forms:

Definition 5.1 (SPARQL SELECT Query).
The structure of a SPARQL SELECT query has the following form:

```
SELECT selectExpression
[WITH META MetaList]
FROM NAMED GraphList
WHERE P
```

Definition 5.2 (SPARQL CONSTRUCT Query).
The structure of a SPARQL CONSTRUCT query has the following form:

```
CONSTRUCT constructExpression
[WITH META MetaList]
FROM NAMED GraphList
WHERE P
```

In these definitions, P refers to a graph pattern that explains how RDF$^+$ literal statements from the given *GraphList* are matched. Matches bind variables that are used for providing results according to the *selectExpression* or the *constructExpression*.

5.2 SPARQL Semantics Revisited

In this subsection we define the semantics of SPARQL queries evaluated on an RDF$^+$ theory. For our definitions we use two building blocks: algebraic semantics of SPARQL [11, 13] and the *how-provenance* calculated via annotated relations (cf. [7]).

The algebraic semantics of SPARQL queries are given based on set-theoretic operations for sets of variable assignments (cf. [11, 13]). Thereby, a variable assignment is a partial function $\mu : V \to U \cup L$, where V is the set of variables given in a SPARQL query. A set of variable assignments can be represented by a relation ϕ over the domain $(U \cup L)^{|V|}$, where the variables V are the attributes and assignments are the tuples of this relation. Such a set of assignments may be assigned information about the so called *how-provenance* [7], i.e. the assignments may be annotated with formulae describing the individual derivation tree used to assign the variables. The how-provenance annotation may be represented by a function $\Phi : (U \cup L)^{|V|} \to \mathfrak{F}$, where $(U \cup L)^{|V|}$ is the set of all tuples of the length $|V|$ over the domain $U \cup L$ and $\mathfrak{F}$ is the set of formulae annotating variable assignments. The set of formulae $\mathfrak{F}$ is given by all Boolean formulas constructed over the set of literal statements $\mathfrak{S}$ and including a bottom element $\bot$ and a top element $\top$. The formulae constitute an algebra $(\mathfrak{F}, \wedge, \vee, \neg, \bot, \top)$. The special element $\bot$ is used as annotation of variable assignments which are not in the relation ϕ. The special element $\top$ may be omitted, but it allows for simplification of complex formulas.

Assume the following SPARQL query to be evaluated on the RDF$^+$ knowledge base K:

Example 5.1.

```
SELECT ?g ?x ?y
FROM NAMED G1, G2
WHERE {
    GRAPH ?g {?x researchTopic ?y .} }
```

```
K = {
(G1, JamesHendler,researchTopic,SemanticWeb, θ1),
(G1, JamesHendler,affiliatedWith,RensselaerPI, θ2),
(G2, JamesHendler,researchTopic,KnowledgeDiscovery, θ3)
...  }
```

For the query of example 5.1, we may find the following variable assignments using standard SPARQL processing and we may indicate, which atomic formulae, i.e. RDF$^+$ quintuples in this simple example, led to these variable assignments. This indication is given by the statement identifiers representing their statements.

Example 5.2.

$\Phi =$

?g	?x	?y	$\mathfrak{F}$
G1	JamesHendler	SemanticWeb	θ_1
G2	JamesHendler	KnowledgeDiscovery	θ_3
G2	RudiStuder	SemanticWeb	θ_5

.

This simple example of how a set of variable bindings has been produced is generalized to SPARQL queries of arbitrary complexity by a recursive definition of simultaneous query evaluation and computation of the annotations. The first step in evaluating a graph pattern is to find matches for the triple pattern contained in the query. Because the RDF$^+$ knowledge base K consists of quintuples, we need to adapt the SPARQL evaluation procedures. The statement identifiers do not need to be matched, as they depend functionally on graph name, subject, predicate and object. Therefore, we consider matching of quadruple patterns $(\gamma, \alpha, \beta, \delta)$. As a simplification of our formalization we assume that the keyword GRAPH together with a URI or a graph variable is used in any given SPARQL query. If it is not used, we may expand a given SPARQL query to include it.

Definition 5.3 (Basic Quadruple Pattern Matching). *Let K be a knowledge base of RDF$^+$ literal statements and μ be a variable assignment.*

The evaluation of the SPARQL query "GRAPH γ $\{\alpha\beta\delta\}$" *over* K, *denoted by* $[\![\text{GRAPH } \gamma\ \{\alpha\beta\delta\}]\!]_K$ *is defined by the annotated relation* Φ, $\text{dom}(\Phi) = \{\mu \mid \text{dom}(\mu) = \text{vars}(\text{GRAPH } \gamma\ \{\alpha\beta\delta\})\}$,

$$\Phi(\mu) = \begin{cases} \theta & \textit{if } \mathrm{r}(\mu, (\gamma, \alpha, \beta, \delta)) = (g, s, p, o) \wedge \\ & \quad (g, s, p, o, \theta) \in K \wedge f_1(g, s, p, o) = \theta, \\ \bot & \textit{else} \end{cases}$$

where $\text{vars}(P)$ *denotes the variables contained in a pattern P and* $\mathrm{r}(\mu, (\gamma, \alpha, \beta, \delta))$ *is the quadruple obtained by replacing the variables in* $(\gamma, \alpha, \beta, \delta)$ *according to* μ.

An example for this definition is given by processing example 5.1, i.e. query and corresponding dataset, delivering the result as indicated in example 5.2.

Basic quadruple pattern matching is not directly applicable, if an expression "GRAPH γ" appears outside a complex triple pattern. In such a case, we first need to distribute the expression "GRAPH γ" appropriately to atomic triple patterns in order to prescribe atomic SPARQL expressions accessible by basic quadruple pattern matching. Because named graphs cannot be nested, this distribution is always possible and unambiguous. In the following we use the function quads(P) to denote the query resulting from this transformation. In example 5.3 this transformation is demonstrated on a conjunction of two triple patterns.

Example 5.3.

P_1 =

```
GRAPH ?src {
    { ?x researchTopic ?y .}
    { ?x affiliatedWith ?z .}
}
```

$quads(P_1)$ =

```
GRAPH ?src { ?x researchTopic ?y .}
GRAPH ?src { ?x affiliatedWith ?z .}
```

Now we define the evaluation of complex graph patterns by operations on sets of variable assignments similar to [11, 13].

Definition 5.4 (Complex graph pattern matching). *Let* P_1, P_2 *be complex graph patterns. The evaluation of graph patterns over K, denoted by* $[\![\cdot]\!]_K$*, is defined recursively:*

1. $[\![\text{GRAPH } \gamma\ \{\alpha\beta\delta\}]\!]_K$ *is given by definition 5.3,*
2. $[\![\text{GRAPH g P}_1]\!]_K = [\![quads(P_1)]\!]_K$,
3. (a) $[\![P_1 \text{ AND } P_2]\!]_K = [\![P_1]\!]_K \bowtie [\![P_2]\!]_K$,
   (b) $[\![P_1 \text{ OPT } P_2]\!]_K = [\![P_1]\!]_K ⟕ [\![P_2]\!]_K$,
   (c) $[\![P_1 \text{ UNION } P_2]\!]_K = [\![P_1]\!]_K \cup [\![P_2]\!]_K$,
4. $[\![P_1 \text{ FILTER } C]\!]_K = \sigma_c([\![P_1]\!]_K)$,

The definition uses the operation AND. In standard SPARQL the operation AND is denoted by the absence of an operator. Like [11, 13] we still use the explicit term AND in order to facilitate referencing to this operator.

The recursion in the SPARQL query evaluation defined here is indeed identical to [11, 13]. Only the basic pattern matching has been changed slightly. Basic pattern matching now considers quadruples and it annotates variable assignments from basic matches with atomic statements from $\mathfrak{S}$ and variable assignments from complex matches with Boolean formulae $F \in \mathfrak{F}$ over S.

As an example, consider the query from example 5.4 evaluated on the knowledge base from example 5.5.

Example 5.4.

```
SELECT ?h1 ?h2 ?x ?y
FROM NAMED G1,G2
WHERE {
    { GRAPH ?h1 {?x affiliatedWith ?y .} AND
      GRAPH ?h2 {?x researchTopic "SemanticWeb" .} }
    FILTER  {?x="JamesHendler"}
}
```

Example 5.5.

```
K = {
(G1, JamesHendler,researchTopic,SemanticWeb, θ1),
(G1, JamesHendler,affiliatedWith,RensselaerPI, θ2),
(G2, JamesHendler,researchTopic,KnowledgeDiscovery, θ3),
(G2, JamesHendler,affiliatedWith,UnivMaryland, θ4),
(G2, RudiStuder,researchTopic,SemanticWeb θ5),
(G2, RudiStuder,affiliatedWith,UnivKarlsruhe θ6) }
```

Let P be the graph pattern contained in the WHERE clause of the query. Then the evaluation of P is defined by an algebraic expression:

$$\begin{aligned} [\![P]\!]_K &= [\![\{P_1 \text{ AND } P_2\} \text{ FILTER } \{?Z = \text{"JamesHendler"}\}]\!]_K \\ &= \sigma_{?x=\text{"JamesHendler"}}([\![P_1 \text{ AND } P_2]\!]_K) \\ &= \sigma_{?x=\text{"JamesHendler"}}([\![P_1]\!]_K \bowtie [\![P_2]\!]_K) \\ &= \sigma_{?x=\text{"JamesHendler"}}(\Phi_1 \bowtie \Phi_2) \end{aligned}$$

where Φ_1 and Φ_2 are relations representing variable assignments and their annotations. In this example and in the preceding definition we have used algebraic operations on sets of annotated bindings. However, we have not yet explained how these operations are used to construct formulas representing the how-provenance. The following definition will specify how complex formulae from $\mathfrak{F}$, which serve as annotations for results of matching complex graph pattern, will be derived.

Definition 5.5 (Algebra of Annotated Relations). *Let Φ_1, Φ_2 be sets of annotated variable assignments. We define $\bowtie, \cup, \setminus$ σ and $=\!\!\bowtie$ via operations on the annotations of the assignments as following:*

- $(\Phi_1 \bowtie \Phi_2)(\mu) = \Phi_1(\mu_1) \wedge \Phi_2(\mu_2)$, *where μ_1 and μ_2 are compatible and $\mu = \mu_1 \cup \mu_2$,*
- $(\Phi_1 \cup \Phi_2)(\mu) = \Phi_1(\mu) \vee \Phi_2(\mu)$,
- $(\Phi_1 \setminus \Phi_2)(\mu) = \Phi_1(\mu) \wedge \neg\left(\bigvee_{\mu_i, \Phi_2(\mu_i) \neq \bot} \Phi_2(\mu_i)\right)$, *where μ_i is compatible to μ,*
- $(\sigma_c(\Phi))(\mu) = \Phi(\mu) \wedge f_c(\mu)$, *where $f_c(\mu)$ denotes a function mapping μ to either $\top$ or $\bot$ according the condition c.*
- $(\Phi_1 =\!\!\bowtie \Phi_2)(\mu) = (\Phi_1 \bowtie \Phi_2)(\mu) \vee (\Phi_1 \setminus \Phi_2)(\mu)$.

Let us now continue the evaluation of the query specified in Example 5.4. In order to evaluate the expression $\sigma_{?x=\text{"JamesHendler"}}(\Phi_1 \bowtie \Phi_2)$ we need to determine Φ_1 and Φ_2 using definition 5.3. The intermediate result is shown in example 5.6. To evaluate the conjunction of two quadruple patterns the operation $\bowtie$ is applied, the result is shown in example 5.7. The annotation $\theta_1 \wedge \theta_2$ of the first row represents that this assignment has been derived from the conjunction of the two literal statements θ_1 and θ_2 (see example 5.5). Application of the σ-operation to the intermediate results gives the annotated relation shown in example 5.8.

Example 5.6.

$\Phi_1 =$

?h1	?x	?y	A_1
G1	JamesHendler	RensselaerPI	θ_2
G2	JamesHendler	UnivMaryland	θ_4
G2	RudiStuder	UnivKarlsruhe	θ_6

$\Phi_2 =$

?h2	?y	A_2
G2	JamesHendler	θ_1
G2	RudiStuder	θ_5

Example 5.7.

$\Phi_1 \bowtie \Phi_2 =$

?h1	?h2	?x	?y	A_3
G1	G1	JamesHendler	RensselaerPI	$\theta_1 \wedge \theta_2$
G1	G2	JamesHendler	UnivMaryland	$\theta_1 \wedge \theta_4$
G2	G2	RudiStuder	UnivKarlsruhe	$\theta_5 \wedge \theta_6$

Example 5.8.

$\sigma_{?x=\text{"JamesHendler"}}(\Phi_1 \bowtie \Phi_2) =$

?h1	?h2	?x	?y	A_4
G1	G1	JamesHendler	RensselaerPI	$(\theta_1 \wedge \theta_2) \wedge \top$
G1	G2	JamesHendler	UnivMaryland	$(\theta_1 \wedge \theta_4) \wedge \top$

The annotations $\Phi(\mu)$ can now be used to assign truth values for μ. I_s (see definition 4.4) assigns truth values to all atomic statements $s_i \in K \subseteq \mathfrak{S}$. We extend the interpretation I_s to capture all the Boolean formulae over statements $\mathfrak{S}$.

Definition 5.6 (Standard Interpretation of Formulae). *Let $F, F_1, F_2 \in \mathfrak{F}$ be Boolean formulae over $\mathfrak{S}$, let $F_a \in \mathfrak{S}$ be an atomic formula. We define the standard interpretation of formulae I_s^f as follows:*

- $I_s^f(F_a) := I_s(F_a)$;
- $I_s^f(\neg F) := \bot$ *if* $I_s^f(F) = \top$; $I_s^f(\neg F) := \top$ *if* $I_s^f(F) = \bot$;
- $I_s^f(F_1 \wedge F_2)$ *is* $\top$ *if* $I_s^f(F_1) = I_s^f(F_2) = \top$, *otherwise* $\bot$
- $I_s^f(F_1 \vee F_2)$ *is* $\top$ *if* $I_s^f(F_1) = \top$ *or* $I_s^f(F_2) = \top$, *otherwise* $\bot$.

For instance, I_s^f returns $\top$ for the assignment shown in the first row of $\Phi_1 \bowtie \Phi_2$ from example 5.7, because the statements θ_1 and θ_2 are in the knowledge base.

Analogously to I_s^f, we can extend a Π-interpretation I_π over RDF$^+$ statements to a Π-interpretation I_π^f over formulae. Remember that RDF$^+$ allows for only one ω per $\theta \in \Theta$ and $\pi \in \Pi$. In order to make use of the how-provenance represented by the annotations we require that for each meta-knowledge property π an algebra $(\Omega_\pi, \wedge, \vee, \neg, \top_\pi, \bot_\pi)$ with three operations $\wedge, \vee, \neg$ and two special elements $\top_\pi, \bot_\pi \in \Omega_\pi$ is defined. The definition of the algebras can be supplied by a modeler according to the intended semantics of the different meta-knowledge properties.

Definition 5.7 (Π-Interpretation of Formulae). *Let $F, F_1, F_2 \in \mathfrak{F}$ be Boolean formulae over $\mathfrak{S}$, let $F_a \in \mathfrak{S}$ be an atomic formula. We define the interpretation I_π^f as follows:*

- $I_\pi^f(F_a) := I_\pi(F_a)$;
- $I_\pi^f(\neg F)$ *is* $\neg I_\pi^f(F)$;
- $I_\pi^f(F_1 \wedge F_2)$ *is* $I_\pi^f(F_1) \wedge I_\pi^f(F_2)$;
- $I_\pi^f(F_1 \vee F_2)$ *is* $I_\pi^f(F_1) \vee I_\pi^f(F_2)$;

For illustration we consider in example 5.9 the definition of fuzzy logic operations to calculate a possibility measure on variable assignments, operations defined on timestamps which calculate the time of the last modification, and set operations defined for extraction sources that construct the combined provenance.

Example 5.9.

$I_{fuzzy}(x_1 \wedge x_2) = \min(I_{fuzzy}(x_1), I_{fuzzy}(x_2))$
$I_{fuzzy}(x_1 \vee x_2) = \max(I_{fuzzy}(x_1), I_{fuzzy}(x_2))$
$I_{fuzzy}(\neg x_1) = 1 - I_{fuzzy}(x_1)$

$I_{time}(x_1 \wedge x_2) = \min(I_{time}(x_1), I_{time}(x_2))$
$I_{time}(x_1 \vee x_2) = \max(I_{time}(x_1), I_{time}(x_2))$
$I_{time}(\neg x_1) = 0$

$I_{source}(x_1 \wedge x_2) = I_{source}(x_1) \cup I_{time}(x_2)$
$I_{source}(x_1 \vee x_2) = I_{source}(x_1) \cup I_{time}(x_2)$
$I_{source}(\neg x_1) = \{\}$

Query forms. For SELECT queries the evaluation on an RDF$^+$dataset is almost the same as in standard SPARQL. The only difference is that the evaluations $I^f_\pi(\Phi(\mu))$ are returned in addition to the bindings themselves. The evaluation will consider all named graphs specified in the WITH META clause, retrieve all meta knowledge statements from these graphs, and calculate values for all found knowledge properties. If there is no meta knowledge statement (θ, π, ω) for a particular RDF$^+$ literal statement $f_5(\theta)$ and a meta-knowledge property π we will use $\bot_\pi$ as default value.

Analogously to standard evaluation, the evaluation of a CONSTRUCT query on an RDF$^+$ dataset results in a single RDF$^+$ graph using a graph template. This is in line with the fact that the graph template consists of a conjunction of triple patterns and, thus, quadruple patterns cannot be stated.[3]

For each binding μ_i satisfying the WHERE-conditions and for each triple pattern t_j in the CONSTRUCT-template add the quintuple $(\hat{g}, s_{i,j}, p_{i,j}, o_{i,j}, \theta_{i,j})$ to the result set where $(s_{i,j}, p_{i,j}, o_{i,j})$ is the triple obtained by replacing the variables in t_j according to μ_i and $\theta_{i,j}$ is the statement identifier $f_1(\hat{g}, s_{i,j}, p_{i,j}, o_{i,j})$.

Notice that the same quintuple might have been produced more than once, because different derivation trees may result in the same binding of variables. Then, there may also be different annotations $\Phi(\mu_i)$ of the different bindings μ_i which have been used to create the identical $(\hat{g}, s_{i,j}, p_{i,j}, o_{i,j}, \theta_{i,j})$.

In order to account for meta-statements for new quintuples, each new quintuple inherits the meta-knowledge properties π associated with the binding which has been used to create that quintuple. For all new quintuples $f_5(\theta_{i,j})$ and all π we add statements $(\theta_{i,j}, \pi, \omega_{i,j})$ to M where $\omega_{i,j} = I^f_\pi(\Phi(\mu_i))$.

If the same quintuple has been obtained via different bindings with different annotations, $I^f_\pi(\Phi(\mu_i))$ will return different values for the involved bindings μ_i. As a consequence there are multiple $(\theta_{i,j}, \pi, \omega_{i,j})$ refering to the same statement identifier $\theta_{i,j}$. We aggregate the values $\omega_{i,j}$ for each meta-knowledge property into a single new meta-knowledge statement $(\theta_{i,j}, \pi, I^f_\pi(\bigvee \omega_{i,j}))$ and discard the original meta-knowledge statements (i.e. I^f_π is also the grouping function used in Definition 4.8).

Now we extend our previous examples to the calculation of meta-knowledge. Meta-knowledge for some of the RDF$^+$ statements presented in example 5.5 is specified in example 5.10.

Example 5.10.

```
$M$ = {
($\theta_1$, mk:fuzzy, 0.9),
($\theta_1$, mk:time, "5/5/2007"),
($\theta_2$, mk:fuzzy, 0.9),
($\theta_2$, mk:time, "5/5/2007"),
($\theta_4$, mk:fuzzy, 0.6),
($\theta_4$, mk:time, "6/6/2001")}
```

We assume that the RDF$^+$ meta-knowledge statements of example 5.10 are returned in a new RDF graph as described in section 4.3.

As an example for an CONSTRUCT statement consider example 5.11.

Example 5.11.

```
CONSTRUCT {?x worksAt ?y}
WITH META G3, G4
FROM NAMED G1, G2
WHERE {
    { GRAPH ?h1 {?x researchTopic "SemanticWeb" .} AND
      GRAPH ?h2 {?x affiliatedWith ?y .} }
    FILTER  {?x="JamesHendler"}
}
```

The preliminary result of query evaluation is shown in example 5.12. It contains variable assignments, the how-provenance of their construction, and corresponding meta knowledge properties that were constructed according to the standard interpretation. Using this information, the query processor returns the RDF$^+$ literal and meta-knowledge statements listed in example 5.13.

Example 5.12.

?h1	?h2	?x	?y	A_4	*fuzzy*	*time*
G1	G1	JamesHendler	RensselaerPI	$(\theta_1 \wedge \theta_2) \wedge \top$	0.9	5/5/2007
G1	G2	JamesHendler	UnivMaryland	$(\theta_1 \wedge \theta_4) \wedge \top$	0.6	6/6/2001

Example 5.13.

```
$K_{res}$ = {
($G_{new}$, JamesHendler, worksAt , RensselaerPI, $\theta_{new1}$)}
($G_{new}$, JamesHendler, worksAt , UnivMaryland, $\theta_{new2}$)}

$M_{res}$ = {
($\theta_{new1}$, mk:fuzzy, 0.9),
($\theta_{new1}$, mk:time, "5/5/2007"),
($\theta_{new2}$, mk:fuzzy, 0.6),
($\theta_{new2}$, mk:time, "6/6/2001") }
```

6. A Run through the Example

This section summarizes the discussed steps of meta knowledge representation and utilization for the sample scenario that was introduced in section 2.

6.1 Tasks for the administrator

In order to represent and utilize meta knowledge, the system administrator has to meet some design choices. In particular, the application-specific knowledge properties must be defined. In our sample scenario, we consider three meta knowledge properties: source, certainty, and timestamp. In the next step, the administrator defines the intended semantics of these properties in order to facilitate query processing with complex expressions and pattern combinations. Using the notion from Section 5.1, we assume that corresponding definitions for meta-knowledge properties are defined according to previously discussed example 5.9.

Finally, data and available associated meta-knowledge are represented in RDF using named graphs [1, 4], and imported into our RDF$^+$-based repository.

6.2 Processing performed by the System

We assume that the administrator manages the small sample knowledge base introduced in section 2. As discussed in section 4, the internal representation of the knowledge in the RDF$^+$-based repository is as follows (assuming suitable namespace abbreviations).

[3] Standard SPARQL does not allow for giving this graph a name. In order to associate meta knowledge, multiple named graphs as outputs are convenient. In order to remain standard compliant, the SPARQL engine may however also return data and meta knowledge in two different batches distinguished by some implementation-specific mechanism.

Example 6.1.

```
{
(G1, JamesHendler researchTopic SemanticWeb, $\theta_1$),
(G1, JamesHendler affiliatedWith RensselaerPI, $\theta_2$)
(G2 JamesHendler researchTopic KnowledgeDiscovery $\theta_2$)
...}
```

The associated meta knowledge is internally represented in the following form:

Example 6.2.

```
{
($\theta_1$, mk:source, <www.rpi.edu/report.doc>),
($\theta_1$, mk:certainty, 0.9)
($\theta_1$, mk:timestamp, "5/5/2007") ...  }
```

Following our sample scenario, we consider the query that retrieves names of Semantic Web experts together with their affiliations.

Example 6.3.

```
CONSTRUCT {?x worksAt ?z}
WITH META Gx
FROM NAMED G1,G2
WHERE {
    { GRAPH ?h1 {?x researchTopic y .} AND
      GRAPH ?h2 {?x affiliatedWith ?z .} }
    FILTER  {?y = "SemanticWeb"}
}
```

Internally, the query processor evaluates this query using graph patterns discussed in 5.1. As a result, the following preliminary variable assignment is produced:

Example 6.4.

?h1	?h2	?x	?y	A_3
G1	G1	JamesHendler	RensselaerPI	$\theta_1 \wedge \theta_2$
G1	G2	JamesHendler	UnivMaryland	$\theta_1 \wedge \theta_4$
G2	G2	RudiStuder	UnivKarlsruhe	$\theta_5 \wedge \theta_6$

It contains possible variable assignments, and the how-provenance (A_3) that explains *how* these source statements have been used.

By combining this information with definitions for meta knowledge properties and available meta knowledge statements, the query processor constructs the result as follows:

Example 6.5.

```
{
($G_{new}$, JamesHendler, worksAt , RensselaerPI, $\theta_{new1}$)
($G_{new}$, JamesHendler, worksAt , UnivMaryland, $\theta_{new2}$)
($G_{new}$, RudiStuder, worksAt , UnivKarlsruhe, $\theta_{new3}$)}

{
($\theta_{new1}$, mk:fuzzy, 0.9),
($\theta_{new1}$, mk:time, "5/5/2007"),
($\theta_{new1}$, mk:source, <www.rpi.edu/report.doc>),
($\theta_{new2}$, mk:time, "6/6/2001") ...  }
```

This result is then serialized in RDF.

6.3 Benefits for the user/developer

The user or application developer can access the knowledge stored in the RDF$^+$-based repository in different ways. On one hand, the repository does not change the existing SPARQL semantics and thus fully supports common SPARQL queries. This is an important advantage for compatibility with existing applications and interfaces. On the other hand, the repository supports the advanced SPARQL syntax with metaknowledge support (section 5.1). To this reason, the user obtains additional access to the valuable meta knowledge that can be used for relevance ranking, conflict resolution, or other similar applications in connection with retrieved knowledge.

In our application scenario, the user may realize that the query answer is potentially contradictive (James Hendler is affiliated with Rensselaer PI and University of Maryland). By inspecting the associated meta knowledge, he would realize that the second fact was generated by mistake. In fact, it is based on outdated information (knowledge from the document "survey.pdf" with timestamp "6/6/2001") that was wrongly combined with knowledge from a more recent source (namely document "report.doc" with timestamp "5/5/2007"). It turns out that the affiliation of James Hendler has actually changed from U Maryland to Rensselaer PI, and the erroneous tuple can be safely excluded from further processing.

7. Complexity

In this section we analyze how the construction of the annotations influences the complexity of the decision problem related to SPARQL. The decision problem associated with the evaluation of a SPARQL query can be stated as following [11]: *Given an RDF dataset D, a graph pattern P and a mapping μ, determine whether μ is in the result of P applied to D.* For this decision problem, which we denote by Eval, an analysis of the complexity is presented in [11, 12]. In the context of RDF$^+$ datasets and annotated variable assignments we have a slightly different decision problem: *Given an RDF$^+$ dataset D^+, an RDF$^+$ graph pattern P^+, a variable assignment μ and an annotation α determine whether α is the correct annotation of μ.* We denote this problem by Eval$^+$. An annotation is correct iff it is identical to the formula obtained by evaluating P^+ as defined in section 5.

Since μ must have an annotation $\alpha \neq \bot$ iff μ is in the result the second decision problem includes the first one. The key difference is to construct *different* annotations for mappings which are in the result. With the following two theorems we show that for pattern which do not use the OPTIONAL operator Eval$^+$ has the same complexity as Eval. For both theorems the RDF counterparts have been established by [11, 12].

Theorem 7.1. Eval$^+$ *can be solved in time $O(|P| \cdot |D|)$ for graph pattern expressions constructed by using only AND and FILTER operators.*

Theorem 7.2. Eval$^+$ *is NP-complete for graph pattern expressions constructed by using only AND, FILTER and UNION operators.*

The theorems indicate that our treatment of meta knowledge does not add to the computational complexity of SPARQL. A proof for each of the theorems can be found at
`http://isweb.uni-koblenz.de/Research/MetaKnowledge`.

8. Implementation

The framework described in this paper has been implemented and is available as an initial prototype. The prototype is available as an open source implementation at `http://isweb.uni-koblenz.de/Research/MetaKnowledge` together with example queries using artificial data from the LeHigh benchmark[4].

9. Related Work

The importance of better understanding the ways by which the result came about is fundamental to many Semantic Web applications and scenarios. The specification of the Semantic Web proof layer was discussed in [10, 14, 9]. Our approach is focused on a different language model (RDF) and provides fine-grained meta knowledge management for retrieval queries with SPARQL that is not directly comparable with proof traces for OWL reasoning.

In the area of database systems, meta knowledge is often represented using an extension of the relational data model, coined *annotated relations*. Its purpose is primarily the description of data origins (provenance) and the process by which it arrived as a query answer [5, 2, 3, 6]. Basically, our methodology follows the same idea. However, our approach is specially designed for RDF graph models and not directly comparable to metadata models for relational database systems. The same holds for the query language (SPARQL instead of SQL) and its semantics. An important difference to isolated database solutions is the serialization ability of RDF and thus seamless exchanging and utilization of meta knowledge from our framework across the Semantic Web.

10. Conclusion and Future Work

In this paper, we presented an original, generic, formalized and implemented approach for managing many dimensions of meta knowledge, like source, authorship, certainty, and others, for RDF repositories. Our method re-uses existing RDF modeling possibilities in order to represent meta knowledge. Then, it extends SPARQL query processing in such a way that given a SPARQL query for data, one may request meta knowledge without modifying the query proper. We achieve highly flexible and automatically coordinated querying for data and meta knowledge, while completely separating the two areas of concern. Our approach remains compatible to existing standards and query languages and can be easily integrated with existing applications and interfaces.

In the future, we will investigate the meta knowledge support for OWL-based knowledge bases with advanced reasoning capabilities. Due to the substantially higher complexity of inferencing and retrieval algorithms (e.g. reasoning in OWL-DL vs. RDF querying with SPARQL) and the distributed nature of knowledge sources in the Semantic Web, the notion of meta knowledge will require further, non-trivial justification. Another interesting research issue is the support for *nested* meta knowledge (i.e. construction of meta knowledge for the result with respect to additional information *about* meta knowledge of its origins).

Our long-term objective is the generic, efficient and effective infrastructure for meta knowledge management as an integral part of the proof layer of the Semantic Web.

[4] available at `http://swat.cse.lehigh.edu/projects/lubm/`

11. REFERENCES

[1] Christian Bizer and Jeremy J. Carroll. Modelling context using named graphs. In *W3C Semantic Web Interest Group Meeting*, Cannes, France, 2004.

[2] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Data Provenance: Some Basic Issues. *20th Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS), New Delhi, India*, pages 87–93, 2000.

[3] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and Where: A Characterization of Data Provenance. *8th International Conference on Database Theory (ICDT), London, UK*, pages 316–330, 2001.

[4] Jeremy J. Carroll and Patrick Stickler. Trix: RDF triples in XML. In *Proceedings of the Extreme Markup Languages 2004*, Montreal, Canada, 2004.

[5] Y. Cui and J. Widom. Practical Lineage Tracing in Data Warehouses. *16th International Conference on Data Engineering (ICDE), San Diego, USA*, 2000.

[6] Li Ding, Pranam Kolari, Tim Finin, Anupam Joshi, Yun Peng, and Yelena Yesha. On homeland security and the Semantic Web: A provenance and trust aware inference framework. In *Proceedings of the AAAI Spring Symposium on AI Technologies for Homeland Security*, 2005.

[7] Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.

[8] Turtle Terse RDF Triple Language. http://www.dajobe.org/2004/01/turtle/.

[9] D. McGuinness and P. Pinheiro da Silva. Explaining Answers from the Semantic Web: the Inference Web Approach. *J. Web Sem.*, 1(4):397–413, 2004.

[10] W. Murdock, D. McGuinness, P. Pinheiro da Silva, C. Welty, and D. Ferrucci. Explaining Conclusions from Diverse Knowledge Sources. *International Semantic Web Conference (ISWC), Athens, USA*, pages 861–872, 2006.

[11] Jorge Perez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of SPARQL. In *ISWC*, volume 4273 of *LNCS*, pages 30–43, 2006.

[12] Jorge Perez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of SPARQL. arXiv:cs/0605124v1 [cs.DB], May 2006.

[13] Jorge Perez, Marcelo Arenas, and Claudio Gutierrez. Semantics of SPARQL. Technical Report TR/DCC-2006-17, Universidad de Chile, October 2006.

[14] P. Pinheiro da Silva, D. McGuinness, and R. Fikes. A Proof Markup Language for Semantic Web services. *Inf. Syst.*, 31(4-5):381–395, 2006.

[15] Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. Working draft, W3C, March 2007. http://www.w3.org/TR/rdf-sparql-query/.

[16] M. Schraefel, N. Shadbolt, N. Gibbins, S. Harris, and H. Glaser. CS AKTive Space: Representing Computer Science in the Semantic Web. *13th International Conference on World Wide Web (WWW), New York, USA*, pages 384–392, 2004.

[17] Bernhard Schueler, Sergej Sizov, and Steffen Staab. Management of Meta Knowledge for RDF Repositories. In *Int. Conf. on Semantic Computing (ICSC)*, pages 543–550, Irvine, CA, September 2007.