Semantic Web 0 (0) 1–21
IOS Press

Reasoning with Data Flows and Policy Propagation Rules

Editor(s): Mathieu d'Aquin, Open University, UK; Sabrina Kirrane, Wirtschaftsuniversität Wien, Austria; Serena Villata, Université Nice Sophia Antipolis, France
Solicited review(s): Simon Steyskal, Wirtschaftuniversität Wien, Technische Universität Wien & Siemens, Austria; David Corsar, University of Aberdeen, UK; Ernesto Damiani, Università degli Studi di Milano, Italy

Enrico Daga [a,*], Aldo Gangemi [b], Enrico Motta [a]
[a] *Knowledge Media Institute, The Open University*
Walton Hall, Milton Keynes, United Kingdom
E-mail: {enrico.daga,enrico.motta}@open.ac.uk
[b] *Université Paris13, France and Institute of Cognitive Sciences and Technologies - CNR, Italy*
E-mail: aldo.gangemi@{univ-paris13.fr,cnr.it}

Abstract.

Data-oriented systems and applications are at the centre of current developments of the World Wide Web. In these scenarios, assessing what policies propagate from the licenses of data sources to the output of a given data-intensive system is an important problem. Both policies and data flows can be described with Semantic Web languages. Although it is possible to define Policy Propagation Rules (PPR) by associating policies to data flow steps, this activity results in a huge number of rules to be stored and managed. In a recent paper, we introduced strategies for reducing the size of a PPR knowledge base by using an ontology of the possible relations between data objects, the Datanode ontology, and applying the (A)AAAA methodology, a knowledge engineering approach that exploits Formal Concept Analysis (FCA). In this article, we investigate whether this reasoning is feasible and how it can be performed. For this purpose, we study the impact of compressing a rule base associated with an inference mechanism on the performance of the reasoning process. Moreover, we report on an extension of the (A)AAAA methodology that includes a *coherency check* algorithm, that makes this reasoning possible. We show how this compression, in addition to being beneficial to the management of the knowledge base, also has a positive impact on the performance and resource requirements of the reasoning process for policy propagation.

Keywords: Data Hub, Data Flows, Policies, Rules, Formal Concept Analysis, RDF Licenses

1. Introduction

Data-oriented systems and applications are at the centre of current developments of the World Wide Web (WWW). Emerging enterprises focus their business model on providing value from data collection, integration, processing, and redistribution. These kind of systems are not new, as the Web has enabled for a long time tools such as news aggregators, which collect articles from various providers, and republish them as collections of short readings, often focusing on specific topics (politics, sport, etc.)[1]. Nowadays, the extraction, publication, and reuse of data on the Web is an established practice, and a large number of APIs provide access to JSON documents, data tables, or Linked Data

*Corresponding author, e-mail: enrico.daga@open.ac.uk

[1] Wikipedia: `https://en.wikipedia.org/wiki/News_aggregator`.

for a variety of use cases, spanning from content and media linkage [22] to science and education [20].

The key aspect on which we are focusing here is the publication of Licenses and Terms and Condition documents associated with those APIs and data artifacts, that declare the associated rights and policies that should guide their use. Data Hubs collect a large variety of data sources and process them in order to implement the workflow that *connects data in their original sources to applications that might want to exploit these data* [12]. These systems create new challenges in terms of the volume of data to be stored and require novel processing techniques (for example stream-based analysis [26]), but more importantly they demand for more sophisticated approaches to data governance [10]. In the Web of (open) data, developers can access a large variety of information, and often publish the results of their processing. Hence, they need to be aware of any usage constraints attached to data sources they want to exploit, and they need support in publishing the appropriate policies alongside the data they distribute.

In this complex scenario, assessing what policies propagate from the licenses associated with the data sources to the output of a given data-intensive process is an important problem. Both policies and data flows can be described within the Semantic Web, relying on standards like the W3C PROV model[2] to describe process executions in a provenance chain and the Open Digital Rights Language[3], which actual purpose is to formalize and validate policies. Particularly, it is possible to specify Policy Propagation Rules (PPR) [7] by associating policies with data flow steps, although this activity results in a large number of rules to be stored and managed. In [7], we studied how a PPR knowledge base can be compressed by using an ontology of the possible relations between data objects, the Datanode ontology[4], and by applying the (A)AAAA methodology, a knowledge engineering approach that exploits Formal Concept Analysis (FCA).

In this article we illustrate how reasoning on policy propagation can be practically performed. Building upon [7], we report on an extension of the (A)AAAA methodology that includes a *coherency check* between the hierarchy of the FCA lattice and the Datanode ontology. This extension was necessary in order to exploit the compressed rule base during reasoning and avoid incorrect results. While the compression of the rule base reduces the number of rules to be managed, it requires the reasoner to compute more inferences. Therefore, we study the impact of rule base compression on the performance of the reasoning process. In other words, this article focuses on two contributions that relate to the aspect of reasoning with (compressed) PPRs, which was missing in [7]: 1- the extension of the (A)AAAA methodology by adding an additional *coherency check* step to the Assessment phase, and 2- the evaluation of the effect of compression on reasoning performance.

The article is structured as follows. Section 2 reviews the relevant literature. Section 3 presents an exemplary use case, and introduces the elements for reasoning on policy propagation, going through the description of the data flow, the representation of policies, and the concept of Policy Propagation Rule (PPR). Section 4 provides a summary of the (A)AAAA methodology, integrated with a novel *Assessment* phase that includes a *coherency check* algorithm that allows effective reasoning with a compressed rule base. We also evaluate the impact of this evolved methodology on the compression factor of the knowledge base of PPRs. In Section 5, we report on experimental results about the impact of a compressed rule base on reasoning. For this purpose, we compare the performance of reasoning with an uncompressed rule base against reasoning with a compressed one. We perform this comparison using two different reasoners, the first computing the inferences at query time, the second materializing them at load time. Finally, we discuss our observations before closing the article with some conclusions and perspectives on future work.

2. Related Work

In recent years, data repositories and registries have been growing, spanning from data cataloguing services (Datahub[5]), data collections (Wikidata[6], Europeana[7]), to platforms that manage the collection and redistribution of data (Socrata[8]). An emerging category of such

[2] W3C PROV, https://www.w3.org/TR/prov-overview/.

[3] ODRL W3C Community Group, https://www.w3.org/community/odrl/.

[4] Datanode, http://purl.org/datanode/ns/ and http://purl.org/datanode/docs.

[5] Datahub. https://datahub.io/

[6] Wikidata. https://www.wikidata.org

[7] Europeana. http://labs.europeana.eu/

[8] Socrata. https://www.socrata.com/

systems are City Data Hubs, which need to support developers not only in obtaining data, but also in assessing the policies associated with data resulting from complex pipelines [12,3,2]. It is therefore for these systems to implement technologies that allow policies associated to derived datasets to be computed. In this article we concentrate on the problem of reasoning with propagating policies.

Policies can be represented on the Web in a machine-readable format. The W3C ODRL Community Group[9] works on the development of a set of specifications to enable interoperability and transparent communication of policies associated with software, services, and data. The Open Digital Rights Language (ODRL)[10] is an emerging language to support the definition, exchange and validation of policies [18]. Although ODRL is also available as an ontology, it only defines the semantics of policies in terms of natural language descriptions. An extension of the ODRL semantics has been proposed in [31] by considering dependencies between actions, and discussing the impact of explicit and implicit dependencies on the evaluation of policy expressions. The idea of establishing dependencies between ODRL actions in order to enhance the evaluation of ODRL expressions is related to our work, where we abstract relations in data flows. The Datanode ontology [6] which we use here is however designed to express a wider range of relations between data artifacts, and not only the ones derivable from actions. For instance, partitive relations influence the attached policies but are not derived from any action on the data. Nevertheless, a PPR reasoner can surely benefit from a well-defined semantics of ODRL actions. Recently, the W3C Permissions & Obligations Expression Working Group[11] followed up on ODRL to develop an official W3C standard for defining permissions and obligations.

The RDF Licenses Dataset [28] is an attempt to establish a knowledge base of license descriptions based on RDF and the ontology provided by ODRL. It also uses other vocabularies aimed to extend the list of possible actions, for instance the Linked Data Rights[12] vocabulary.

[9] W3C ODRL Community Group https://www.w3.org/community/odrl/

[10] ODRL Vocabulary & Expression, https://www.w3.org/TR/2016/WD-vocab-odrl-20160721/.

[11] W3c Permissions & Obligations Working Group, https://www.w3.org/2016/poe/wiki/Main_Page.

[12] Linked Data Rights (LDR): http://purl.oclc.org/NET/ldr/ns#.

Process executions can be described in the Semantic Web using the Provenance Ontology (PROV-O) [24]. PROV-O describes workflow executions in terms of *agents*, *actions* and *assets* involved. The Datanode ontology has been designed to describe Semantic Web applications by means of the relations between the data involved in their processes [6]. The ontology is a taxonomy of possible relations that may occur between data objects, which might be part of a process execution, such as the ones described with PROV-O. It can therefore be used to further qualify the implications of the actions performed in such a process. Datanode can describe process implications in a data-oriented way, namely as network of data objects. While policies and process executions can be represented, in the present paper we aim at studying the process of reasoning upon the propagation of policies across a data flow.

Rule-based representation and reasoning over policies is required in order to enable secure data access and usage in distributed environments, particularly in the Semantic Web [25,13,4]. Defeasible logic is used to reason with deontic statements, for example to check compatibility of licenses or to validate constraints attached to components on multi-agent systems [29]. The problem of licenses' compatibility has been extensively studied in the literature [16,15] and tools that can perform such assessment do exist [23]. Our previous work introduces a form of policy reasoning, namely *policy propagation* [7]. A Policy Propagation Rule (PPR) is a Horn clause defined by associating a Datanode relation with an ODRL policy. Reasoning with Horn rules is an effective way of dealing with policies, particularly because Horn rules allow tractable defeasible reasoning [1]. While in this article we only focus on policy propagation, PPRs can in principle be integrated with rule-based reasoners for policy validation.

Formal Concept Analysis (FCA) [33] has the capability of classifying collections of objects depending on their *features*. We apply FCA in conjunction with the Datanode ontology to detect a common behaviour of relations in terms of policy propagation, with the purpose of compressing a PPR knowledge base. We refer the reader to [9] for a description of the Contento tool, that implements FCA as well as other functionalities for evolving concept lattices in Semantic Web ontologies, also part of the approach we present here.

The approach described in this paper clearly relates to principles and methods of knowledge engineering [32]. In [27], knowledge acquisition is considered

as an iterative process of model refinement. More recently, problem solving methods have been studied in relation to the task of understanding process executions [14]. These contributions form the background of the approach we are following in the present work. The problem of compressing propositional knowledge bases has been extensively studied in the past, focusing on the optimization of a Horn minimization process to improve the performance of rule execution [17,5]. Differently, we deal with compression as a mean to reduce the number of minimal rules to be managed (each PPR being already an atomic rule), by means of an additional knowledge base (the Datanode ontology).

It is worth noting that our problem is not one of policy enforcement, but of providing the right information about policies that might affect the terms of use of a given asset produced by a complex data flow. This problem is also different from the one of minimizing access control policies (example, the abstraction by subsumption proposed in [19]), as the abstraction required is on the propagation of the policy, not the policy itself. Reasoning on policy propagation does not require the policies to be validated per se. On the contrary, we claim that validating the policies of a data artifact, which is the result of some manipulation, should consider the policies inherited from the input data, according to the particular actions performed.

To our knowledge, the problem of propagation of usage policies in data flows has not been tackled before the contribution in [7]. In [10] we proposed an approach for integrating policy propagation in the data governance activity of Data Hubs, where policies and data flows are managed by Data Hub managers. However, in [7] as well as in the present work, we do not focus on the quality of the data flow representations, and assume a machine-readable description of the policies of the input asset, as well as the existence of an accurate data flow.

3. Reasoning on policy propagation

In this section, we describe our approach for reasoning on policy propagation, and we present a use case as an example.

3.1. Approach

We define the problem of policy propagation as identifying the set of policies associated with the output of a process, implied by the policies associated with the input data source. In order to perform reasoning on policy propagation, we need:

a) descriptions of policies attached to data sources;
b) a description of the data flow (the actions performed on the data), and
c) policy propagation rules (which actions do propagate a given policy).

Description of policies. We assume the policies of data sources are described as licenses or "terms and conditions" documents, and that they are expressed in RDF according to the ODRL ontology[13]. An ODRL `odrl:Policy` is an *entity to capture the statements of the policy*, specifying a set of `odrl:Rules`, each including a deontic aspect (`odrl:permission` or `odrl:prohibition`), which are defined for a set of `odrl:Actions` and a `odrl:target` `odrl:Asset`. Permissions, in turn, can comprise a `odrl:duty` (or more). For example, the RDF Licenses Dataset [28] is a source of such descriptions. In our work, we also developed ad-hoc RDF documents to satisfy this requirement, when necessary.

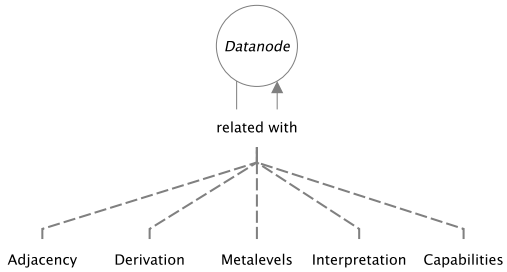


Fig. 1. Top hierarchy of the Datanode ontology.

Description of the data flow. Data flows are represented with the Datanode ontology [6]. The terms are defined under the `http://purl.org/datanode/ns/` namespace (we use the prefix `dn:` for readability). The ontology defines a unique type - `dn:Datanode` - and 115 relations, starting from a single top property: `dn:relatedWith`, having `dn:Datanode` as `rdfs:domain` and `rdfs:range`. An instance of `dn:Datanode` is any data object that can be the input or output of a process.

[13]ODRL 2.1: `https://www.w3.org/ns/odrl/2/ODRL21`.

The ontology groups the relations under five main dimensions[14], summarized in Figure 1:

Adjacency. `dn:adjacentTo` represents proximity between two datanodes in a data container. For example, proximity may result from being parts of the same dataset - `dn:disjointPartWith`, or being an annotation of the dataset - `dn:hasAnnotation`, or an attachment - `dn:attachedTo`.

Derivation. This branch specializes `dn:hasDerivation` in a number of different forms. Examples cover activities like mining - `dn:hasExtraction`, selection - `dn:isSelectionOf`, reasoning - `dn:hasInference`, remodelling - `dn:remodelledFrom`, or the activity of making snapshots of data or caches - `dn:hasSnapshot`, `dn:hasCache`, to mention a few.

Metalevels. This dimension covers the relations between a data object and its metadata. The property `dn:metadata` is used to designate a relation with information that applies to the datanode as a whole. This relation specializes as `dn:describes`, `dn:hasAnnotation` and `dn:hasStatistics`.

Interpretation. This is designed to capture the possibility that a datanode might contribute to inferences that can be made in another one. Two datanodes *might* be "understood" together, i.e. their content can be compared, or the interpretation (inferences) of one may affect the interpretation (inferences) of another. The more intuitive examples are `dn:consistentWith` and `dn:inconsistentWith`. However, this is also the area of the ontology that covers partitive relations: `dn:isPartOf` and the two specializations `dn:isPortionOf` and `dn:isSectionOf`. In Datanode, *portion* refers to a part of the population of a dataset (such as the rows of a spreadsheet), while *section* refers to a set of values for a certain dimension in a dataset (for example, a column of a spreadsheet).

Capabilities. Capability is intended as *the power or ability to generate an outcome*[15]. Capability is covered with two separate branches starting from `dn:overlappingCapabilityWith` and `dn:differentCapabilityFrom`, respectively. Two datanodes may have similar (or different) potential. For example, `dn:overlappingVocabularyWith` and `dn:overlappingPopulationWith` express the similarity between two data objects in terms of vocabulary or population of a dataset. Under this scope, we also positioned `dn:optimizedInto` (also a kind of derivation), to state the empowerment of an existing capability.

It is worth noting that Datanode relations often have multiple ancestors. For example, `dn:hasStatistics` is both a `dn:hasComputation` and a `dn:describedBy` kind of relation, which in turn are subsumed by `dn:hasDerivation` and `dn:metadata` respectively. Similarly, `dn:hasAnnotation` relates a datanode to some attached metadata, therefore it is subsumed by `dn:attachedTo` and `dn:metadata`. We refer to [6] for a discussion on the development of Datanode.

In this work, we use the representations of data flows extracted from the descriptions of several Semantic Web applications prepared in [6].

Policy Propagation Rules. A Policy Propagation Rule (PPR) establishes a binding between a Datanode relation r and a policy p. A PPR is a Horn clause of the following form:

$$has(X,p) \wedge propagates(p,r) \wedge relation(r,X,Y) \rightarrow has(Y,p)$$

where X and Y are data objects, p is a policy and r a Datanode relation between X and Y. When the policy p holds for a data object X, related to another data object Y by the relation r, then the policy p will also hold for the data object Y. For example a PPR could be used to represent the fact that downloading a file F distributed with an attribution requirement will result in a local copy D, which also needs to be used according to the attribution requirement. Therefore, the above abstract rule could be instantiated as follows:

[14] In this section we only summarize the basic features of the ontology, and we omit to specify inverse relations (for example `dn:isDerivationOf`), for clarity. The interested reader is referred to [6] and to the online documentation: `http://purl.org/datanode/docs`

[15] Definition from `http://en.wiktionary.org/wiki/capability`.

$$has(F, attribution) \wedge \\ propagates(attribution, isCopyOf) \wedge \\ relation(isCopyOf, F, D) \rightarrow has(D, attribution)$$

In fact, we can reduce a PPR to a more compact form, i.e. a binary association between a policy p and a relation r:

$$propagates(p, r)$$

as the other components of the rule can be automatically derived for any possible X and Y.

With these elements established, we can trace the policies propagated within the data flow connecting input and output.

3.2. Example use case

We described the components required to reason upon policy propagation in data flows. We now introduce a motivating scenario. The following are those name spaces that will be referred to in this example:

```
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
odrl: <http://www.w3.org/ns/odrl/2/>
cc: <http://creativecommons.org/ns#>
dn: <http://purl.com/datanode/ns/>
ppr: <http://purl.com/datanode/ppr/ns/>
ex: <http://purl.org/datanode/ex/>
```

We selected EventMedia [21] as an exemplary data-oriented system. EventMedia exploits real-time connections to enrich content describing events and associates it with media objects[16]. The application reuses data exposed by third parties, aggregating data about events and exposing them alongside multimedia objects retrieved on the Web. Aggregated data are internally represented using the LODE ontology [30]. In order to associate the right policies to these data, a description of the policies of the input data, a description of the data flow, and a knowledge base of PPRs are needed.

Table 1 lists the licenses or terms of use documents associated with the input data objects[17]. Listing 1 lists the set of policies associated to the content of the Flickr API, stated in the Flickr APIs Terms of Use[18].

Listing 1: Policies representation extracted from the Flickr APIs Terms of Use.

```
ex:FlickrTerms a odrl:Offer;
 rdfs:label "Flickr APIs Terms of Use";
 rdfs:seeAlso
  <https://www.flickr.com/services/api/tos/>;
 odrl:assigner <https://www.flickr.com> ;
 odrl:prohibition [
    odrl:target ex:Flickr ;
    odrl:action odrl:sell, odrl:grantUse,
            cc:CommercialUse ];
 odrl:permission [
    a odrl:Permission;
    odrl:target ex:Flickr ;
    odrl:action odrl:use;
    odrl:duty odrl:attribute ]
  .
```

Figure 2 illustrates the EventMedia data flow and Listing 2 the equivalent RDF description. Data are processed from event directories and enriched with additional information and media from sources like DBpedia[19], Flickr[20] or Foursquare[21]. In the figure, circles are data objects and arcs are Datanode relations. We will follow the path that connects the `ex:output` data object to two of the input data objects, namely `ex:Flickr` - that represents the Flickr API[22] (this path is highlighted in the figure), and Eventful[23] - a portal to search for upcoming events and related tickets. Apart from using the LODE ontology, `ex:output` is remodelled from an aggregation of various sources, named as `ex:collection`. The population (entities) of `ex:collection` includes `ex:events`, a `dn:combinationFrom ex:Eventful` with other sources (central path in the figure). Moreover, `ex:collection` includes descriptions of media from `ex:Flickr`, expressed by the path `dn:hasPortion / dn:isCopyOf / dn:isSelectionOf`. The data selected from `ex:Flickr` also refer to (some of) the entities aggregated in `ex:events`. This is expressed

[16] See `http://eventmedia.eurecom.fr/`.

[17] The *Upcoming* service is not available at the time of writing, however a snapshot of the documentation can be consulted from the Web Archive, reporting a non-commercial use clause: `https://web.archive.org/web/20130131064223/http://upcoming.yahoo.com/services/api/`. The application was firstly produced in 2014, when the EventMedia dataset description article was firstly submitted to the Semantic Web Journal. The description produced refers to the submitted version, which could be changed in the published version.

[18] Flickr API Terms of Use: `https://www.flickr.com/services/api/tos/`.

[19] DBpedia: `http://dbpedia.org`.

[20] Flickr: `http://www.flickr.com`.

[21] This description has been initially elaborated in [6].

[22] Flickr API: `https://www.flickr.com/services/api/`.

[23] Eventful: `http://eventful.com/`

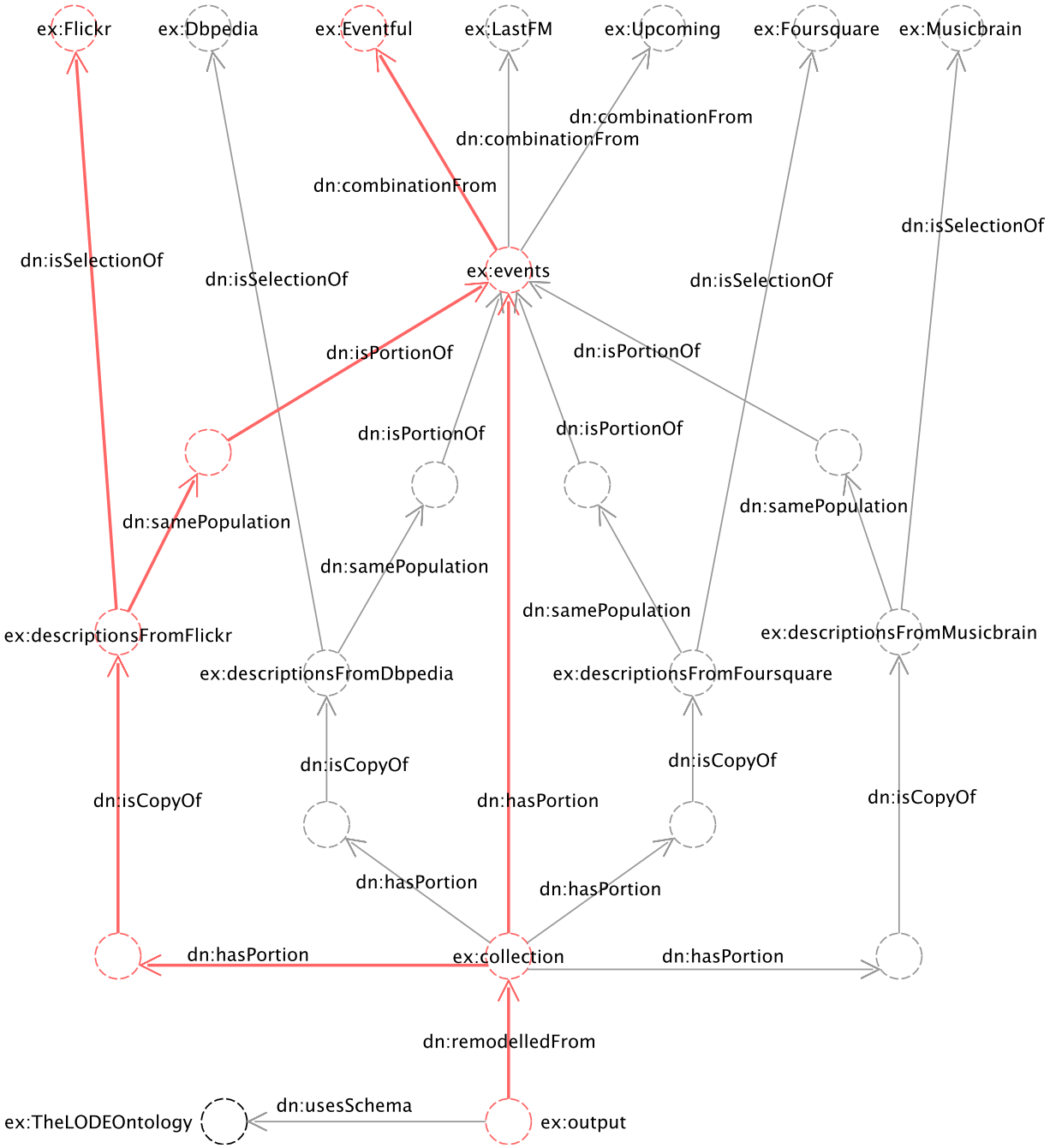


Fig. 2. The data flow of EventMedia. Input sources are the top nodes. The node at the bottom depicts the output data, which is a remodelling of the data collected from various sources according to a specific schema.

by the path `ex:descriptionsFromFlickr dn:samePopulation / dn:isPortionOf ex:events`. Therefore, the data flow is a backtrace of the abstract process of the EventMedia system, from the `ex:output` data object towards the input data sources.

Listing 2: The EventMedia data flow in RDF.

```
ex:events
    dn:combinationFrom ex:Eventful ,
      ex:LastFM , ex:Upcoming .

ex:collection dn:hasPortion
 [dn:isCopyOf ex:descriptionsFromFlickr],
 [dn:isCopyOf ex:descriptionsFromDbpedia],
 [dn:isCopyOf ex:descriptionsFromMusicbrain],
 [dn:isCopyOf ex:descriptionsFromFoursquare],
 ex:events .

ex:descriptionsFromDbpedia
    dn:isSelectionOf ex:Dbpedia ;
    dn:samePopulation
        [ dn:isPortionOf ex:events ] .

ex:descriptionsFromFlickr
    dn:isSelectionOf ex:Flickr ;
    dn:samePopulation
        [ dn:isPortionOf ex:events ] .

ex:descriptionsFromFoursquare
    dn:isSelectionOf ex:Foursquare ;
    dn:samePopulation
        [ dn:isPortionOf ex:events ] .

ex:descriptionsFromMusicbrain
    dn:isSelectionOf ex:Musicbrain ;
    dn:samePopulation
        [ dn:isPortionOf ex:events ] .

:output
    dn:isRemodelledFrom ex:collection ;
    dn:usesSchema ex:TheLODEOntology .
```

Table 1

Sources of Terms and conditions associated with the data sources of EventMedia.

Source	T&C
Flickr	Flickr APIs Terms of Use[24]
Dbpedia	Creative Commons CC-BY-SA 3.0
Eventful	Eventful API Terms of Use[25]
LastFM	LastFM Terms of Service[26]
Upcoming	Non Commercial Use Requirement
Musicbrain	Creative Commons CC0
Foursquare	Foursquare Developers Policies[27]

The data flow described so far can be leveraged by a reasoner in conjunction with the ODRL policies of the inputs, and the PPRs, to infer the policies associated with `ex:output`. Listing 3 shows the policies propagated from the inputs to the output of the EventMedia data flow, some of the deriving from the restrictions applied to Flickr data, shown previously in Listing 1

Listing 3: Example of policy associated with the output of EventMedia.

```
ex:outputPset a odrl:Set ;
  odrl:prohibition [
    odrl:target ex:output ;
    odrl:action odrl:modify ,
        cc:commercialUse , odrl:sell ];
  odrl:permission [
    odrl:target ex:output ;
    odrl:action odrl:use;
    odrl:duty odrl:attribute ]
  .
```

In [7] we considered the set of relations defined by Datanode and the policies defined in the RDF Licenses Dataset to generate a knowledge base of 3865 propagation rules. With the goal of improving the management of the rules, we studied to what extent it is possible to reduce the number of rules to be stored. This reduction requires to be complemented by inferences produced by a reasoner, relying on the axioms of the Datanode ontology. In the present work, we study whether this reasoning is practically feasible, and make the hypothesis that compressing the size of the rule base will not negatively impact the efficiency of the reasoner in computing the propagated policies.

4. (A)AAAA Methodology: overview and *coherency check*

Firstly introduced in [7], the (A)AAAA methodology covers all the phases necessary to set up a compact knowledge base of PPRs[28]. The methodology is based

[24] Flickr APIs Terms of Use. `https://www.flickr.com/services/api/tos/`

[25] Eventful API Terms of Use. `http://api.eventful.com/terms`

[26] LastFM Terms of Service. `http://www.last.fm/api/tos`

[27] Foursquare Developers Policies. `https://developer.foursquare.com/overview/community`

[28] In our work, it has been applied with the support of the command line tool PPR-A-FIVE: `https://github.com/enridaga/ppr-a-five`.

on two assumptions: 1) Policy Propagation Rules are associations between policies and data flow steps, and 2) an ontology is available to organize data flow steps in a semantic hierarchy, e.g., for expressing the fact that the relation *is a copy of* is a sub-relation of *is a derivation of*[29]. The inferences that can be derived from the ontology allow us to remove rules from the knowledge base. The methodology permits to measure the impact of the application of the ontology and supports its evolution with the purpose of maximizing the compression of the knowledge base of PPRs [7]. With respect to the methodology already presented, the novel contribution of this article is the introduction of a *coherency check* method in the *Assessment* phase. In what follows we summarize the methodology, focusing on the *coherency check* element, and refer the reader to [7] for a general overview.

The methodology is composed of the following phases:

A1 - Acquisition.

The initial task is to set up a knowledge base of PPRs. We used the Datanode ontology to extract a list of 115 possible relations between data objects, and combined them with 113 policies derived from the ones defined in the RDF License Dataset. The combination of relations and policies lead to a matrix of 12995 cells. This phase required a manual supervision of all associations between policies and relations in order to establish the initial set of propagation rules. This was performed with the support of the Contento tool [9].

A2 - Analysis.

The objective of the second phase is to detect common behaviors of relations with respect to policy propagation. We achieve this by applying FCA, providing as input the binary matrix representation of the knowledge base R consisting of PPRs. The output of the FCA algorithm is an ordered set of *concepts* C. In FCA terms, each concept groups a set of objects (the concept's *extent*) and maps it to a set of attributes (the concept's *intent*). In our case, each concept represents a set of relations propagating the same set of policies. These concepts are organized hierarchically in a *lattice*, ordered from the top concept T, which includes all the objects and potentially no attribute, to the bottom concept B, including all the attributes with potentially an empty extent (set of objects). All other concepts are ordered from the top to the bottom. For example, usually a first layer of concepts right below T would include large groups of objects all having few attributes in common. Layers below would have more attributes and less objects, until the bottom B is reached. In our case, the top concept T would include all relations and no policy, while the bottom concept B includes all the policies but no relation. The concepts identified by FCA group relations that have a common behavior, as they propagate the same policies. The output of the process is an ordered *lattice* of concepts: clusters of policies that are propagated by the same set of relations.

A3 - Abstraction.

In this phase, we apply a method for subtracting rules in order to reduce the size of the knowledge base. The abstraction process is based on applying an ontology that organizes the relations in a hierarchy (the Datanode ontology). For instance, the relation `dn:hasCopy` is a sub-relation of `dn:hasDerivation`. Intuitively, a number of policies propagated by `dn:hasDerivation` should be also propagated by `dn:hasCopy` and by all the other sub-relations in that branch of the hierarchy. By grouping all the relations below `dn:hasDerivation` in a transitive closure, we obtain a group of relations similar to the ones in the FCA concepts, that we call the `dn:hasDerivation` *branch*. We compute the branch of each one of the relations in the ontology hierarchy. Since we expect branches of the ontology to be reflected in the clusters of relations obtained by FCA, we therefore search for matches between the branches and the concepts of the lattice. When a match occurs, we subtract the rules that can be inferred from the PPR knowledge base.

A general estimation of the effectiveness of the approach is given by the *compression factor* (CF). We calculate the CF as the number of abstracted rules divided by the total number of rules:

$$CF = \frac{|A|}{|R|}$$

with R the set of rules, and A the set of rules that can be subtracted. Concrete examples of the application of this phase can be found in [7].

[29] In our work we rely on Datanode as reference ontology, even if this is not required by the methodology itself.

A4 - Assessment.

The objective of this phase is to assess to what extent the ontology and the FCA lattice are coherent. In particular, we want to:

1. detect mismatches (*coherency check*) to be resolved before using the compressed rule base with a reasoner, and
2. identify *quasi matches* that could become a full match by performing changes in the rule base or the ontology

Coherency check. The abstraction process is based on the assumption that it is possible to replace asserted rules with inferences implied by subsumed relations in the ontology. This requires that all policies propagated by a given relation must be propagated by all the sub-relations in the original (uncompressed) rule base. A coherency check process is necessary to identify whether this assumption does hold for all the relations in each one of the concepts of the lattice.

In case it does not, we want to collect and report all the mismatches in order to be able to fix them at a later stage in the methodology. Listing 4 shows the algorithm used to detect such problems on a given concept in the lattice.

Listing 4: Coherency check algorithm.

```
c = // a concept
M = []
S=SuperConcepts(c)
ForEach s in S
   E=Extent(c)
   E1=Extent(s)
   ForEach(e1 in E1)
      if not(Contains(E,e1))
         ForEach e in E
            if Contains(Branch(e),e1)
              M = [e,e1]|P // Mismatch detected
return M
```

We know from the definition of a FCA lattice that super-concepts will include a larger set of relations propagating a smaller number of policies. Given a concept c, the algorithm extracts the relations (extent) of each of any super-concept (S denotes the set of all super-concepts s of c). In case these relations are not present in (the extent of) c, it is mandatory for them not to be sub-relations of any relation in the extent of c. In case they are, this means that a sub-relation is not inheriting all the policies of the parent one, thus invalidating our assumption. Mismatches M are identified and reported. Listing 5 shows the results obtained by applying the algorithm to Concept 71. In this example, a number of sub-relations of `dn:isVocabularyOf` do not propagate some of the policies of Concept 71.

Listing 5: Coherency check result for Concept 71: mismatches.

Concept	Branch	Relation
71	dn:isVocabularyOf	dn:attributesOf
71	dn:isVocabularyOf	dn:datatypesOf
71	dn:isVocabularyOf	dn:descriptorsOf
71	dn:hasVocabulary	dn:hasDatatypes
71	dn:hasVocabulary	dn:hasDescriptors
71	dn:hasVocabulary	dn:hasRelations
71	dn:isChangeOf	dn:isAdditionOf
71	dn:isChangeOf	dn:isDeletionOf
71	dn:isVocabularyOf	dn:relationsOf
71	dn:isVocabularyOf	dn:typesOf

Quasi matches. The result of the *Abstraction* phase includes a set of measures between concepts and portions of the ontology. Table 2 shows an example of the measures obtained. The measures defined in the *Abstraction* phase are now considered to quantify and qualify the way the ontology aligns with the propagation rules: precision (Pre) and recall (Rec) indicate how close a relation is to being a suitable abstraction for policy propagation. For example, Concept 67 matches with two branches of the ontology hierarchy: `hasPart` and `isPartOf`. The Pre is 1 for the `hasPart` branch, meaning that all the relations subsumed by `hasPart` (`hasSection`, `hasPortion`, etc.) also propagate the policies in Concept 67. Conversely, the Pre with respect to `isPartOf` is 0.86, meaning that some of the relations in this branch apparently do not propagate the policies in Concept 67. Concept 36 covers the branches `hasCopy` and `isCopyOf`, meaning that the related policies are transferred between copies of a given data artifact, re-

Table 2

Excerpt from the table of measures computed during the abstraction phase.

c=Concept ID, ES=Extent Size, IS=Intersection Size, BS=Branch size, Pre=Precision, Rec=Recall, F1=F-Measure.

c	ES	IS	BS	Pre	Rec	F1	Branch
79	52	52	115	0.45	1	0.62	relatedWith
77	46	19	21	0.9	0.41	0.56	hasDerivation
75	44	8	11	0.73	0.18	0.29	samePopulationAs
67	35	7	7	1	0.2	0.33	hasPart
67	35	6	7	0.86	0.17	0.28	isPartOf
36	16	3	3	1	0.19	0.32	hasCopy
36	16	3	3	1	0.19	0.32	isCopyOf
24	12	6	6	1	0.5	0.67	hasVocabulary
9	8	1	1	1	0.12	0.21	hasReification
0	4	4	115	0.03	1	0.06	relatedWith

gardless of the direction of the relation (that specifies which one of the two object was the original). Some general considerations can be made by inspecting these measures. When $Rec = 1$, the whole extent of the concept is in the branch. The branch might also include other relations, which do not propagate the policies included in the concept. When $Pre = 1$, we can perform the subtraction of rules. The perfect match between a concept and a branch of the ontology would be $F1 = 1$. A low recall indicates that a high number of exceptions still need to be kept in the rule set. It also reflects a high ES, from which we can deduce a low number of policies in the concept. As a consequence of that, inspecting a partial match with high precision and low recall highlights a problem that might be easy to fix, as the number of relations and policies to compare will be low. For example, row 2 of Table 2 refers to a comparison between Concept 77 and the `hasDerivation` branch of the ontology hierarchy. Concept 77 includes 46 relations (extent size ES). The *hasDerivation* branch has 21 relations (branch size BS), 19 of which are included in Concept 77 (intersection size IS). Therefore, all except 2 relations in the *hasDerivation* branch propagate the policies of Concept 77 ($BS - IS = 2$). We only need to check whether those 2 relations in the `hasDerivation` branch might also propagate the policies in Concept 77 and then change the knowledge base to obtain a *full match* - when all the relations in a branch are a subset of the extent of a given Concept, all propagating the related policies. The methods to perform this change are the subject of the next section.

At this stage we can make the following considerations:

- The presence of mismatches between the lattice and the ontology will cause the reasoner to return wrong results. They must, therefore, be eliminated.
- The size of the matrix that was manually prepared in the *Acquisition* phase is large (13k cells), and even with the support of the Contento tool it is still possible that errors or misjudgments are made at that stage of the process.
- The Datanode ontology was not designed for the purpose of representing a common behavior of relations in terms of propagation of policies. It should be possible to refine the ontology in order to make it cover the current use case in a better way (and to further reduce the number of rules).

A5 - Adjustment

In this phase we perform operations that change the ontology (or the PPR knowledge base) in order to repair mismatches, correct inaccuracies, refine the hierarchy of relations, and improve the compression factor as a consequence. Six operations can be performed: *Fill*, *Wedge*, *Merge*, *Group*, *Remove*, *Add*. The *Fill* operation modifies the PPR knowledge base by adding all the rules necessary to make an ontology branch being fully covered by a concept, therefore evolving a *quasi match* into a *full match*. All the other actions are targeted to add, remove or reposition relations in the ontology hierarchy (further details about each operation can be found in [7]). The *Assessment* phase of the methodology reported possible mismatches between the FCA output and the ontology hierarchy. These errors must be repaired if we want the compressed rule base to be used by a reasoner. For example, Listing 5 shows the set of mismatches detected for concept 71. In this list, the `dn:isVocabularyOf` branch contains a number of relations that do not propagate the related policies, breaking the assumption that all the policies of `dn:isVocabularyOf` are also propagated by all the other relations in his branch. With the *Fill* operation, we can add all the necessary rules to remove this mismatch.

After each operation, we run our process again from the *Analysis* phase to the *Assessment*, in order to evaluate whether the change fixed the mismatch and/or how much the change affected the compression factor. The process is repeated until all mismatches have been fixed, and there are no other quasi matches that can be adjusted to become full matches. Moreover, when new policies are defined in the dataset of licenses, the process has to be repeated in order to insert the new propagation rules. However, this is only required after changes in the licenses, as changes in the associations between policies and data objects do not affect the PPRs, e.g., changing the license of a data source or adding new data flows.

As reported in Table 3 we performed the process 27 times with the objective to improve the compression and remove errors from the PPR knowledge base, identified by the *coherency check* algorithm. Figure 4 shows how the compression factor CF increases with the number of adjustments performed, while Figure 5 illustrates the progressive reduction of mismatches.

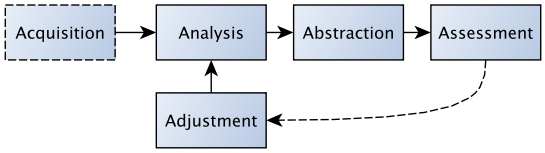


Fig. 3. (A)AAAA Methodology.

Details about the changes performed are provided in Table 3 (identified by the symbol $+$), which also includes statistics about number of mismatches ($\neq$), the impact on number of rules (R), number of concepts generated by FCA (C), number of rules abstracted (A), remaining rules (R_+), and compression factor (CF). Moreover, Table 3 highlights the improvements obtained before (published in [7]) and the further compression obtained after the introduction of the *coherency check* method in the *Assessment* phase (after change 15).

Table 3
List of changes performed.

+	C	$\neq$	R	A	R_+	CF
0	80	15	3363	1925	1438	0.572
1	80	16	3370	1953	1417	0.58
2	80	16	3370	1953	1417	0.58
3	80	16	3480	2283	1197	0.656
4	80	18	3482	2299	1183	0.66
5	78	12	3500	2376	1124	0.679
6	78	14	3608	2484	1124	0.688
7	78	16	3716	2592	1124	0.698
8	96	16	3822	2698	1124	0.706
9	93	15	3824	2706	1118	0.708
10	93	15	3824	2706	1118	0.708
11	93	15	3824	2706	1118	0.708
12	93	15	3824	2706	1118	0.708
13	76	15	3837	2765	1072	0.721
14	76	15	3844	2778	1066	0.723
15	78	15	3865	2817	1048	**0.729**
16	78	13	3866	2818	1048	0.729
17	78	13	3874	2826	1048	0.729
18	63	11	3878	2830	1048	0.73
19	63	11	3882	2834	1048	0.73
20	63	9	3892	2844	1048	0.731
21	55	9	3897	2849	1048	0.731
22	60	8	3898	2850	1048	0.731
23	60	3	3908	2860	1048	0.732
24	54	0	3914	2870	1044	0.733
26	34	0	4225	3451	774	**0.817**

The first column identifies the change performed (starting from the initial state).
C = Number of concepts in the FCA lattice
$\neq$ = Number of mismatches between the FCA lattice and the ontology
R = Number of rules before the process
A = Number of rules abstracted (subtracted)
R_+ = Size of the compressed rule base (without the abstracted rules)
CF = Compression Factor

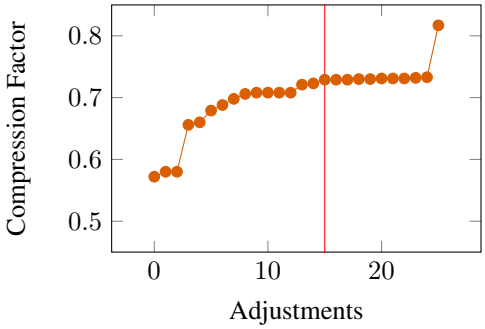


Fig. 4. Progress of the CF.

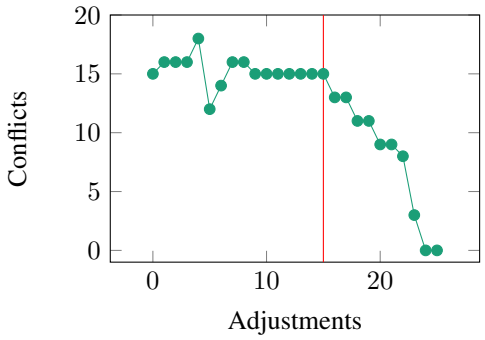


Fig. 5. Progress in the number of conflicts.

Thanks to this methodology we have been able to fix many errors in the initial data, to refine Datanode by clarifying the semantics of many properties and adding new useful ones. The inclusion of a *coherency check* phase is required for a safe use of the compressed rule base with a reasoner. However, the introduction of this approach allowed us to reduce the size even more. As final result we obtained: 4225 rules in total, 34 concepts, 3451 rules abstracted and 774 rules remaining, boosting the CF up to **0.817**.

The version of the ontology prior to performing such changes can be found at `http://purl.org/datanode/0.3/ns/` and the modified version of the ontology can be found at `http://purl.org/datanode/0.5/ns/`. As previously mentioned, the *Acquisition* phase has been performed with the Contento tool [9,8]. The tools used in the other phases of the methodology, from the *Abstraction* to the *Adjustment* phases, can be found at `https://github.com/enridaga/ppr-a-five`.

Table 4
Data flows used in the experiments.

Data flow	has policy	has relation	relations	data objects	policies	sources	output policies
AEMOO-1	6	18	13	10	6	1	6
DBREC-1	6	2	2	2	6	1	3
DBREC-2	6	8	7	5	6	1	6
DBREC-3	6	12	7	8	6	1	6
DBREC-4	6	14	8	9	6	1	3
DBREC-5	6	14	10	10	6	1	6
DBREC-6	6	10	10	6	6	1	3
DBREC-7	6	9	6	10	6	1	6
DBREC-8	6	5	4	5	6	1	6
DISCOU-1	7	22	10	14	7	1	5
DISCOU-11	5	13	9	12	5	1	0
EventMedia-1	37	25	8	24	25	6	4
REXPLORE-1	16	14	8	14	8	3	3
REXPLORE-2	32	23	4	18	14	6	6
REXPLORE-4	32	18	8	14	15	6	3

Highlighted are the maximum and minimum values for each of the data flow inputs. In one case (DISCOU-11), none of the policies attached to the source are propagated to the output.

5. Experiments

The methodology described in the previous section allows to reduce the number of rules that need to be stored and managed. The results of applying this methodology on the PPR knowledge base derived from the RDF Licenses Dataset, show how the compression factor can be dramatically increased after several iterations. Our assumption in this work is that it might positively affect the performance of reasoning on policy propagation. Here, we therefore assess through realistic cases the performance of reasoners when dealing with a compressed knowledge base of PPRs, as compared to when dealing with the uncompressed set.

We took 15 data flow descriptions from previous work [6], referring to 5 applications that rely on data obtained from the Web. Each data flow represents a data manipulation process, consuming a data source (sometimes multiple sources), and returning an output data object. Given a set of policies P_i associated with the input data, the objective of a reasoner is to find the policies P_o associated with the output of the data flow.

The experiments have the objective to compare the performance of a reasoner when using an uncompressed or a compressed rule base respectively. Therefore, each reasoning task is performed twice: at first time, to provide the full knowledge base of PPRs; the second time, to provide the compressed knowledge base in conjunction with the hierarchy of relations of the Datanode ontology (required to produce the inferences).

Reasoners infer logical consequences from a set of asserted facts and inference rules (knowledge base). A reasoner can compute the possible inferences from the rules and the facts any time it is queried, thus exploring the inferences required to provide the complete answer. Alternatively, a reasoner can compute all possible inferences at the time the knowledge base is loaded, and only explore the materialized facts at query time. In order to appropriately address both of those reasoning strategies, we run the experiments with two different reasoners. The first reasoner performs the inference at query time using a backward chaining approach; is implemented as Prolog program and we will refer to it as the *Prolog* reasoner. The second reasoner computes all the inferences at loading time (materialization); is implemented as an RDFS reasoner in conjunction with SPIN rules, and we will refer to it as the *SPIN* reasoner. Both reasoners are implemented in Java within the PPR Reasoner project[30]. Both reasoners have the capability of executing PPRs and expand the results according to the ontology hierarchy.

[30]PPR Reasoner: `https://github.com/enridaga/pprreasoner`. The experiments were performed within the *ppr-evaluation* module, that includes instructions about how to reproduce them.

The *Prolog* implementation is a program relying on JLog, a Prolog interpreter written in Java[31]. The program incorporates a *meta rule* that traverses the set of PPRs, encoded as facts. At the same time, it supports the subsumption between relations. Listing 6 shows an excerpt of the program.

Listing 6: Excerpt of the Prolog reasoner program.

```
i_rdfs_sub_property_of(X,X) .
i_rdfs_sub_property_of(X,Y) :-
        rdfs_sub_property_of(X,Z),
        i_rdfs_sub_property_of(Z,Y) .
i_propagates(X,Y) :- propagates(X,Y) .
i_propagates(X,Y) :-
        i_rdfs_sub_property_of(X,Z),
        propagates(Z,Y) .
i_has_policy(T,P,_) :- has_policy(T,P) .
i_has_policy(T,P,L) :-
        i_has_relation(S,T,R),
        not(visited(S,L)),
        i_propagates(R,P),
        i_has_policy(S,P,[S|L]) .
i_has_policy(T,P) :- i_has_policy(T,P,[]) .
```

The *SPIN* reasoner is built upon the RDFS reasoner of Apache Jena[32] in combination with SPIN[33], a rule engine that allows to define rules using SPARQL. The core part of the reasoner executes PPRs as a SPARQL meta query (Listing 7).

Listing 7: Construct meta-query of the *SPIN* reasoner.

```
CONSTRUCT {
  ?this ppr:policy ?policy
} WHERE {
  ?int ?relatedWith ?this .
  ?int ppr:policy ?policy .
  ?relatedWith ppr:propagates ?policy
}
```

We performed the experiments with the data flows listed in Table 4. Each data flow describes a process executed within one of the 5 systems selected as exemplary data-oriented applications. These data flows were formalized before the present work (in [6]), and were reused for the experiments without changes. However, information about the policies of the input was added. Table 4 illustrates the properties of these data flows, and compares them along several dimensions. The *has policy* column reports the number of statements about policies, from a minimum of 5 to 37 policies. The size of the data flow is reported in the *has relation* column of the table, as it is measured in number of Datanode relations used, spanning from 2 to the maximum of 25. The *relations* column reports the number of distinct relations, the same applying to *data objects*, *policies*, *sources* and the propagated *output policies*. Highlighted are the maximum and minimum values for each of the dimensions. In one case (DISCOU-11), none of the policies attached to the source are propagated to the output.

Each experiment takes the following arguments:

- *Input*: a data flow description
- *Compression*: $True/False$
- *Output*: the output resource to be queried for policies

In case *compression* is $False$, we provide the complete knowledge base of PPRs as input of the reasoning process without including information on subsumption between the relations described in the dataflow. Conversely, when *compression* is set to $True$, the compressed PPR knowledge base is used in conjunction with the Datanode ontology. It is worth noting that the (A)AAAA methodology is also an ontology evolution method, as most of the operations targeted to improve the compression of the rule base are performed on the ontology by adding, removing and replacing relations in the hierarchy. In these experiments, we are considering the evolved rule base (and ontology), that has been harmonized by fixing mismatches between the rule set and the ontology.

The experiments were executed on a MacBook Pro with an Intel Core i7/3 GHz Dual Core processor and 16 GB of RAM. In case a process was not completed within five minutes, it was interrupted. Each process was monitored and information about CPU usage and RAM (RSS memory) was registered at intervals of half a second. When terminating, the experiment output would include: total time (t), resources load time (l), setup time (s), and query time (q). The size of the input for each experiment is reported in the diagrams in Figure 6.

We consider performance on two main dimensions: time and space.

Time performance is measured under the following dimensions:

L Resources load time.

[31] http://jlogic.sourceforge.net/
[32] http://jena.apache.org/
[33] http://spinrdf.org/

S Setup time. It includes L, in addition to any other operation performed before being ready to receive queries (e.g., materialization).
Q Query time.
T Total duration: $T = S + Q$.

Space is measured as follows:

Pa Average *CPU* usage.
M Maximum memory required by the process

Each experiment was executed 20 times. We compared the results of the experiments with and without compression, and verified they included the same policies. In the present report, we show the average of the measures obtained in the different executions. In order to evaluate the accuracy of the computed average measure from the twenty executions of the same experiment, we calculated the related *Coefficient of Variation* (CV)[34]. CV is a measure of spread that indicates the amount of variability relative to the mean. A high CV indicates a large difference in the observed measures, thus reducing the significance of the computed mean. Diagrams 7a and 7b display the CV of all the measures for the *Prolog* and *SPIN* reasoner, respectively. In almost all the cases the CV for the *Prolog* reasoner was below 0.1, with the exception of memory usage M, that in many cases showed a fluctuation between 0.2 and 0.4. Experiments with the *SPIN* reasoner reported a much more stable behaviour in terms of consumed resources, the CV being assessed below 0.1 in almost all the cases, except the Query time of some experiments (the peak is on DBREC-4). However, Q with the *SPIN* reasoner were fluctuating around an average of 10ms, making the observed variation irrelevant. Finally, we consider the computed mean of the observed measures in these experiments to be significant.

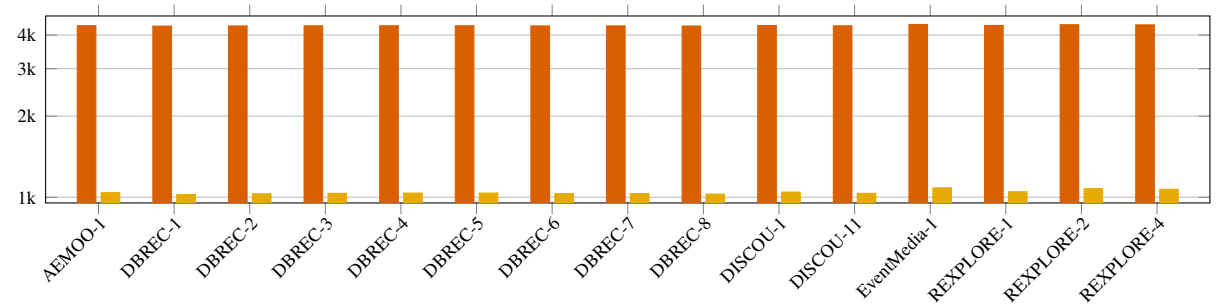
Before discussing the results, it is worth reminding the reader that this evaluation is not targeted to compare the two implementations of a PPR reasoner, but to observe the impact of our compression strategy on the *approaches* of the *Prolog* and *SPIN* implementations, assuming that any other implementation is likely to make use of a combination of the two reasoning strategies they respectively implement.

Figures 8 and 9 illustrate the results of the experiments performed with the *Prolog* and the *SPIN* reasoner, respectively. For each data flow, the bar on the left displays the time with an uncompressed input, and the one on the right the time with a compressed input. We will follow this convention in the other diagrams as well. Figure 8c displays a comparison of the total time between an uncompressed and compressed input with the *Prolog* reasoner. In all cases, there has been a significant increase in performance with the compressed rule base: in three cases (DBREC-5, DISCOU-1, REXPLORE-4) the uncompressed version of the experiment could not complete within the five minutes, while the compressed version returned results in less then a minute. The total time of the experiments with the *SPIN* reasoner (Figure 9c) is much smaller (fractions of a second), having the maximum total time of approximately 2 seconds (EventMedia-1). However, in this case too, we report an increase in every case in performance for all the data flows, with some cases performing much better than others (DBREC-3, DBREC-4). The total time T of the experiment can be broken up into setup time S (including load time L) and query time Q. This observation is depicted in Figures 8a and 9a, and in both cases the impact of the rule reduction process is evident. An interesting difference between the two implementations can be seen by comparing Figures 8b and 9b. The cost of the query time in the *Prolog* reasoner is very large compared to the related setup time S. The *SPIN* reasoner, conversely, showed a larger setup time S with a very low cost on query time Q. The reason is that the second materializes all the inferences at setup time, before query execution. This accounts for the lack of difference in query time between the uncompressed and compressed version of the experiments with the SPIN reasoner.

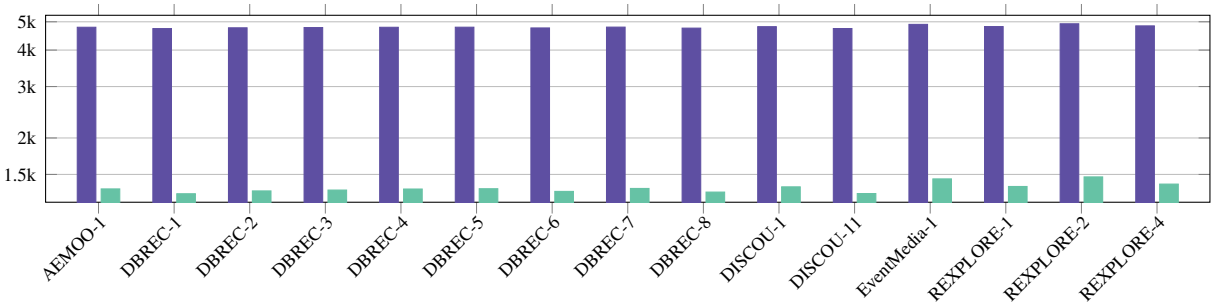
We did not observe changes in Pa for the *Prolog* reasoner (Figure 8d), while the differences in memory consumption M is significant (Figure 8e), demonstrating a performance improvement caused by the compressed input. A decrease in space consumption was also observed using the *SPIN* reasoner (Figures 9d and 9e), even if smaller, and negative in only 2 cases with regard to memory consumption M (DBREC-1 and DBREC-6).

A summary of the impact of the compression on the different measures is depicted in Figures 10 and 11. The first bar on the left of both diagrams illustrates the reduction of the size of the *Input*, while the others how much each measure is reduced. A serious improvement has been achieved in the case of the *Prolog* reasoner, implementing a backward chaining algorithm executed at query time. A PPR reasoner could also be implemented to perform inferencing at loading time

[34]Coefficient of Variation, also known as Relative Standard Deviation (RSD). `https://en.wikipedia.org/wiki/Coefficient_of_variation`



(a) *Prolog* reasoner: input size computed as number of Prolog facts with the original (dark orange) and compressed (light yellow) input for each data flow.



(b) *SPIN* reasoner: input size computed as number of RDF triples with the original (dark purple) and compressed (light green) input for each data flow.

Fig. 6. Input size for the *Prolog* (6a) and *SPIN* (6b) reasoners. It can be deduced that the size of the data flow has a small impact on the general size of the input.

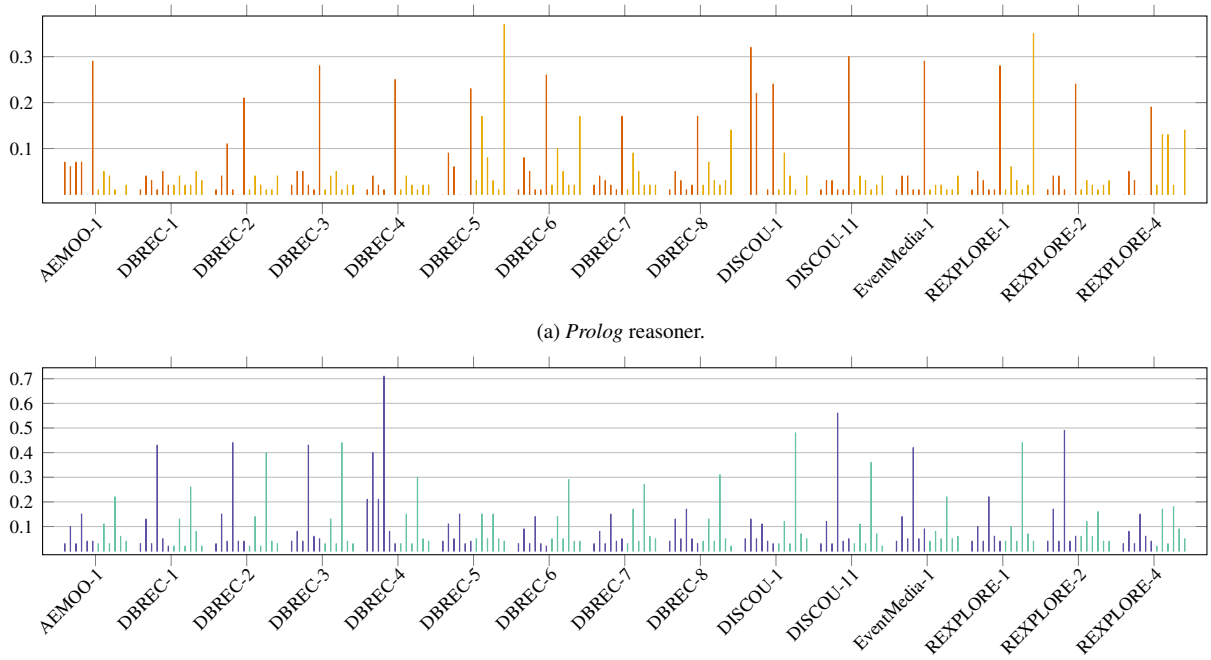
(materialization). The experiments with the *SPIN* implementation is therefore used to show that the effect on reasoning performance exists in both cases, even if in different ways depending on the approach to inferencing. The main conclusion from our experiments is therefore that the methodology presented in [7] and extended with *coherency check* leads to a compressed PPR knowledge base that is not only more manageable for the knowledge engineers maintaining them, but also improves our ability to apply reasoning for the purpose of policy propagation. In addition, it appears clearly that, when dealing with a compressed PPR knowledge base, an approach based on materialization of inferences at load time is preferable to one based on computing the inferences at query time.

6. Conclusions

In this article, we presented an approach for reasoning on the propagation of policies in a data flow. This method is grounded on a rule base of Policy Propagation Rules (PPRs). Rules can easily grow in number, depending on the size of the possible policies and the one of the possible operations performed in a data flow. The (A)AAAA methodology can be used to reduce this size significantly, as demonstrated in [7], by relying on the inference properties of the Datanode ontology, applied to describe the possible relations between data objects. We presented an evolved version of the methodology, which was required to be sure the inferred policies were correct when using the compressed rule base. However, while this activity reduces the size of the input of the reasoner, it requires more inferences to be computed. Therefore, we performed experiments to assess the impact of the compression on reasoning performance. The present article provides two major contributions:

- the (A)AAAA methodology has been extended by including a *coherency check* algorithm, and
- experimental results demonstrating that a compressed knowledge base makes the reasoning on policy propagation more efficient.

This is a preliminary step on studying compression in knowledge management and its impact on reasoning in a more general point of view. Reasoning on policy propagation requires a formalisation of the data flow,



(a) *Prolog* reasoner.

(b) *SPIN* reasoner.

Fig. 7. Coefficient of Variation (CV) of the observed measures. Each experiment was run 20 times. In the diagram we see the CV of measures observed. For each data flow, the diagram shows twelve bars. The first six bars refer to the uncompressed rule base, the second six to the compressed rule base. Each group shows the CV for, in order: T, L, S, Q, Pa and M.

and producing such representation can be time consuming. Recent work by the authors investigate how it is possible to support users in the formalisation of data flows derived from scientific workflows [11]. It would be of interest to explore methods for supporting and automating the generation of such data flows from other pre-existing artefacts (e.g., code bases and their documentation). Future work includes defining new measures to describe the complexity of a data flow and how it affects reasoning on policy propagation, as well as studying the validation of data flows with respect to policies, particularly when multiple sources are used. Finally, we are currently setting up an experimental evaluation (including a user study) to assess the quality of the knowledge base of PPRs produced with this approach, the correctness of the reasoning results with respect to users' expectations, and the effectiveness of the associated methodology in the environment of the MK:Smart Data Hub [12].

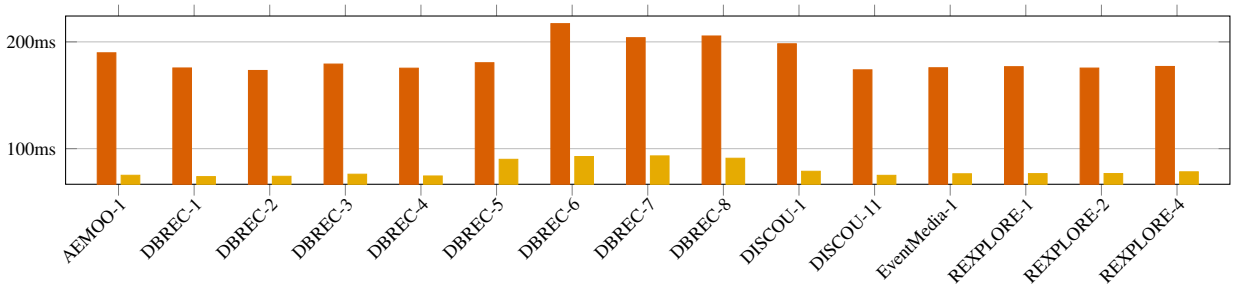
References

[1] G. Antoniou and G. Wagner. Rules and defeasible reasoning on the Semantic Web. In M. Schroeder and G. Wagner, editors, *Rules and Rule Markup Languages for the Semantic Web, Second International Workshop, RuleML 2003, Sanibel Island, FL, USA, October 20, 2003, Proceedings*, volume 2876 of *Lecture Notes in Computer Science*, pages 111–120. Springer, 2003. 10.1007/978-3-540-39715-1_8.

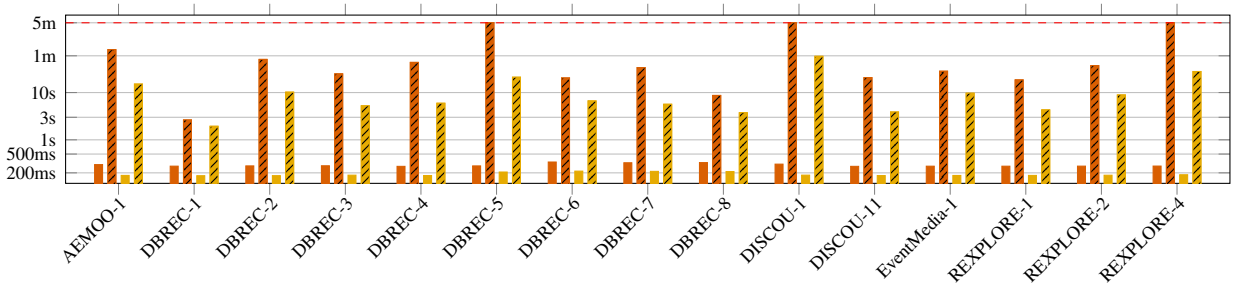
[2] S. Bischof, C. Martin, A. Polleres, and P. Schneider. Collecting, integrating, enriching and republishing open city data as linked data. In M. Arenas, Ó. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. T. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, and S. Staab, editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, volume 9367 of *Lecture Notes in Computer Science*, pages 57–75. Springer, 2015. 10.1007/978-3-319-25010-6_4.

[3] J.-M. Bohli, A. Skarmeta, M. Victoria Moreno, D. Garcia, and P. Langendorfer. SMARTIE project: Secure IoT data management for smart cities. In *2015 International Conference on Recent Advances in Internet of Things (RioT), 7-9 April 2015, Singapore*. IEEE, 2015. 10.1109/RIOT.2015.7104906.

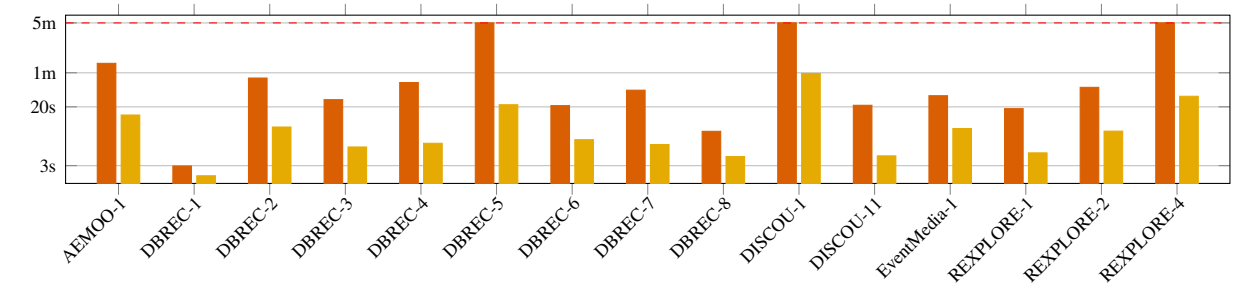
[4] P. A. Bonatti and D. Olmedilla. Rule-based policy representation and reasoning for the semantic web. In G. Antoniou, U. Aßmann, C. Baroglio, S. Decker, N. Henze, P. Patranjan, and R. Tolksdorf, editors, *Reasoning Web, Third International Summer School 2007, Dresden, Germany, September 3-7, 2007, Tutorial Lectures*, volume 4636 of *Lecture Notes in Computer Science*, pages 240–268. Springer, 2007. 10.1007/978-3-540-74615-7_4.



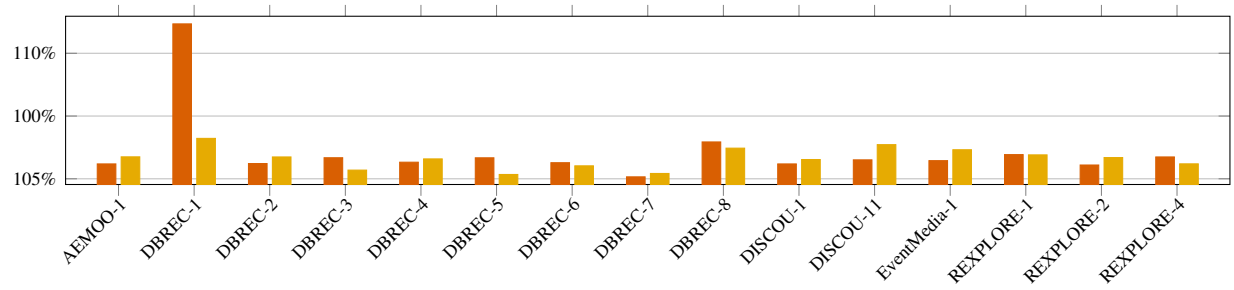
(a) *Prolog* reasoner: load time L.



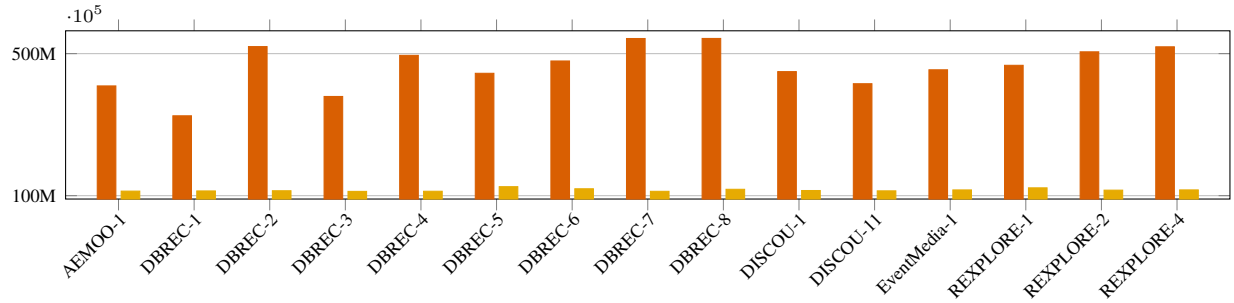
(b) *Prolog* reasoner: comparison of setup time S and query time Q (bars with pattern).



(c) *Prolog* reasoner: total time T.

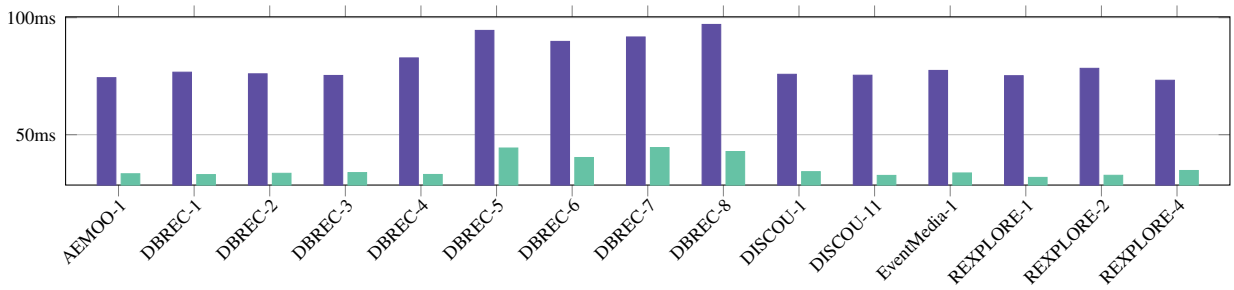


(d) *Prolog* reasoner: average CPU Pa.

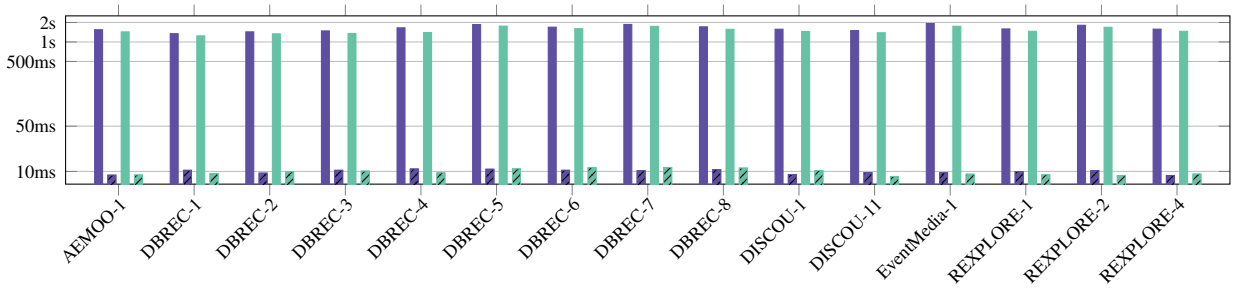


(e) *Prolog* reasoner: max memory consumption M.

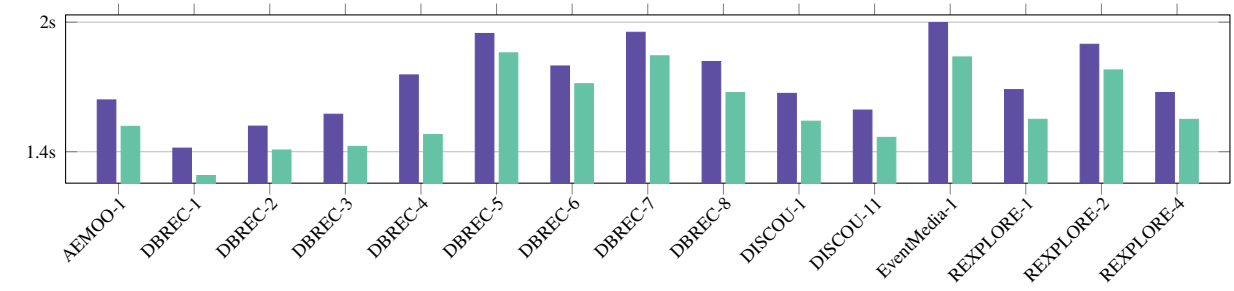
Fig. 8. *Prolog* reasoner: performance measures. Each diagram reports on the performance of this reasoner with an uncompressed or compressed rule base with respect to a given measure. The bars on the left (in dark orange) refer to the uncompressed rule base, while the bars on the right (in light yellow) the compressed one.



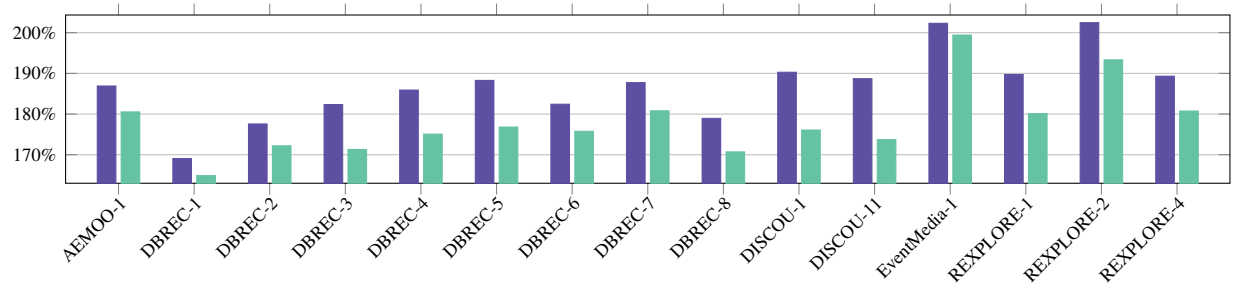
(a) *SPIN* reasoner: load time L.



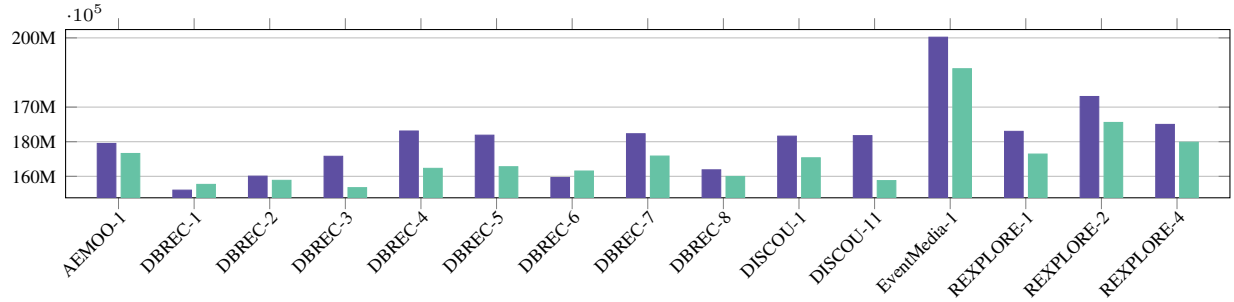
(b) *SPIN* reasoner: comparison of setup time S and query time Q (bars with pattern).



(c) *SPIN* reasoner: total time T.



(d) *SPIN* reasoner: average CPU Pa.



(e) *SPIN* reasoner: max memory consumption M.

Fig. 9. *SPIN* reasoner: performance measures. Each diagram reports on the performance of this reasoner with an uncompressed or compressed rule base with respect to a given measure. The bars on the left (in dark purple) refer to the uncompressed rule base, while the bars on the right (in light green) the compressed one.

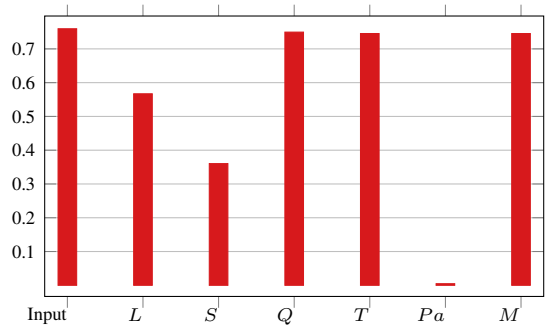


Fig. 10. *Prolog* reasoner: impact of compression on reasoner performance. The bars show the factor by which each measure has been reduced by applying a compressed input.

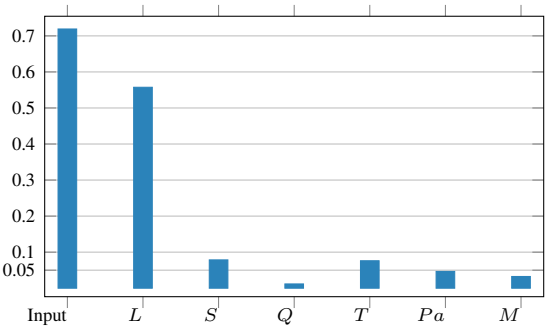


Fig. 11. *SPIN* reasoner: impact of compression on reasoner performance. The bars show the factor by which each measure has been reduced by applying a compressed input.

[5] E. Boros, O. Cepek, and P. Kucera. A decomposition method for CNF minimality proofs. *Theoretical Computer Science*, 510:111–126, 2013. 10.1016/j.tcs.2013.09.016.

[6] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. Describing semantic web applications through relations between data nodes. Technical Report kmi-14-05, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, 2014. URL `http://kmi.open.ac.uk/publications/techreport/kmi-14-05`.

[7] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. Propagation of policies in rich data flows. In K. Barker and J. M. Gómez-Pérez, editors, *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015, Palisades, NY, USA, October 7-10, 2015*, pages 5:1–5:8. ACM, 2015. 10.1145/2815833.2815839.

[8] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. Bottom-up ontology construction with Contento. In S. Villata, J. Z. Pan, and M. Dragoni, editors, *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015.*, volume 1486 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL `http://ceur-ws.org/Vol-1486/paper_76.pdf`.

[9] E. Daga, M. d'Aquin, E. Motta, and A. Gangemi. A bottom-up approach for licences classification and selection. In F. Gandon, C. Guéret, S. Villata, J. G. Breslin, C. Faron-Zucker, and A. Zimmermann, editors, *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, volume 9341 of *Lecture Notes in Computer Science*, pages 257–267. Springer, 2015. 10.1007/978-3-319-25639-9_41.

[10] E. Daga, M. d'Aquin, A. Adamou, and E. Motta. Addressing exploitability of smart city data. In *IEEE International Smart Cities Conference, ISC2 2016, Trento, Italy, September 12-15, 2016*, pages 1–6. IEEE, 2016. 10.1109/ISC2.2016.7580764.

[11] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. An incremental learning method to support the annotation of workflows with data-to-data relations. In E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, editors, *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, volume 10024 of *Lecture Notes in Computer Science*, pages 129–144, 2016. 10.1007/978-3-319-49004-5_9.

[12] M. d'Aquin, A. Adamou, E. Daga, S. Liu, K. Thomas, and E. Motta. Dealing with diversity in a smart-city datahub. In T. Omitola, J. G. Breslin, and P. M. Barnaghi, editors, *Proceedings of the Fifth Workshop on Semantics for Smarter Cities a Workshop at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, volume 1280 of *CEUR Workshop Proceedings*, pages 68–82. CEUR-WS.org, 2014. URL `http://ceur-ws.org/Vol-1280/paper8.pdf`.

[13] R. Gavriloaie, W. Nejdl, D. Olmedilla, K. E. Seamons, and M. Winslett. No registration needed: How to use declarative policies and negotiation to access sensitive resources on the semantic web. In C. Bussler, J. Davies, D. Fensel, and R. Studer, editors, *The Semantic Web: Research and Applications, First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, Proceedings*, volume 3053 of *Lecture Notes in Computer Science*, pages 342–356. Springer, 2004. 10.1007/978-3-540-25956-5_24.

[14] J. M. Gómez-Pérez and Ó. Corcho. Problem-solving methods for understanding process executions. *Computing in Science and Engineering*, 10(3):47–52, 2008. 10.1109/MCSE.2008.78.

[15] G. Governatori, A. Rotolo, S. Villata, and F. Gandon. One license to compose them all - A deontic logic approach to data licensing on the web of data. In H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 151–166. Springer, 2013. 10.1007/978-3-642-41335-3_10.

[16] G. Governatori, H. Lam, A. Rotolo, S. Villata, G. A. Atemezing, and F. L. Gandon. Checking licenses compatibility between vocabularies and data. In O. Hartig, A. Hogan, and J. Sequeda, editors, *Proceedings of the 5th International Workshop on Consuming Linked Data (COLD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014.*, volume 1264 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014. URL `http://ceur-ws.org/Vol-1264/cold2014_GovernatoriLRVAG.pdf`.

[17] P. L. Hammer and A. Kogan. Optimal compression of propositional horn knowledge bases: Complexity and approximation. *Artificial Intelligence*, 64(1):131–145, 1993. 10.1016/0004-3702(93)90062-G.

[18] R. Iannella, S. Guth, D. Pähler, and A. Kasten, editors. *ODRL Version 2.1 Core Model*. W3C ODRL Community Group Final Specification, 5 March 2015. URL http://www.w3.org/community/odrl/model/2.1/.

[19] S. Javanmardi, M. Amini, R. Jalili, and Y. GanjiSaffar. SBAC: "A Semantic-Based Access Control Model". In *11th Nordic Workshop on Secure IT-systems (NordSec'06), Linkping, Sweden*, volume 22, 2006.

[20] R. Kawase, M. Fisichella, K. Niemann, V. Pitsilis, A. Vidalis, P. Holtkamp, and B. P. Nunes. OpenScout: Harvesting business and management learning objects from the web of data. In L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, J. P. M. de Oliveira, F. Lima, and E. Wilde, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 445–450. International World Wide Web Conferences Steering Committee / ACM, 2013. URL http://dl.acm.org/citation.cfm?id=2487962.

[21] H. Khrouf and R. Troncy. EventMedia: A LOD dataset of events illustrated with media. *Semantic Web*, 7(2):193–199, 2016. 10.3233/SW-150184.

[22] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets semantic web - how the BBC uses dbpedia and linked data to make connections. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, editors, *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer, 2009. 10.1007/978-3-642-02121-3_53.

[23] H. Lam and G. Governatori. The making of SPINdle. In G. Governatori, J. Hall, and A. Paschke, editors, *Rule Interchange and Applications, International Symposium, RuleML 2009, Las Vegas, Nevada, USA, November 5-7, 2009. Proceedings*, volume 5858 of *Lecture Notes in Computer Science*, pages 315–322. Springer, 2009. 10.1007/978-3-642-04985-9_29.

[24] T. Lebo, S. Sahoo, and D. McGuinness, editors. *PROV-O: The PROV Ontology*. W3C Recommendation, 30 April 2013. URL https://www.w3.org/TR/prov-o/.

[25] H. Li, X. Zhang, H. Wu, and Y. Qu. Design and application of rule based access control policies. In L. Kagal, T. Finin, and J. Hendler, editors, *Proceedings of the Semantic Web and Policy Workshop, held in conjunction with the 4th International Semantic Web Conference, 7 November, 2005, Galway Ireland*, pages 34–41, 2005. URL https://www.csee.umbc.edu/csee/research/swpw/papers/zhang.pdf.

[26] A. Margara, J. Urbani, F. van Harmelen, and H. E. Bal. Streaming the web: Reasoning over dynamic data. *Journal of Web Semantics*, 25:24–44, 2014. 10.1016/j.websem.2014.02.001.

[27] E. Motta, T. Rajan, and M. Eisenstadt. Knowledge acquisition as a process of model refinement. *Knowledge acquisition*, 2 (1):21–49, 1990. 10.1016/S1042-8143(05)80021-4.

[28] V. Rodríguez-Doncel, S. Villata, and A. Gómez-Pérez. A dataset of RDF licenses. In R. Hoekstra, editor, *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, volume 271 of *Frontiers in Artificial Intelligence and Applications*, pages 187–188. IOS Press, 2014. 10.3233/978-1-61499-468-8-187.

[29] M. Sensoy, T. J. Norman, W. W. Vasconcelos, and K. P. Sycara. OWL-POLAR: A framework for semantic policy representation and reasoning. *Journal of Web Semantics*, 12:148–160, 2012. 10.1016/j.websem.2011.11.005.

[30] R. Shaw, R. Troncy, and L. Hardman. LODE: linking open descriptions of events. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *The Semantic Web, Fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. Proceedings*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009. 10.1007/978-3-642-10871-6_11.

[31] S. Steyskal and A. Polleres. Towards formal semantics for ODRL policies. In N. Bassiliades, G. Gottlob, F. Sadri, A. Paschke, and D. Roman, editors, *Rule Technologies: Foundations, Tools, and Applications - 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings*, volume 9202 of *Lecture Notes in Computer Science*, pages 360–375. Springer, 2015. 10.1007/978-3-319-21542-6_23.

[32] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998. 10.1016/S0169-023X(97)00056-6.

[33] R. Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*, pages 1–33. Springer, 2005. 10.1007/11528784_1.