Graph Databases Lifecycle Methodology and Tool to Support Index/Store Versioning

Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, Nadia Rauch
DISIT Lab, Dep. of Information Engineering, University of Florence, Italy
http://www.disit.dinfo.unifi.it , {pierfrancesco.bellini, ivan.bruno, paolo.nesi, nadia.rauch}@unifi.it

***Abstract*— Graph databases are taking place in many different applications: smart city, smart cloud, smart education, etc. In most cases, the applications imply the creation of ontologies and the integration of a large set of knowledge to build a knowledge base as an RDF KB store, with ontologies, static data, historical data and real time data. Most of the RDF stores are endowed of inferential engines that materialize some knowledge as triples during indexing or querying. In these cases, deleting concepts may imply the removal and change of many triples, especially if the triples are those modeling the ontological part of the knowledge base, or are referred by many other concepts. For these solutions, the graph database versioning feature is not provided at level of the RDF stores tool, and it is quite complex and time consuming to be addressed as black box approach. In most cases the indexing is a time consuming process, and the rebuilding of the KB may imply manually edited long scripts that are error prone. Therefore, in order to solve these kinds of problems, this paper proposes a lifecycle methodology and a tool supporting versioning of indexes for RDF KB store. The solution proposed has been developed on the basis of a number of knowledge oriented projects as Sii-Mobility (smart city), RESOLUTE (smart city risk assessment), ICARO (smart cloud). Results are reported in terms of time saving and reliability.**

Keywords — RDF Knowledge base versioning, graph stores versioning, RDF store management, knowledge base life cycle.

I. INTRODUCTION

Large graph databases are getting a strong push in their diffusion for setting up new kind of big data services for smart cities, digital libraries, competence modeling, health care, smart education, etc. This fact is mainly due to their capability in modeling knowledge and thus on creating Knowledge-Based, KB, systems [Grosan and Abraham, 2011]. Graph databases may be implemented as RDF stores (Resource Description Framework) [Klyne and Carrol, 2004], to create interactive services in which reasoning and deductions can be elaborated including inference engines on top of the store. An RDF store is grounded on the concept of triple that puts in relationship two entities. For example, *Carl knows Paolo*, consisting of a subject, a predicate and an object, which in turn are represented with URI. Predicates, as "*knows*", may be specified by using vocabulary that defines relations. A vocabulary defines the common characteristics of things belonging to classes and their relations. A vocabulary, also called ontology, is defined by using RDFS (RDF Schema, RDF Vocabulary Description Language) or the OWL extension (Ontology Web Language). Recently RDF store have been also addressed as noSQL stores for big data [Bellini et al., 2013a]. A large set of ontologies and related data sets are now accessible, see for example the large number of LOD (linked open data) accessible and related each other via URI [Berners-Lee, 2006], [Bizer et al., 2011]. RDF stores may be made accessible via an entry point to pose semantic queries formalized for example in SPARQL [Hartig et al., 2009] (SPARQL Protocol and RDF Query Language, recursive definition). Non trivial RDF stores based solutions are typically produced by exploiting **multiple ontologies**, loading data triples and testing/validating the obtained results. This means that they are built by using some ontology building methodology [Noy and McGuinness, 2001], [Lopez, 1999], integrated with a knowledge base development life cycles.

The RDF store may grow over time adding new triples, and may have the capacity to learn if endowed of an **inferential reasoner/engine**, i.e., producing new knowledge that are new triples. Thus, the inferential engine associated with the RDF store materializes new triples during reasoning (for example at the time of indexing or querying). These facts are the main motivations to low performances in indexing, and critical performances in deleting triples of RDF stores as graph databases since they are involved in removing the materialized triples in the store. These features impact on store performances, and thus, in literature, many benchmarks for the evaluation of RDF stores are present. Some of them use real data as from dbPedia, UniProt, WordNet, other use synthetically generated data as LUBM [Guo et al., 2005] (university domain), BSBM [Bizer et al., 2009] (e-commerce domain), SP2Bench [Schmidt et al., 2009] (library domain). More recently, in Linked Data Benchmark Council LDBC EU project, two new benchmarks have been developed: one based on Social Network [Erling et al., 2015] and the second on Semantic Publishing. While LUBM and SP2Bench benchmarks are based on real data, and evaluate only the queries performed after the data load. BSBM and LDBC benchmarks evaluate a mix of insert/update/delete/query workloads. When RDF stores are used as a support for a KB, some of the changes in the RDF store can be destructive for the graph model, such as changes in the triples modeling the ontology on which millions of instances are related. In order to keep the performance acceptable, the RDF store has to be rebuilt from scratch or from some partial version to save time in releasing the new version. Thus, the lifecycle may present multiple cycles in which the RDF store is built incrementally via progressive refinements mediating among: (i) reusing ontological models, (ii) increasing the capability of making

deductions and reasoning on the knowledge base, (iii) maintaining acceptable query performance and rendering performances, (iv) simplifying the design of the front-end services, (v) satisfying the arrival of additional data and models and/or corrections, etc. A commonly agreed lifecycle model to build KBs is not available yet and many researchers have tried to embed KB development steps into some conventional software lifecycle models [Batarseh, Gonzalez, 2013]. In general, development of KB systems is a multistep process and proceeds iteratively, using an evolutionary prototyping strategy. A number of lifecycle models have been proposed specifically for KB systems [Milette 2012].

In the lifecycle model, a change in the ontology may generate the review and regeneration of a wide amount of RDF triples. The problem of ontology versioning as addressed in [Klein et al., 2002], [Noy and Musen, 2004] can be easily applied if the ontology is not used as a basis for creating a large RDF KB store. Moreover, in [Volkel et al., 2005], the versioning of RDF KB has been addressed similarly to the CVS solutions by using commands as: *commit*, *update*, *branch*, *merge*, and *diff*. The differences are computed at semantic level on files of triples. Thus, [Zegins et al., 2007] presented a solution for versioning RDF models assuming the possibility of estimating the delta between two RDF models by performing a set of adds and deletes to a model to transform it to the other. At database level, the key performance aspects of an RDF KB store version management are the storage space and the time to create a new version [Tzitzikas et al., 2008]. Therefore, possible approaches could be to store: (a) each version as an independent triples store [Klein et al., 2002], [Noy and Musen, 2004], [Volkel et al., 2005]; (b) the deltas in terms of triples between two consecutive versions and implementing a computationally expensive and time consuming chain of processes to maintain and apply deltas [Zegins et al., 2007].

In this paper, a versioning system for RDF KB proposes to integrate both (a) and (b) solutions. It manages versioning of RDF stores by: (i) keeping trace of the set of triples to build each version, (ii) storing each version and related set of triples, (iii) providing an automated tool for keeping trace of triple files, descriptions for store building and stores, (iv) allowing the versioning of the RDF KB store, (v) reducing the critical manual error prone operations. This approach allows to make indexing versioning for RDF stores that materialize triples at indexing (as OWLIM [http://www.ontotext.com/]) or at querying (as Virtuoso [http://virtuoso.openlinksw.com/]) without influencing the RDF store reconstruction. The resulting time for returning to a previous version and to reconstruction of a new one is satisfactory and viable, since some of the RDF stores are very time consuming in indexing, while other do not allow the deletion of triples. Therefore, the paper presents an RDF KB methodology life-cycle suitable for big data graph databases, and a versioning tool for RDF KB stores that has been developed and tested for SESAME OWLIM and Virtuoso; and thus it can be simply extended to other RDF stores. The solutions have been developed for Km4City project [Bellini et al., 2013b], and adopted for other RDF KB oriented projects as Sii-Mobility Smart City national project and RESOLUTE H2020 European Commission Project. They are large KB oriented projects in the Smart City, smart cloud, smart railway domains, developed at the DISIT Lab of the University of Florence http://www.disit.org/6568 .

The paper is organized as follows. Section II presents the RDF Knowledge Base life-cycle model and methodology for development. In Section III, the RDF KB indexing flow and requirements for the RDF Indexing Manager tool are presented. Section IV describes the RDF Index Manager tool, detailing the architecture, and the XML formal model for index descriptors. In Section V, experimental results are reported providing data related to real cases, in terms of time and managed complexity. Conclusions are drawn in Section VI.

II. A KNOWLEDGE BASE LIFE-CYCLE

Building a RDF KB is a challenging practice that needs a well-defined methodology and lifecycle to keep under control the entire development process. RDF KBs are mainly developed thanks to a cycle approach that allows checking and validating the advances made, and if needed, to make adjustments when a problem is identified. As stated above, the lifecycle proposed in this paper has been derived from the DISIT Lab experience cumulated while developing a number of big data RDF KBs.

The proposed methodology and lifecycle for RDF KB is reported in **Figure 1**. The life-cycle presents 4 vertical pillars and one horizontal block that represents the **RDF Store usage and Maintenance**. The life-cycle spans from the ontology creation to the RDF Store usage on the front-end where also real time data are added.

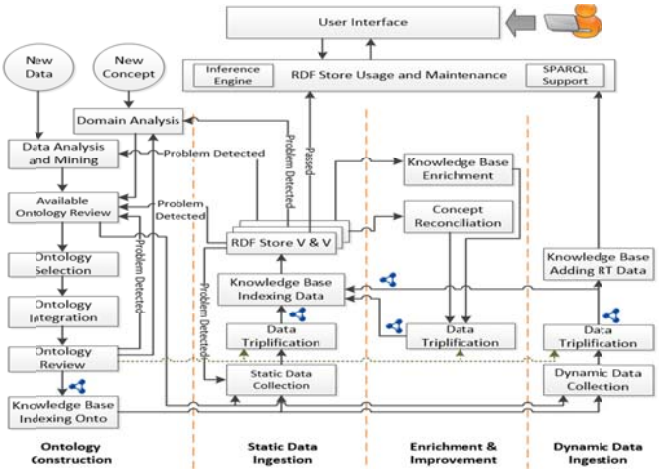

Figure 1. RDF KB Life Cycle Model

The pillars refer to the:

Ontology construction, from domain analysis the setup of the RDF Store containing triples of the selected ontologies and possible additional triples to complete the domain model (Knowledge Base O, KB-O). For example, the Km4City ontology reuses: *dcterms* to set of properties and classes for modeling metadata; *foaf* dedicated to relations among people or groups; *schema.org* for a description of people and organizations; *wgs84_pos* representing latitude and longitude; *GoodRelations* for a description of business entities and their

locations; *OWL-Time* for temporal modeling; *OTN* for transport aspects; *GIS Dictionary*, to represent the spatial component of geographic features; etc. [Bellini et al., 2013b]. The combined ontology is reviewed and possible problems may lead to more or less deep redefinition of the process.

Static Data Ingestion: this phase is related to the loading of the data instances of the ontological classes and attributes. Despite their name, static data may change rarely over time, for example, the position of bus stops may be considered static data even if they change seasonally. They come from several sources (static, statistical, historical, etc.), and have to be converted in triples according to the KB-O coming from the previous phase. Then, they are finally indexed by using several sets of triples, maybe thousands. The indexing produces a KB including the former KB-O, plus many data instances; thus, allowing performing the Verification and Validation, V&V, of the RDF KB.

The V&V phase may be the moment in which some problems are detected. They may constrain the expert to: (i) wrong data or incomplete data to need a review of the data mapping to the ontology (restart from the first step of this phase of data collection), (ii) missing ontology aspects and classes, thus leading to the review of the ontology built (returning to Ontology Review), (iii) problems in data collected that may be wrongly mapped to ontology classes (returning to Data Analysis and Mining), (iv) mistake in data mapping that may lead to revise the whole Domain Analysis, and successive steps. If this phase is passed, the RDF Store passes to the phase of RDF Store Usage and Maintenance. Additional static data sets may be added to the KB-O if the ontological model supports them without deletion, otherwise a review is needed.

Enrichment and Improvement, E&I: this phase allows solving problems that may be present in the produced RDF Store. E&I processes may take advantage from the access to the partially integrated KB, exploiting for examples solutions of Link Discovering [Ngomo, 2011], [Isele, Bizer, 2013], and/or making tuned semantic queries. Additional processes of E&I may be added to the RDF Store if the model supports them without performing some delete otherwise a model review is needed.

Dynamic Data Ingestion: when the RDF store is in use, collected data from real time information (for example, bus delay with respect the arrival time, weather forecast, likes on the user profile, status of sensors, status of cloud processes, etc.) can be added to the RDF Store and saved into the repository of the historical triples. Additional dynamic data sources may be added to the RDF Store if the model supports them without performing some delete otherwise a model review is needed. Please note that dynamic data should not need to validate and verify process since the data to be added in real time are new instances of data already mapped and integrated as historical data.

A. Data & Domain Analysis and Ontology Construction

Brief descriptions of major interesting blocks pertaining to the proposed life cycle and methodology (see Figure 1) are now provided.

Data Analysis and Mining: Each data set (static or real-time) to be addressed in the RDF KB is analyzed and checked to assess if the information related to each single data field is well described in terms of type, range, and context. The data collected is analyzed to understand the concepts in terms of their structure, relationships and information in domain.

Domain Analysis: this step is executed in parallel or in alternative to the above data analysis steps. In this phase, the concepts of the domain addressed by the application are studied to understand concepts, terminology, their relationships, and the general rules that are related to them. Several methodologies are accessible to help the analysts in identifying concept from the literature review of the domain – as well as thumb rules: substantive are classes, verbs are relationship, details are attributes, etc.

Available Ontology Review: This phase is very important. Once the major aspects of the domain have been identified. The phase consists of studying other related ontologies at the state of the art to see if they can model the whole identified domain and data concepts or at least a part of them. The realistic solution is to start from one or a set of available ontologies and complete the expected model with some specific classes and relationships. This task, it could be performed every time new static and/or dynamic data kind have to be addressed, or for addressing identified problems. When this ontology review is performed starting from an active domain ontology (in the RDF KB), it may happen that the expert may discover that no changes are needed at ontological level (e.g., a new class is not needed since the concept for hosting data is already in place), thus resulting in a direct jump to the phases of static or dynamic data management.

***Ontology Selection*:** on the basis of the actions previously performed on concepts and data against ontologies, it is possible to make a selection of the most suitable ontologies to be taken as seeding concepts. The process of selection has to take into account also the licensing aspects, which impose some constraints. For example, some of licenses of the ontologies do not allow being tuned/modified. If the study has not led to any results, it is always possible to write a specific domain ontology.

Ontology Integration: as a result of the previous steps the main ontologies have been identified and thus they have to be integrated/glued with each other. In addition, the missing concepts have to be formalized by completing the fitting of the KB with the domain analysis performed. ***Ontology Review***: Once the ontology was created/modified, a first revision took place even without the massive loading of instance. Thanks to tools like *Protégé* [http://protege.stanford.edu/], which allow to apply a reasoner to the ontology in order to verify that knowledge is modeled as desired. A number of metrics and criteria may be also applied to verify if the ontology has been developed with common criteria. E.g., [Noy and McGuinness, 2001], [Gómez-Pérez, 2004], [Rector et al., 2004].

***Knowledge Base Indexing Onto, KBIO*:** the task in which the RDF Store index containing the selected ontologies, vocabularies, and custom defined concepts are integrated as

triples. They may be some files and some tens of classes. This process usually starts from an empty RDF store and takes a few seconds since the ontologies are comprised of a small number of triples and the RDF is empty; differently from what happen when millions of triples of data sets are indexed and they lead to many materialized triples.

B. From Ontology to KB via Data Ingestion, major tasks

Static Data Collection: on the basis of the created domain ontology, the analysed data (addressed in task Data Analysis and Mining) have to be processed. Static data are typically obtained from open data, statistical data, private data that do not change over time so rapidly. The process of static data ingestion may be performed by means of parallel and distributed architectures executing processes as ETL (Extract, Transform and Load), Java, microgrid, harvesting, crawling, etc. It may include: file access, REST/WS calls, data mapping, quality improvement, e.g., [Bellini et al., 2013b]. ***Data Triplification*:** A task in which the data (static, dynamic) are mapped to triples on the basis of the domain ontology model.

***Knowledge Base Enrichment*:** task focused on enriching the RDF KB Store by adding links to external LOD. For example, referring from the street title to VIP to its dbPedia definition (from Avenida Winston Churchill, to its page on dbPedia: http://live.dbpedia.org/page/Winston_Churchill). For example for the Km4city KB a tool has been created, that allows to identify famous names inside the KB and search for the same name on dbPedia, to finally create triple *cite*/*isCitedBy* thanks to the *CITO Ontology* [http://purl.org/spar/cito].

***Concept Reconciliation*:** task related to solve the lack of coherence among indexed entities referring to the same concept but coming from different data sets. This process is a critical step during the KB realization and helps to create new knowledge and new connections between data that would otherwise remain unconnected. For example, different services located at the same street number, several profile aspects of the same person, different representations of the same part of the brain. This task typically produces a number of triples solving the problem of missing links possible. Triples includes relationships of *owl:sameAs.*

Dynamic Data Collection: Dynamic data are subject to a lighter ingestion process with respect to static data. In fact, they are picked up and immediately mapped into RDF triples, in order to speed up as much as possible the process that allows making them available to users adding them to the RDF store (*Knowledge Base Adding RT Data).* At the same time, Real Time triples are stored as Historical Dynamic Data for successive construction of versioned data stored.

Knowledge Base Indexing Data, KBID: This task takes in charge a high number of triples coming from different data sets:

- **Static data**: for example one or more file containing a set of triples for each single data set;
- **Historical Dynamic data**: several files and triples for each real time data collection channel. For example, the collected weather forecasts of the past two months, the last 200 measures of traffic flow sensor DG32453165, the data regarding the Cloud Host and VM in the last week;
- **Reconciliated data**: triples connecting concepts and data into the RDF KB;
- **Enrichments data**: triples connecting data entities of the RDF KB to external LOD RDF stores. When the enrichment tasks are performed on real time data, they have to be performed in real time as well. For examples if the enrichment is performed on an Opera Name, or about a VIP person name.

In order to pass from the ontological model to a real RDF KB store, many data sets (static, statistical and historical), should be included / indexed in the RDF KB. Very often, indexing process of large files may take several hours. Often files of triples are linked each other and the order of indexing of these data may becomes essential. In some cases, the historical data can lead to very huge number of triples, thus compromising / influencing the performance of the whole RDF Store. This implies that the RDF KB has to be periodically polished by removing most of the cumulated historical data. This activity is quite natural for smart city and smart cloud applications. For example in cloud monitoring systems as NAGIOS, data are dense in the close time and sparse in the past.

C. RDF Store Verification and Validatiton, V&V

Once the RDF KB Store containing triples coming from data (static, historic, reconciliation and enrichment) has been produced, it is possible to precede with the validation and verification of the RDF Store vs the ontological definition. Please note that, the RDF store index has to be accessible to perform the following V&V processes via semantic queries end analyzing consistency. They can be automatically performed through a set of validation processes implemented as SILK [Isele, Bizer, 2013] as well as SPARQL processes.

The verification and validation process has the duty to detect inconsistencies and incompleteness: (i) verify if the data indexing has been correctly performed, (ii) detect eventual reconciliations to be performed identifying missing connections, (iii) identify eventual enrichments to be performed, (iv) identify eventual mismatch from data loaded and the ontology (for example counting the triples to be indexed and those indexed in reality), (v) verify if the expected inferences are exploited at the query time, etc. The above mentioned criteria allow identifying different kind of problems that may lead to revise the ontological model, the data ingestion process, etc. etc.

III. RDF INDEXING FLOW AND REQUIREMENTS

As described in the previous section, there are several reasons for which into the RDF KB life cycle the process may lead to (i) revise the ontology (and thus to revise the data mapping and triplification invalidating the indexing and the materialization of triples); (ii) revise the data ingestion including a new data mapping, quality improvement, reconciliation, enrichment and triplification. As stated in the previous section, the life-cycle model foresees two steps where the Knowledge Base Indexing has to be performed: KBIO, KBID. On the other hand, as pointed out in the introduction, in most of the RDF store models, the versioning is not an internal feature. This is due to

the fact that it cannot be easily performed at level index and stored triples for their complexity in removing them, due to the triples materialization by inference. According to the proposed RDF KB life cycle, the modeling of a chain of connected versions of *indexes/RDF Stores*, with incremental complexity may be very useful to keep under control the evolving index with the aim of saving time by exploiting intermediate versions in generating the RDF Store/index for the successive deployment. For example, in the case of Smart City, the layered versions of the index may include the ontology, static and dynamic data, historical data, etc.

To better describe the process of RDF Index versioning, it is necessary to put in evidence the differences between the "*index*" and "*index descriptor*". An RDF KB store is in substance an "*index*", while content can be accessed via URI cited in the triples elements. The index is created by loading the triples into the RDF store, and as a result a binary index is built, maybe materializing additional triples according to the ontological model and the specific RDF store inferential engine adopted. The recipe to create the RDF Store index, that is the collection of atomic files containing triples (including triples of ontologies as well as those related to data sets: static, historical, dynamic), can be called as the "*index descriptor*", that may be used to generate a script for index generation. The script syntax can be different from an RDF Store to another, since their commands for loading and indexing can be different. This approach implies to have aside each pair "*index*" and "*index descriptor*" also the history of files containing triples with their versions, last update dates, and dependencies from other files. For example, see **Figure 2,** where the reconciliation of triples connecting parking locations (File 1, ver 1.5) with respect to civic numbers depends on the ontology and on the parking area data sets. Thus leading to create a set of triples connected with dashed lines.

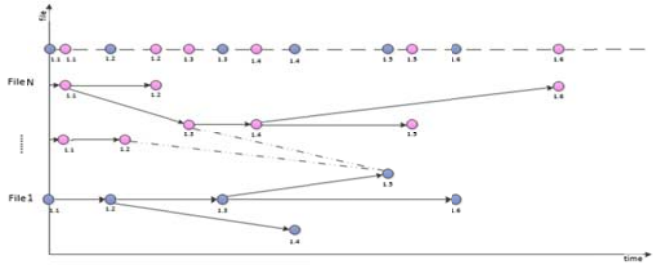

Figure 2. Example of set of file versioning

Definition. Let $F = \{f_1, f_2, f_3, \dots\}$ be the set of triple files that are available for indexing and $DS = \{ds_1, ds_2, ds_3, \dots\}$ is the set of datasets and ontologies that are available for ingestion. The function $ds: F \to DS$ associate the file to the dataset it belongs to, function $time: F \to \mathbb{N}$ associate each file with the time when it was created and function $dep: F \to \wp(F)$ associate each triple file with a set of files that it depends on (e.g. ontologies), $\wp(X)$ is the power set of set X. The *dep* function must not introduce a cyclic dependency among files. Moreover, a file should not depend on files created in the future:

$$\forall f \in F, \forall\, s \in dep(f).\, time(s) < time(f)$$

Example $\boldsymbol{DS} = \{km4c, otn, roads, services, busses\}$,
$\boldsymbol{F} = \{kf_1, kf_2, of_1, rf_1, rf_2, sf_1, sf_2, bf_1, bf_1, bf_2,\}$,
$\boldsymbol{ds} = \{(kf_1 \to km4c), (kf_2 \to km4c), (of_1 \to otn),$
$(rf_1 \to roads), (rf_2 \to roads), (sf_1 \to services),$
$(sf_2 \to services), (bf_1 \to busses), (bf_2 \to busses)\}$
$\boldsymbol{time} = \{(kf_1 \to 2), (kf_2 \to 5), (of_1 \to 1), (rf_1 \to 3),$
$(rf_2 \to 8), (sf_1 \to 2), (sf_2 \to 8), (bf_1 \to 3), (bf_2 \to 8)\}$
$\boldsymbol{dep} = \{(kf_1 \to \{of_1\}), (kf_2 \to \{of_1\}), (rf_1 \to \{kf_1\}),$
$(rf_2 \to \{kf_2\}), (sf_1 \to \{kf_1\}), (sf_2 \to \{kf_2\}),$
$(bf_1 \to \{kf_1\}), (bf_2 \to \{kf_2\})\}$

Definition A subset S of F is *indexable* iff

$$\forall\, f, f' \in S, f \neq f' \to ds(f) \neq ds(f')$$

Meaning that files need to be associated with different datasets. **Example** the set $\{kf_1, of_1, rf_1, sf_1\}$ is *indexable* while $\{kf_1, rf_1, rf_2\}$ is not indexable because $ds(rf_1) = ds(rf_2) = roads$.

Definition The function $C: \wp(F) \to \wp(F)$ associates a subset of F with closure of the subset with respect to the *dep* function. It can be computed using the recursive function:

$$C(S) = \begin{cases} S \cup C(dep(S) \backslash S) & S \neq \emptyset \\ \emptyset & S = \emptyset \end{cases}$$

Where:

$$dep(S) = \bigcup_{s \in S} dep(s)$$

Example $C(\{rf_1, sf_2\}) = \{kf_1, kf_2, of_1, rf_1, sf_2\}$
Definition Let $I = \{i_1, i_2, i_3, \dots\} \cup \{\varepsilon\}$ be the set of indexes produced and ε is the empty index. The function $from: I \to I$ associates an index with the index it was started from and the function $files: I \to \wp(F)$ associate an index with the set of files to be added to the index we are starting from. Consider that the "*from*" function must not introduce a cyclic dependency among indexes.
Example: $\boldsymbol{I} = \{i_1, i_2, i_3, i_4\}$
$\boldsymbol{from} = \{(i_1 \to \varepsilon), (i_2 \to i_1), (i_3 \to i_2), (i_4 \to i_2)\}$
$\boldsymbol{files} = \{(i_1 \to \{kf_1\}), (i_2 \to \{rf_1, bf_1\}), (i_3 \to \{sf_1\}),$
$(i_4 \to \{sf_2\})\}$

Definition Function $\phi: I \to \wp(F)$ provides for each index the set of files that are indexed, it is defined recursively as:

$$\phi(i) = \begin{cases} \phi(from(i)) \cup files(i) & i \neq \varepsilon \\ \emptyset & i = \varepsilon \end{cases}$$

Example $\phi(i_1) = \{kf_1\}$, $\phi(i_2) = \{kf_1, rf_1, bf_1\}$, $\phi(i_3) = \{kf_1, rf_1, bf_1, sf_1\}$, $\phi(i_4) = \{kf_1, rf_1, bf_1, sf_2\}$

Definition An index $i \in I$ is *correct* if $C(\phi(i))$ is indexable meaning that in the closure of files in the index are not present different versions of files of the same dataset. **Example** the indexes i_1, i_2, i_3 are correct while i_4 is not correct because $C(\phi(i_4)) = \{kf_1, kf_2, of_1, rf_1, bf_1, sf_2\}$ is not indexable.

Figure 3 shows possible evolutions of an RDF Store with their corresponding index-descriptors and indexes. The figure wants to highlight that simultaneously can be carried out different versions of pairs: index, index-descriptor, each of which

containing different data. The different index colors indicate that each index may contain different data, according to the evolution with which it has been created. For example, considering the index, index-descriptor pair version labeled 1, including ontologies and vocabularies, we can assume that the pairs number 1.1 could be incrementally generated, starting from version 1 by adding geographic information and bus stops; and version 1.1.1 by adding services. Subsequently the need to create another alternative branch occurred since the bus stops changed positions, and thus version 1.2 was created by adding geographic information and the new bus stops; and from that version 1.2.1 adding again the services. Please note that version 1.2.2, represents an example of index generation by starting from version 1.2 by cloning index and index-descriptor, and adding other data set triples.

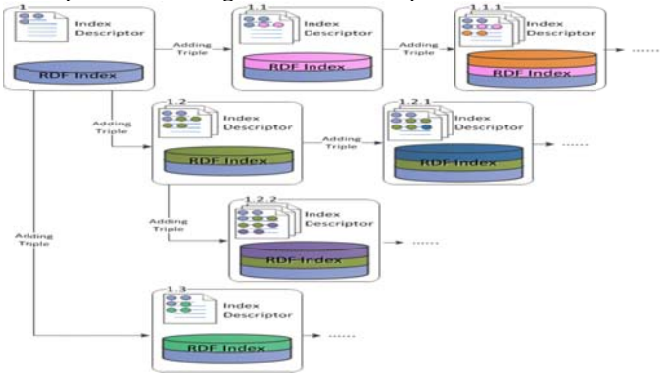

Figure 3. RDF KB store index versioning: reporting index-descriptors and indexed

In this last scenario, an existing index may be extended with new generated data set or updated by including new corrected versions of data and/or ontologies. Since the RDF KB building is an evolving process, it is not possible to predict whether one has to keep a specific previously created version of the index or not. Any small change could be used to generate a new version, while the suggestion is to save versions every time a consolidated point is available similarly to virtual machine snapshots. Moreover, since the triples associated with each single data set are accessible, reconstruction of partial intermediate versions are also possible, saving time in generating triples. Furthermore, each times some ontologies change, most of triples must be generated again, and therefore, for the same dataset, more triples versions could exist.

*A. Requirements for **RDF Index Manager Tool***

On the basis of the above presented model, the RDF KB indexing versioning activities described can be supported by means of an RDF Index Manager (RIM), that should allow:

- Keep tracing *RDF KB Store Versions*, *RKBSV*, in terms of *files of triples*, *index-description*, and RDF *Index*;
- Maintaining a repository of *RKBSVs* where they could be stored and retrieved;
- Selecting a *RKBSV* from the repository for modification, to examine changes and the history version, to be used as base for building a new version;
- Managing the *index descriptor* as a list of files containing triples;
- Generating a RDF KB index on the basis of an *RKBSV* independently from the RDF store kind automatically, and in particular for SESAME OWLIM and Virtuoso;
- Monitoring the RDF KB index generation and the feeding state;
- Suggest the closest version of the *RKBSV* with respect to the demanded new index in terms of files of triples;
- Avoiding manually managing the script file of indexing, since it is time consuming and an error prone process.

IV. RDF INDEX MANAGER TOOL

The **RDF Index Manager** tool satisfy the above presented requirements, creates and manages *index descriptors*, and files of triples, and generates automatically the corresponding *indexes* independently from the RDF store type. The *index descriptor*, as mentioned before, is a list of ontologies and related data sets described with their triple files and version. The chosen approach with generation and update is to: (i) build the entire index (*build all*) by loading triples when ontologies and related data set change, (ii) extending the index when only new data sets and triples have to be added *(incremental building)*, (iii) make a physical copy (*clone*) of a consolidated RDF index when an index descriptor is built starting from an older consolidated descriptor. The big amount of triples to load in the index suggested exploiting the bulk data loading supported by many RDF stores.

The main functionalities provided by the tool are described as following: Setup of a new index descriptor, to create an empty index descriptor; Clone a previous index descriptor to create a new version that it is populated with the same data sets and triples version of the parent with some addition. A clone of the parent RDF index is made and used to build the new store loading the new additional data sets; Copy a previous index with updated versions to create a new version populated with same data sets of parent and new versions of triples. This allows speeding up the creation of an update version of the index descriptor. A new RDF index will be created and loaded from scratch; Edit the index descriptor to add a data set (ontology, static, historical and reconciliations), select triples version; update triples version of a data set; remove a data set; Import/Export the index descriptions as XML representations that could be used for backup/restore and share; RDF Index Generation by producing a scripted procedure (for Windows and Linux) according to the index descriptor and the selected RDF store kind. The procedure may be incremental or for reconstructing the index from scratch; Monitoring the RDF Index Generation by controlling the store feeding as: the queue of data set to be loaded, the data set already in the store, time indicators (time spent, max time to upload a data set, etc..), progression and output of building process; Logging building data related to RDF store building for further access (i.e. statistical and verification analysis).

A. *Architecture, RDF Index Generation and evolution*

The RDF Index Manager is constituted by the following components. The *RDF Store Manager* manages different versions of RDF Stores exploiting the *Version Manager* which provides the triples files version for all the data sets. The *Index Manager Application Server and GUI* which is the user interface for creating, loading and editing the index, building and putting in execution the scripts for RDF store feed, monitoring the whole creation process. It also provides users management, user control access and configuration settings. The *Index Manager API REST Interface* consists of a set of REST calls to be invoked by the indexing script during the RDF store building to keep trace of the indexing process status. The *Index Builder Manager* generates the scripts according to the RDF Store kind. The section contains a list of ontologies/file and each file is described by: an unique identifier corresponding to the name, the reference to the index, the version of triples to use, the operation to perform add, update, remove and commit, and an entity for setting if it was inherited by a cloning (Clone). The historical data differs from other section for the presence of time interval that defines the triples to use (date and time for TripleStart and TripleEnd).

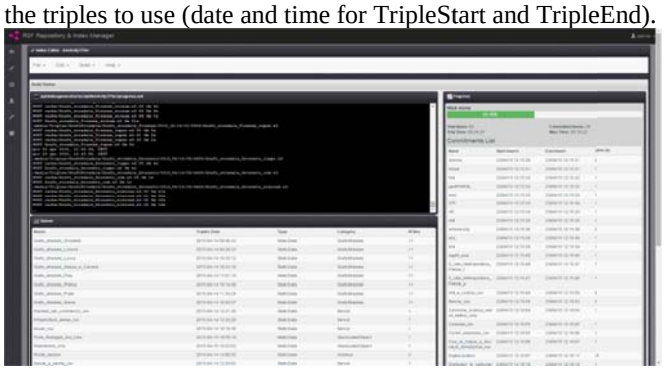

Figure 4. RDF Index Buidling Monitor

For the RDF Index generation the RDF Index Manager produces a script according to the index descriptor and the RDF store target. The script is structured in the following steps: (i) setup of script, (ii) initialization of RDF store, (iii) bulk uploading of triples into the store, (iv) RDF store finalization, (v) create possible additional indexes as textual indexes, geographical indexes that need additional database commands, and (vi) update index building status.

The RDF Index Manager has been realized as a PHP 5.5.x web application with MySQL support running under Ubuntu. The **Figure 4** shows the Building Monitor View when a batch script is running. This view provides different information panels: the output of script in real-time on top, the queue of data set to insert, the progress and the total time spent for the committed data set. Such information allows also evaluating the time necessary to build a repository using the two RDF Stores.

V. EXPERIMENTAL RESULTS

In Table 1, examples of results are reported. The data refer to the comparison of the usage of the RIM and versioning in building a Smart City RDF store. The RDF stores currently managed are Virtuoso 7.2 as open source RDF store and the commercial OWLIM SE ver. 4.3 and GraphDB 6.1. The measures reported have been performed by means of an incremental building of the RDF Store for the three solutions. The building started with 12 files of triples including ontologies (first column), then each column of the table refers to the added triples/files (street graphs, smart city services, enrichment and reconciliations, historical data of real time data for 1 month). The time estimated for the cases of total indexing include: create, load, finalize; while those for incremental indexing include: clone, load, finalize. The three RDF store kinds have a different behavior. OWLIM and GraphDB create inferred triples at the indexing; this determines a higher number of triples with respect to Virtuoso, i.e., 73.4 wrt 46.2 million; and a higher indexing time. In both cases, the percentage of saved time, for non small RDF stores, is very high, greater than the 22% up to the 97% of saved time. For small stores, Virtuoso can be indexed in shorter time, and thus it could be better to rebuild instead of cloning and versioning.

	Ontologies	**+ street graphs**	**+ smart city Services**	**+Enrich& Reconciliations**	**+Historical data 1 month**
Indexing process					
Final number of triples	15809	33547501	34462930	34557142	44218719
Final number of Files	12	137	178	185	27794
Added triples with respect to previous version	15809	33531692	915429	94212	9661577
Added Files with respect to previous version	12	125	41	7	27609
OWLIM SE 4.3					
Indexing Time without RIM (s)	18	6536	6198	7516	12093
Indexing Time with RIM (s)	11	6029	514	343	5745
% of saved time, RIM versioning	**38,9**	**7,8**	**91,7**	**95,4**	**52,5**
Final Number of triples (including geo + inferred)	16062	57486956	59395432	59486748	73441126
disk space in Mbyte	310	8669	8936	9039	13110
VIRTUOSO 7.2					
Indexing Time without RIM (s)	146	806	964	1000	2487
Indexing Time with RIM (s)	156	833	421	296	1932
% of saved time, RIM versioning	**-6,8**	**-3,3**	**56,3**	**70,4**	**22,3**
Final Number of triples (including geo, no inferred)	21628	35452613	36301322	36420445	46232510
disk space in Mbyte	68	1450	1632	1631	2294
GraphDB 6.1					
Indexing Time without RIM (s)	9	7818	7929	7671	12915
Indexing Time with RIM (s)	2	6791	454	214	4849
% of saved time, RIM versioning	**77,8**	**13,1**	**94,3**	**97,2**	**62,45**
Final Number of triples (including geo + inferred)	15809	57486415	59394891	59487551	73441929

disk space in Mbyte	96	4276	4466	4643	5714

Table 1 – Saving time using Index Manager with respect to rebuilding. Data collected on Ubuntu 64bit, 16 core x 2 Ghz, 500 Gbyte HD

VI. CONCLUSIONS

Graph databases are used in many different applications: smart city, smart cloud, smart education, etc., where large RDF KB store are created with ontologies, static data, historical data and real time data. Most of the RDF stores are endowed of inferential engines that materialize some knowledge as triples during indexing or querying. In these cases, the delete of concepts may imply the removal and change of many triples, especially if the triples are those modeling the ontological part of the knowledge base, or are referred by many other concepts. For these solutions, the graph database versioning feature is not provided at level of the RDF stores tool, and it is quite complex and time consuming to be addressed as black box approach. In most cases, the RDF store rebuilt by indexing is time consuming, and may imply manually edited long scripts that are error prone. In order to solve this kind of problem, in this paper, a lifecycle methodology and our RIM tool for RDF KB store versioning are proposed. The results have shown that saving time up to 95% are possible depending on the number of triples, files and cases to be indexed.

ACKNOWLEDGMENT

The authors would like to thank to the coworkers that have contributed to the experiments in the several projects, and in particular to Km4City: Giacomo Martelli, Mariano Di Claudio. Thanks also to Ontotext for providing a trial version of their tools.

REFERENCES

[Batarseh, Gonzalez, 2013] Batarseh, Feras A., and Avelino J. Gonzalez. "Incremental lifecycle validation of knowledge-based systems through CommonKADS." IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol.43, n.3, 2013, pp.643-654.

[Bellini et al., 2013a] P. Bellini, M. Di Claudio, P. Nesi, N. Rauch, "Tassonomy and Review of Big Data Solutions Navigation", as Chapter 2 in "Big Data Computing", Ed. Rajendra Akerkar, Western Norway Research Institute, Norway, Chapman and Hall/CRC press, ISBN 978-1-46-657837-1, 2013

[Bellini et al., 2013b] P. Bellini, M. Benigni, R. Billero, P. Nesi and N. Rauch, "Km4City Ontology Bulding vs Data Harvesting and Cleaning for Smart-city Services", International Journal of Visual Language and Computing, Elsevier, http://dx.doi.org/10.1016/j.jvlc.2014.10.023, 2013

[Berners-Lee, 2006] T. Berners-Lee, "Linked Data", http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[Bizer et al., 2009] C. Bizer, A. Schultz. "The Berlin SPARQL Benchmark". International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, Pages 1-24, 2009

[Bizer et al., 2011] Bizer, C., Jentzsch, A., Cyganiak, R.: State of the LOD cloud. http://lod-cloud.net/state/ Retrieved July 5, 2014.

[Erling et al., 2015] O. Erling, A. Averbuch, J.L. LarribaPey, Hassan Chafi, Andrey Gubichev, Arnau Prat, Minh-Duc Pham, Peter Boncz, The LDBC Social Network Benchmark: Interactive Workload. Proceedings of SIGMOD 2015, Melbourne.

[Gómez-Pérez, 2004] Gómez-Pérez, A. Ontology Evaluation. Handbook on Ontologies. S. Staab and R. Studer Editors. Springer. International Handbooks on Information Systems. Pp: 251 – 274. 2004.

[Grosan and Abraham, 2011] Grosan, C., and A. Abraham. Intelligent Systems: A Modern Approach, Springer-Verlag, Berlin, 2011.

[Guo et al., 2005] Y. Guo, Z. Pan, and J. Heflin. "Lubm: A benchmark for owl knowledge base systems". J. Web Semantics, 3(2-3):158–182, 2005.

[Hartig et al., 2009] O. Hartig, C. Bizer, J.-C. Freytag. 2009. Executing SPARQL Queries over the Web of Linked Data. In Proc. of ISWC '09, Springer, pp.293-309.

[Isele, Bizer, 2013] R. Isele, C. Bizer. "Active learning of expressive linkage rules using genetic programming". Web Semantics: Science, Services and Agents on the World Wide Web 23 (2013): pp.2-15

[Klein et al., 2002] M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov. "Ontology versioning and change detection on the web". In Procs of the 13th European Conf. on Knowledge Engineering and Knowledge Management (EKAW02), pages 197–212. Springer, 2002.

[Klyne and Carrol, 2004] G. Klyne, J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation", 2004

[Lopez, 1999] M. Fernandez Lopez, "Overview of Methodologies for Building Ontologies", in: IJCAI99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, Stockholm, 1999.

[Milette 2012] L. Milette, Improving the Knowledge-Based Expert System Lifecycle, UNF report, 2012.

[Ngomo, 2011] Ngomo, A. C. N., & Auer, S. Limes-a time-efficient approach for large-scale link discovery on the web of data. integration, 15, 3. (2011).

[Noy and McGuinness, 2001] Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." Technical Report SMI-2001-0880, Standford Medical Informatics. 2001.

[Noy and Musen, 2004] N. F. Noy and M. A. Musen. "Ontology versioning in an ontology management framework". IEEE Intelligent Systems, 19(4):6–13, 2004.

[Rector et al., 2004] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R.,; Wang, H., Wroe, C. "Owl pizzas: Practical experience of teaching owl–dl: Common errors and common patterns". In Proc. of EKAW 2004, pp: 63 – 81. Springer. 2004.

[Schmidt et al., 2009] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel. "Sp2bench: A sparql performance benchmark". In ICDE, pages 222–233, 2009.

[Tzitzikas et al., 2008] Tzitzikas, Yannis; Theoharis, Yannis; Andreou, Dimitris, On Storage Policies for Semantic Web Repositories That Support Versioning, pp.705-719, LNCS 5021 The Semantic Web: Research and Applications, Springer, 2008

[Volkel et al., 2005] M. Volkel, W. Winkler, Y. Sure, S. R. Kruk, and M. Synak. "SemVersion: A Versioning System for RDF and Ontologies". In Procs. of the 2nd European Semantic Web Conf., ESWC'05., Heraklion, Crete, May 29 June 1 2005.

[Zegins et al., 2007] D. Zeginis, Y. Tzitzikas, and V. Christophides. "On the Foundations of Computing Deltas Between RDF Models". In Procs of the 6th Intern. Semantic Web Conf., ISWC/ASWC'07, pages 637–651, Busan, Korea, November 2007.