On the Diversity and Availability of Temporal Information in Linked Open Data

Anisa Rula[1], Matteo Palmonari[1], Andreas Harth[2], Steffen Stadtmüller[2], and Andrea Maurino[1]

[1] University of Milano-Bicocca
{rula,palmonari,maurino}@disco.unimib.it
[2] Karlsruhe Institute of Technology,
{harth,Steffen.Stadtmueller}@kit.edu

Abstract. An increasing amount of data are published and consumed on the Web according to the Linked Data paradigm. In consideration of both publishers and consumers, the temporal dimension of data is important. In this paper we investigate the characterisation and availability of temporal information in Linked Data at large scale. Based on an abstract definition of temporal information we conduct experiments to evaluate the availability of such information using the data from the 2011 Billion Triple Challenge (BTC) dataset. Focusing in particular on the representation of temporal meta-information, i.e., temporal information associated with RDF statements and graphs, we investigate the approaches proposed in the literature, performing both a quantitative and a qualitative analysis and proposing guidelines for data consumers and publishers. Our experiments show that the amount of temporal information available in the LOD cloud is still very small; several different models have been used on different datasets, with a prevalence of approaches based on the annotation of RDF documents.

Keywords: temporal information, temporal annotation, linked data, semantic web

1 Introduction

The management of temporal information has been deeply studied in the field of temporal databases [19] and has found several applications in the World Wide Web [10], e.g., to improve search engines' ranking methods [1]. Most applications have to manage temporal information in order to capture, model, explore, retrieve, and summarize information evolving over time. This is particularly true on the web, where information change often very frequently [4]. Also Linked Open Data[3] (LOD) cannot be assumed to be static, as data are frequently added or removed, and RDF descriptions change over time [25]. The problem of managing

[3] http://richard.cyganiak.de/2007/10/lod/

change in LOD is in fact receiving an increasing attention. A resource versioning mechanism for LOD has been proposed [7], which supports time-series of varying descriptions. Also the crucial problem of maintaining links over evolving datasets has been addressed [23] and an approach to monitor LOD, based on an analysis of changes, has been recently proposed [18]. The capability of capturing, evaluating and managing temporal information in the LOD plays a crucial role when addressing a number of significant problems:

- *Temporal Validity of Statements.* Assessing the validity of statements retrieved at a given time supports a user to be confident that she is consuming true information;
- *Data Fusion and Integration.* Data quality is a key driver to support the fusion of data coming from heterogeneous sources; in particular, temporal information associated with data has to be considered in order to guarantee that the fused data are as much up-to-date as possible[22].
- *Temporal Query Answering and Search.* Users might want to query a knowledge base formulating constraints on temporal aspects associated with data [24, 26]; moreover, temporal information is often used in information retrieval to rank results to a search submitted by the user [1].
- *Temporal Data Exploration.* Timelines associated with data can be used to improve the user experience when accessing information [26, 1];
- *Temporal Entity Matching.* The analysis of temporal information can support entity resolution in some complex scenarios where the values of the attributes considered by a matcher change over time [21].

Availability of temporal data for Linked Data applications, which leverage such temporal information, is needed. As an example, a data fusion approach, which has been demonstrated in a scenario where more DBpedia datasets are considered, measures how much the data are up-to-date by looking at specific temporal metadata [22]. Another example is given by a temporal query engine based on the efficient representation of the temporal validity of statements [24]; this information is assumed to be available, with the extraction of the validity being dependent on the way time is represented in a dataset. If we want to broaden the coverage of the proposed methods handing more datasets, a deeper understanding of the availability and characterisation of temporal information in LOD is needed. In this paper we investigate the characterisation and availability of temporal information in LOD at large scale, by conducting a quantitative and qualitative analysis. To the best of our knowledge, such a systematic analysis is missing, despite the proposal of several approaches to model and query temporal information in RDF [12, 5, 26], support versioning for LOD [23], and monitor changes in LOD [25, 18] (to overcome the problem of an incomplete change history, only a selected set of documents have been actively surveyed in [25]). Based on a clarification of the concept of temporal information, we identify a specific kind of temporal information, called temporal meta-information in the paper. Temporal meta-information is particularly relevant with respect to the application scenario described above because it associates RDF statements and graphs with information about their creation, modification and validity. Since

an analysis of the whole LOD cloud is probably unachievable, we use the large Billion Triple Challenge (BTC) dataset for our investigation. In particular, we focus on the characterisation and availability of temporal meta-information, reviewing the models proposed in the literature for modelling such information and analysing their usage in the BTC dataset. Based on the obtained results we believe that the availability of temporal information is still very scarce in the LOD cloud, thus preventing the development of effective solutions leveraging temporal information at large scale. Moreover, we found that none of the models proposed to manage temporal information has been widely adopted, although temporal annotations of documents seem to prevail so far. Finally, we provide some guidelines to data consumers and data publishers in order to take advantage of the representation approaches proposed so far.

The paper is organised as follows: Section 2 introduces the preliminary definitions we adopted in this paper; in Section 3 we introduce the notion of temporal information and we investigate their availability in the BTC dataset, analysing the temporal properties adopted and the pay-level-domain they occur in. In Section 4 we review the approaches proposed in the literature for the representation of temporal meta-information, discuss their adoption in well-known datasets, and in Section 5 we conduct experiments to quantitatively investigate the adoption of these models in the LOD using the BTC dataset and we discuss our findings. Section 6 we draw the conclusions.

2 Preliminaries

RDF triples and RDF graphs. Given an infinite set $\mathcal{U}$ of URIs (resource identifiers), an infinite set $\mathcal{B}$ of blank nodes, and an infinite set $\mathcal{L}$ of literals, a triple $\langle s, p, o\rangle \in (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is called an RDF triple; s, p, o are called, respectively, the subject, the predicate and the object of the triple. An RDF graph G is a set of RDF triples. A named graph is a pair $\langle G, u\rangle$, where G is a graph and $u \in \mathcal{U}$. RDF data are often stored using the N-quad format; a quad is a quadruple $\langle s, p, o, c\rangle$ where c defines the context of an RDF triple $\langle s, p, o\rangle$; the context describes provenance of a triple, often represented by - but not limited to - an RDF graph. An RDF triple (or simply triple in the following) is also called *statement*. Statements and graphs will be called also *truth-valuable RDF elements*, as they can be both associated with a truth value, under an interpretation function [11].

Temporal entities. We distinguish two types of temporal entities used for representing temporal information in RDF data: *time points*, represented by a single variable t, and *time intervals* that are are closed intervals defined by endpoints, abstractly represented with the standard notation $[t^b; t^e]$, where t^b and t^e represent the time points respectively beginning and ending the interval with $t^b \leq t^e$ (in this paper we do not consider representations of time, where interval are not bound by time points).

Concrete Representation of Time Points. According to a well-accepted best practice, time points are usually represented on the Web by means of *date for-*

mats. RFC 2616 defines three different date formats that are used in the HTTP protocol[4]. The first *datetime* format, e.g., `Sun, 07 Sep 2007 08:49:37 GMT`, is defined by the standard RFC 822 [6] and is the most preferred. The second *datetime* format, e.g., `Sunday, 07-Sep-07 08:49:37 GMT`, is defined by the standard RFC 850 [16] and differs from RFC 822 by using a complete name for the weekday and two digit year. The third datetime format, e.g., `Sun Sep 7 08:49:37 2007`, is defined by ANSI C's *asctime* format and differs from RFC 822 by using a different order. ISO 8601 defines a numerical date format [17]; an example of date according to this format is 2007-09-07T08:49:37.sZ. Based on this standard, dates can be also modelled as primitive datatypes in XML Schema [9]. The primitive types, `date, dateTime, gYearMonth, gYear, gMonthDay, gDay` and `gMonth` defined by these specifications are usually used in RDF data. An alternative representation of time for Linked Data has been proposed in [5] where URIs are used to provide temporal entities. Such formalization has been encoded in OWL time ontology [13].

RDF statements and documents. Some of the URIs that occur in RDF statements identify resources that are, in fact, documents (e.g., XML documents, PDF documents, or HTML pages). For the purpose of this paper it is relevant to distinguish between *generic documents* and documents publishing RDF data, called *RDF documents* in the following; like other generic documents, RDF documents can be described by RDF descriptions; however, differently from other documents, they contain also statements or graphs, which include truth-valuable RDF elements (statements and graphs). In other words, a description about a RDF document can provide a meta-description about the content of the RDF document[5].

3 Temporal Information and Temporal Properties

In this section, we first propose an abstract definition of temporal information, introducing the concept of temporal meta-information. Then we provide an analysis of temporal information in LOD, focusing on the availability of such information and on the temporal properties that occur more frequently.

Temporal information. At the abstract level a *temporal information* can be described as a ternary relation $T(x, a, t)$, where x is a resource, a statement, or a graph, a is a predicate symbol, and t is a temporal entity. We call *temporal property* any property symbol used in a temporal information. Since a temporal information $T(x, a, t)$ can be also seen as a temporal annotation for the element x, the terms temporal information and temporal annotation will used as interchangeable, depending on the context.

[4] `http://www.ietf.org/rfc/rfc2616.txt`

[5] Although documents publishing RDF data encompass RDF graphs, and other Web documents containing RDF-alike data (e.g. in RDFa), in this paper we consider only RDF documents consisting of RDF graphs, which are the ones described in more structured datasets (as the ones we will consider in our investigation).

The different approaches proposed to concretely represent temporal information will be discussed in the next section. Here we only notice that the definition is agnostic with respect to the type of entity (e.g., truth valued vs non truth valued) which temporal information are associated with. Therefore we introduce a new concept to specifically refer to temporal information associated with truth-valuable elements: a temporal information $T(x, a, t)$ is a *temporal meta-information* if and only if x is a truth-valuable element. The concept of meta-temporal information, which is defined according to semantic criteria, allows to distinguish when temporal information is associated with objects in a domain of interest (e.g. the birth date of a person, but also the creation date of a PDF document) or with truth-valuable RDF elements (e.g, the temporal validity of statement, or the last update of a RDF document).

3.1 Dataset and Experimental Setup

To give more insights about the usage of temporal representation approaches in Linked Data cloud, we analyse the latest release of the Billion Triple Challenge[6] (BTC) dataset which was crawled from the Web in May/June 2011 using a random sample of URIs from the BTC 2010 dataset as seed URIs. The BTC corpus contains over 2.1 bn statements in N-Quads[7] format with over 47 k unique predicates, collected from 7.4 m RDF documents. A crawling-based approach is per design biased towards datasets that are well-interlinked, while more isolated datasets are less likely to be found. We expect this aspect to not have any negative effects on the findings of our analysis, which targets specifically prominent and well interlinked part of the LOD cloud. Considering the size of the corpus, we use Apache Hadoop[8] to analyse the data. Hadoop allows for the parallel and distributed processing of large datasets across clusters of computers. We run the analysis on the KIT OpenCirrus[9] Hadoop cluster. OpenCirrus is a collaboration of several organisations to provide an open cloud-computing research test bed designed to support research. For our analysis we used 54 work nodes, each with a 2.27 GHz 4-Core CPU and 100GB RAM, a setup which completes a scan over the entire corpus in about 15 minutes.

3.2 General Analysis

To gather a broad selection of temporal information in BTC, we employ a string-based search method. We assumed that a temporal information if present, it is contained in the object node of a quad. We used regular expressions to identify temporal information in the object node of every quad in the BTC. Although there exist a set of best principles for publishing and interlinking structured data over the Web [2], there are still lacking of applicability of those guidelines [14].

[6] http://km.aifb.kit.edu/projects/btc-2011/
[7] http://sw.deri.org/2008/07/n-quads/
[8] http://hadoop.apache.org/
[9] https://opencirrus.org/

PLD	quad. (m)	Tquad (k)	doc (k)	Tdoc (k)
scinets.org	56.2	3,391	51.9	44.3
legislation.gov.uk	33.1	1,249	246.4	246.4
ontologycentral.com	55.3	1,029	4.6	4.4
bibsonomy.org	34.5	881	234.7	177.3
loc.gov	7.8	854	345.3	302.9
bbc.co.uk	6.3	679	173.5	83.6
livejournal.com	169.8	530	239.2	238.9
rdfize.com	37.6	495	204.7	204.6
data.gov.uk	13.8	479	178.8	91.9
dbpedia.org	28.4	423	596.6	124.1
musicbrainz.org	2.5	359	0.3	0.3
tfri.gov.tw	153.3	272	154.4	78.2
archiplanet.org	16.3	186	79.2	53.5
freebase.com	27.8	173	572.9	109.1
vu.nl	6.8	156	294.2	26.7
fu-berlin.de	5.7	139	291.6	37.4
bio2rdf.org	20.2	129	744.7	71.6
blogspace.com	0.9	124	0.2	0.2
opera.com	24.1	124	160.3	124.1
myexperiment.org	1.5	114	26.1	13.7

Table 1: Top twenty PLDs with respect to temporal quads.

Temporal Property	quad (m)	doc (k)
dcterms:#modified	3.4	44
dcterms:modified	2.3	842
dcterms:date	1.5	247
dc:date	1.4	188
dcterms:created	0.6	450
dcterms:issued	0.2	222
lj:dateCreated	0.2	238
swivt:#creationDate	0.2	197
lj:dateLastUpdated	0.22	225
wiki:Attribute3ANRHP _certification_date	0.18	53
tl:timeline.owl#start	0.17	31
tl:timeline.owl#end	0.15	24
bio:date	0.14	143
po:schedule_date	0.14	15
swrc:ontology#value	0.096	37
cordis:endDate	0.078	0.002
nl:currentLocationDateStart	0.076	26
po:start_of_media_availability	0.074	10
foaf:dateOfBirth	0.068	68
liteco:dateTime	0.062	62

Table 2: Top twenty temporal properties with respect to temporal quads.

In a recent paper Hogan et al. [15] provides metrics to measure the conformance between the application of the principles by the RDF publishers and the ones defined in [2]. Along this line, we noticed that the RDF publishers do not use the date format as defined by the standards such as RFC 822, ISO 8601 or XML Schema Representation. Therefore, we conduct an extraction based on the standard RFC 822 and ISO 8601 with the format pattern `(EEE), dd MMM yy (HH:mm:(ss) (Z|z))` and `yyyy-MM-(dd('T'HH:mm:(ss).(s)(Z|z)))` respectively[10]. The parser is implemented in a class named SimpleDateFormat[11] in Java. The extraction results in 12,863,547 temporal quads and 1,670 unique temporal properties. Furthermore, to provide an overview over the dataset, we extract all pay-level domains (PLDs) present in the BTC with respect to temporal entities associated with generic documents, resources or statements. Herein, we use pay-level domains (PLDs) to distinguish individual data providers [20]. Table 1 lists the 20 PLDs that have the higher number of temporal quads, reporting also their number of quads, documents and temporal documents. We extracted the PLDs from the contexts of the temporal quads and summed the occurrences. We see that although `scinets.org` is listed on top of the PLDs list, it does not provide a high number of temporal information compared to

[10] The value in the parentheses is optional.

[11] `http://docs.oracle.com/javase/6/docs/api/java/text/SimpleDateFormat.html`

its total number of quads. With respect to the temporal quads and documents, we provide the dominance of the PLDs for each category. On one hand, we can notice that the two PLDs containing the highest number of temporal quads over the number of quads are: `musicbrainz.org` and `blogspace.com`; on the other hand, the three PLDs with highest number of published documents having temporal properties over the total number of published document are: `legislation.gov.uk`, `rdfize.com` and `blogspace.com`.

In order to give more insights about the temporal properties we provide in Table 2 a list of the top 20 more frequent temporal properties, reporting the number of quads and documents they occur in. It can be noticed that the Dublin Core (DC)[8] properties are the one occurring with higher frequency. Remarkably, although the two properties `dcterms:#modified` and `dcterms:modified` seem similar, they do not belong to the same vocabulary. The `dcterms:#modified` is defined neither in the DC vocabulary nor in other vocabularies used in LOD, meaning that the property is only a broken version of the second one. Therefore, we perform a further analysis to check which dataset was responsible for this misleading practice, obtaining the whole number of its occurrences. The problem can be considered a data quality assessment problem. Along this line, the problem can be related either to the RDF publisher that has made an inappropriate usage of the property or to the application that consumes the data and provide an inaccurate data. In order to have a better understanding of the temporal properties usage, we also provide in Figure 1 a distribution of the top 10 temporal properties over the most relevant PLDs they occur in. As shown in Figure 1, we find that `dcterms:#modified` is 100% used in `scinets.org`.

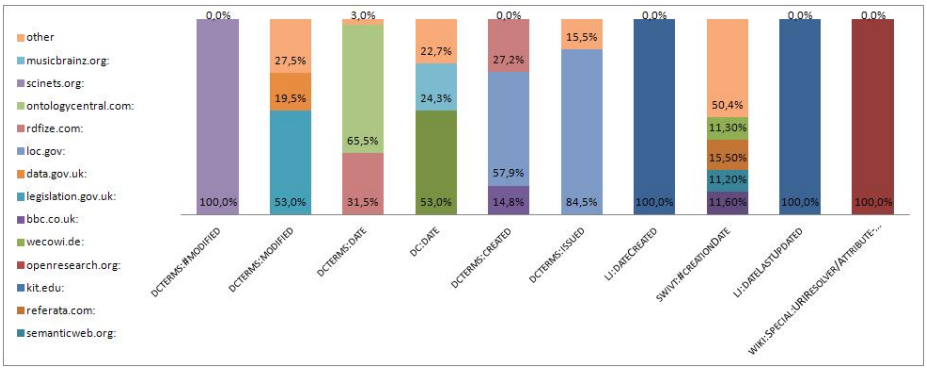

Fig. 1: Distribution of top ten temporal properties with respect to main PLDs.

4 Temporal Meta-information Description Models

In this section we focus on *temporal meta-information*. The peculiarity of this information, with respect to temporal information defined by a relation $T(x, a, t)$,

is that the element x can be a statement, or a graph. Because of the tight constraints given by the triple-based structure of RDF descriptions, the concrete RDF-based representation of an even simple temporal annotation like $T(x, a, t)$, with x being a document and $t = [t^b; t^e]$ a temporal interval, requires some sophisticate mechanisms. Several approaches for providing a concrete representation of a temporal annotation have been proposed. We identify three core perspectives adopted for the concrete representation of temporal meta-information in RDF:

- Document-centric perspective, where time entities, usually points, are associated with RDF documents.
- Fact-centric perspective, where time entities (points or intervals usually intervals are associated with facts; since facts can be represented by one or more statements, we further distinguish the fact-centric perspective into:
  - Sentence-centric perspective, which associates statements with time entities (points or intervals).
  - Relationship-centric perspective, which encapsulates time entities (points or intervals) into objects representing n-ary relations.

In the following we explain in detail the approaches proposed according to the above core perspectives.

4.1 Document-centric perspective

Graphs, i.e. RDF documents, can be associated with temporal-meta information following two approaches: the first one uses HTTP-metadata, and in particular the Last-modified field of an HTTP response header; the second one expresses temporal meta-information using RDF statements with temporal properties taken from available vocabularies such as Dublin Core. Temporal meta-information following these approaches, and in particular, Last-modified and ETage properties of HTTP headers have been used for the detection of changes in Web documents publishing RDF data [25].

Protocol-based representation A protocol-based representation adopts point-based time modelling; the temporal meta-information is not persistently associated with a Web document but can be extracted from the HTTP header returned in response to an HTTP GET request for the document. The temporal meta-information associates a time point, represented by a date, to a Web document G using a predicate a defined in the HTTP protocol according to the schema defined in Listing 1.1.

Listing 1.1: Protocol-based representation

```
HTTP Request
GET G

HTTP Response Header
Status: HTTP/1.1 200 OK
a: t
```

Metadata-based representation Let u_G be a named graph, a_G be a temporal property, t a time point, the metadata-based representation associate a temporal meta-information to an RDF document as shown in Listing 1.2.

Listing 1.2: Metadata-based representation

```
s p o u_G .
u_G a_G t u_G .
```

This approach exploits the concept of named graphs [3].

4.2 Fact-centric perspective

The Fact-centric perspective is identified originally by the temporal RDF model described by Gutierrez et al. [12] which allows the labelling of RDF statements with time intervals, thus expressing the temporal validity of statements. The labelling model adopts a point-based, discrete and linearly ordered temporal domain.

Temporal RDF-based representation Let $\langle s, p, o\rangle$ be an RDF statement and $[t^b; t^e]$ a time interval with a starting point t^b and an ending point t^e, a Temporal RDF-based representation is a temporal annotated statement having the form $\langle s, p, o\rangle[t^b{:}t^e]$.

The above definition requires an extension to the RDF triples and thus can not be used in the basic RDF data model. In other words, the temporal facts are encoded using binary relations, and the fact-based approach requires the addition of one or more temporal attributes to each binary relation. To be able to make statements about statements in the RDF data model, we need to introduce approaches that are able to implement the time dimension. In this section, we present two subcategories of the Fact-centric perspective that approach differently the attachment of time to triples: the Sentence-centric perspective and the Relationship-centric perspective.

Sentence-centric perspective

In the Sentence-centric perspective we identify the following two approaches:

Reification-based representation Let $\langle s, p, o\rangle$ be a statement, s^{st} a modelling of a statement as a resource and a^b_S, a^e_S be two temporal properties with respect to the beginning and the ending point of a time interval $[t^b{:}t^e]$,

a Reification-based representation adds one resource s^{st} and four statements. s^{st} is annotated with time interval as defined in Listing 1.3.

Listing 1.3: Reification-based representation

```
s^st rdf:type rdf:Statement .
s^st rdf:subject s .
s^st rdf:predicate p .
s^st rdf:object o .
s^st a_S^b t^b .
s^st a_S^e t^e .
```

Notice that a property a_S can have a time point or a time interval as property value.

Whenever we want to add a temporal meta-information to an RDF triple, reification is required. However, an alternative approach defined in [24] exploits the RDF named graphs [3] by adding time interval to them, thus creating temporal graphs. The approach associate temporal validity to the statements contained in a default graph by distributing them to the appropriate temporal graphs. Information about the temporal validity of the triple is asserted in the default graph.

Applied Temporal RDF-based representation Let u_{TG} be a temporal graph, a_S^b and a_S^e be two temporal properties with respect to the beginning and the ending point of a time interval $[t^b{:}t^e]$, $\langle s, p, o\rangle$ be a statement in the default graph u_G, the applied temporal RDF-based representation gives a temporal meta-information to the statement by adding it to a temporal graph as shown in Listing 1.4.

Listing 1.4: Temporal RDF-based representation

```
u_TG a_S^b t^b u_TG .
u_TG a_S^e t^e u_TG .
s p o u_TG .
```

Relationship-centric perspective

The N-ary relationship approach[12] expresses relations with arity greater than two as a class rather than as a property. The approach embeds time as an additional attribute, thus describing an instance of the relation or differently saying describing an instance of the class. Remarkably, this approach supports the representation of temporal meta-information as well as other types of information; in fact the validity of fact is embedded into the attributes of objects that represent the relations, i.e., into domain descriptions.

N-ary-relationship-based representation Let $\langle s, p, o\rangle$ be an RDF statement, a_R^b and a_R^e be two temporal properties with respect to the beginning and

[12] http://www.w3.org/TR/swbp-n-aryRelations/

the ending point of a time interval $[t^b{:}t^e]$, o^p be the instance of the class property and $\overline{p}$ be the old property with the new object, an N-ary-base temporal meta-information apply a temporal meta-information to a statement as shown in Listening 1.5

Listing 1.5: Notation for N-ary-relationship-based representation

```
s p̄ o^p .
o^p p o .
o^p a^b_R t^b .
o^p a^e_R t^e .
```

A second temporal model proposed in [27] is based on fluents and timeslices which represent the perdurantist view of the world. Fluents are properties that hold at a specific moment in time, i.e., object properties that change over time. The properties have timeslices as domain and range. Timeslices stands for entities that are extended through temporal dimensions.

4D-fluents-based representation Let $\langle s, p, o\rangle$ be an RDF statement, a^b_R and a^e_R be two temporal properties with respect to the beginning and the ending point of a time interval $[t^b{:}t^e]$, s^t and o^t be the timeslices of s and o, 4D-fluents-based representation associates a temporal meta-information to a relation as shown in in Listing 1.6

Listing 1.6: 4D-fluents-based representation

```
s^t rdf:type :TimeSlice .
s :hasTimeslice s^t .
s^t a^b_R t^b .
s^t a^e_R t^e .
o^t rdf:type :TimeSlice .
o :hasTimeslice o^t .
o^t a^b_R t^b .
o^t a^e_R t^e .
s^t p o^t .
```

5 A Quantitative and Qualitative Analysis of Temporal Meta-information in LOD

In this section we provide analysis for each of the perspectives described in the previous section, with the aim of evaluating the adoption of the heterogeneous approaches proposed to represent temporal meta-information. Along this line, the quantitative analysis is augmented with a qualitative discussion in Section 5.3 that aim to highlight based on the experiments and also on the literature, the advantage and disadvantage of the approaches for each perspective with the aim of addressing also recommendations to the data publishers and data consumers.

5.1 Document-centric perspective

To identify the use of the *Protocol-based representation* we had to ascertain how many of the URIs that identified documents in the BTC returned date information in the HTTP header. However, to conduct an exploratory experiment over the whole amount of documents from the BTC is not feasible, due to the amount of documents (>7m). Therefore, we selected a random sample of 1000 URI due to the fact that for some analysis (in particular on Metadata-, Reification- and N-ary-relationship-based representation) a manual check assessment is needed.

For each document URI in the sample we performed an HTTP lookup to check the last-modification temporal meta-information in the HTTP header response. We found that only 95 out of 1000 URIs returned last-modification temporal meta-information.

To identify the *Metadata-based representation*, we selected a new sample of 1000 URIs that appeared in subject position of quads having temporal information. From this sample we excluded all URIs that make use of a hash symbol # to separate the local name, since we expected these URIs to identify resources- or fact-based perspective than document-based perspective. For the remaining URIs we made an HTTP request and analysed the response code to determine whether the URI identified a generic document or a resource. Following the described approach we found that 432 (43.2%) out of the original 1000 URIs identified documents. Such documents are not limited to RDF ones but they include also html, mp3, xml, pdf documents and so on. As a consequence, we manually checked RDF documents with only the temporal meta-information such as modified and updated which resulted in 51 documents. We further, analysed the 51 RDF documents and discovered that 43 of them are associated with both a protocol-based last-modified temporal meta-information and a metadata-based temporal meta-information. Then for each of the 43 identified documents we compared temporal entities of protocol-based last-modified and metadata-based temporal meta-information. Result of such comparison shows that protocol-based last-modified temporal meta-information is more up-to-date w.r.t. metadata-based temporal meta-information with an average of 364 days.

5.2 Fact-centric perspective

We analysed the *Reification-based representation* in the BTC by looking for how often temporal related quads were used with respect to reified statements. We identified documents containing triples with predicates defined in the RDF reification vocabulary (i.e., `rdf:subject`, `rdf:predicate`, and `rdf:object`). From the identified documents we extracted the properties that express temporal meta-information and use the same subject as reification statements. Reified statements containing temporal meta-information are 2,637 (0.02%).

To account for *Relationship-centric perspective* we again have to use the following approximation, since N-ary relations are hard to be identified just by analysing the data structure. As a consequence we extracted all quads from the BTC using a temporal property and subject, where it is the object of another

quad in the same document. Notice that the possibility to join two triples x and y where $x.subject = y.object$ is a necessary but not sufficient condition, to identify N-ary relations. Therefore, all the temporal triples that are used with N-ary relations in the BTC are contained in a set that we name *scoped set* represented by 7m of temporal quads. From the scoped se, we selected three different random samples of 100 triples and we manually verified if respective documents identified an N-ary relation. Results of such manual analysis show that 10 ± 2 out of 100 triples in the sample are used with an N-ary relation.

Perspective	Approach	Occurrence temp. quads	Occurrence overall quads	Occurrence overall docs
Document-centric	Protocol-based	NA	NA	9.5%
	Metadata-based	43.2%	0.0016%	4.8%
Fact-Centric	Reification-based	0.02%	0.0000008%	0.006%
	N-ary-relationship-based	12.24%	0.0005%	0.6%

Table 3: Temporal information representation approaches and the respective occurrence of i) quads having temporal information; ii) overall quads in the BTC; iii) all documents in the BTC.

5.3 Results and Discussion

Table 3 shows the results of our findings according to the mechanisms adopted in the LOD to annotate data with temporal meta-information.

In the **Document-centric perspective** we identified two approaches used for annotating documents with temporal meta-information: the protocol-based representation and the metadata-based representation. Both approaches are widely used within the document-centric perspective and are more extensively adopted than the Fact-centric perspective. As we hypothesised, the number of temporal meta-information associated with documents is greater than facts. Still, the temporal meta-information used in the metadata-base representation (0.26%), are not high enough compared with the overall number of documents in the sample we used. As mentioned in Section 5.1, the number of available protocol-based last-modified temporal meta-information is smaller than the number of the metadata-based temporal meta-information. However, the temporal meta-information in the protocol-based representation, when available are more up-to-date than the ones in the metadata-based representation. ***Data consumers:*** The applications that consume temporal meta-information will first check for temporal meta-information in the protocol-based representation because they are more up-to-date and in case these information are not available then the applications should be able to check other temporal meta-information in the metadata-based representation. ***RDF data publishers:*** Publishers should carefully update the temporal meta-information whenever the data in the document is changed and both temporal meta-information in protocol- and metadata-based representation should be consistent. *Examples of datasets* providing tem-

poral information according to this perspective are: Protein knowledge base (UNIPROT), legislation.gov.uk).

Differently from the Document-centric perspective the Sentence-centric perspective and the Relationship-centric perspective associate validity time entities to facts with a more fine grained and accurate specification of temporal information. Although the Sentence-centric perspective is relevant for attaching time validity to triples, both the approaches considered in this perspective have limited usage. First, the use of the **Reification-based representation** show a high complexity w.r.t. query processing [15]. Based on the results given in Table 3, this approach appears only in a very small number of quads. ***Data consumers:*** Consumers should be able to evaluate based on the application scenario (e.g., based on the expected types of queries) if it is the case to either build their applications over such representation or to choose a different, and more efficient approach (e.g. Applied temporal RDF-based representation). ***RDF data publishers:*** Publishers should be aware that the Linked Data principles discourage their usage since reification-based representation is considered to be cumbersome for SPARQL query [2], even though they may be useful for representing temporal meta-information. *Examples of datasets*: Timely Yago. Second, the performance of **Applied temporal RDF-based representation** has been reported to have still some efficiency bottleneck [24], especially in the worst case, when the number of graphs (which are associated with temporal annotations) is almost equivalent to the number of triples. ***Data consumers:*** Although we found that the usage of the applied temporal RDF-based representation is relatively uncommon, this approach should deserve more attention because it supports expressive temporal queries based on τ-SPARQL, and can be applied to datasets that provide temporal information according to a Reification-based representation. ***RDF data publishers:*** Publishers should take in consideration the worst case when using the applied temporal RDF-based representation. Therefore, they should use it only when it is possible to group a considerable number of triples into a single graph. *Examples of datasets*: EvOnt (22 million triples).

The **N-ary-relationship-based representation** is very similar to the above approaches with the difference that time is embedded in an object that represent a relation. As Table 3 shows, 0.6% of documents in the BTC contains at least one case of N-ary-relationship-based representation, which is greater than the reification-based representation but still represent a small amount if compared to the overall number of documents. ***Data consumers:*** Consumer applications can extrapolate the temporal validity of facts from representations based on this approach. The lack of a clear distinction between plain temporal information and temporal meta-information provides high flexibility, but at the same makes difficult to predict the kind of temporal information that can be leveraged. ***RDF data publishers:*** Many situations require temporal meta-information associated with relations that can be modelled only as complex objects. Therefore, we recommend to publishers to use N-ary-based representation for complex modelling tasks because it allows flexibility on representing temporal meta-information associated with relation. *Examples of datasets*: Free-

base[13]). The **4D fluents-based representation** supports advanced reasoning functionalities, but, probably also because its complexity, has not been really adopted in LOD, as show in Table 3. *Examples of datasets*: none, to the best our knowledge; it is adopted in the PROTON[14] and DOLCE[15] ontologies).

6 Conclusion

The key contribution of this paper is the investigation of temporal information in LOD, which is important for several research and application domains. As time introduces a further dimension to data it cannot be easily represented in RDF, a language based on binary relations; as a result, several approaches for representing temporal information have been proposed. Based on the qualitative and quantitative analysis using the Billion Triple Challenge 2011 dataset, we came to the conclusion that the availability of temporal information describing the history and the temporal validity of statements and graphs is still very limited. If the representation of temporal validity of RDF data is somewhat more complex and can be expected to be considered in specific contexts, information about the creation and modification of data can be published with quite simple mechanisms. Yet, this information would have great value, e.g., when data coming from different sources need to be integrated and fused.

References

[1] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *1st International Temporal Web Analytics Workshop at WWW*, pages 1–8, 2011.

[2] C. Bizer, R. Cyganiak, and T. Heath. How to publish Linked Data on the Web. linkeddata.org Tutorial, 2008. `http://linkeddata.org/docs/how-to-publish`.

[3] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *14th International Conference on World Wide Web*, pages 613–622, 2005.

[4] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *26th International Conference on Very Large Data Bases*, pages 200–209, 2000.

[5] G. Correndo, M. Salvadores, I. Millard, and N. Shadbolt. Linked timelines: Temporal representation and management in linked data. In *1st International Workshop on Consuming Linked Data at WWW*. 2010.

[6] D. H. Crocker. Standard for the format of ARPA internet text messages. RFC 822, 1982. `http://www.rfc-editor.org/rfc/rfc822.txt`.

[7] H. V. de Sompel, R. Sanderson, M. L. Nelson, L. Balakireva, H. Shankar, and S. Ainsworth. An http-based versioning mechanism for linked data. In *3d Linked Data on the Web Workshop at WWW*, 2010.

[13] `http://www.freebase.com/`

[14] `http://proton.semanticweb.org/`

[15] `http://www.loa.istc.cnr.it/DOLCE.html`

[8] Dublin Core Metadata Initiative. Dublin core metadata element set, version 1.1, 2008. http://www.dublincore.org/documents/dces/.
[9] D. C. Fallside and P. Walmsley. XML schema part 0: Primer second edition. World Wide Web Consortium, 2004. http://www.w3.org/TR/xmlschema-0/.
[10] F. Grandi. Introducing an annotated bibliography on temporal and evolution aspects in the world wide web. *SIGMOD Record*, 33(2):84–86, 2004.
[11] C. Gutierrez, C. Hurtado, A. O. Mendelzon, and J. PÃ©rez. Foundations of semantic web databases. volume 77, pages 520 – 541, 2011.
[12] C. Gutiérrez, C. A. Hurtado, and A. A. Vaisman. Temporal rdf. In *2nd Extended Semantic Web Conference*, pages 93–107, 2005.
[13] J. Hobbs and F. Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, 3(1):66–85, 2004.
[14] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In *3d Linked Data on the Web Workshop at WWW*, 2010.
[15] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *Journal of Web Semantics*, 2012.
[16] M. R. Horton. Standard for interchange of USENET messages. RFC 850, Internet Engineering Task Force, 1983. http://www.ietf.org/rfc/rfc0850.txt.
[17] ISO. Data elements and interchange formats-information interchange-representation of dates and times. ISO 8601, 2004. http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=40874.
[18] T. Käfer, J. Umbrich, A. Hogan, and A. Polleres. Towards a dynamic linked data observatory. In *Linked Data on the Web Workshop at WWW*, 2012.
[19] N. Kline. An update of the temporal database bibliography. *SIGMOD Record*, 22(4):66–80, 1993.
[20] H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov. Irlbot: scaling to 6 billion pages and beyond. In *17th International Conference on World Wide Web*, pages 427–436, 2008.
[21] P. Li, X. L. Dong, A. Maurino, and D. Srivastava. Linking temporal records. *PVLDB*, 4(11), 2011.
[22] P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: Linked Data Quality Assessment and Fusion. In *2nd International Workshop on Linked Web Data Management at EDBT*, 2012.
[23] N. Popitsch and B. Haslhofer. Dsnotify - a solution for event detection and link maintenance in dynamic datasets. *Web Semantic*, 9(3):266–283, 2011.
[24] J. Tappolet and A. Bernstein. Applied temporal rdf: Efficient temporal querying of rdf data with sparql. In *6th Extended Semantic Web Conference*, pages 308–322, 2009.
[25] J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, and S. Decker. Towards dataset dynamics: Change frequency of linked open data sources. In *3d Linked Data on the Web Workshop at WWW*, 2010.
[26] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *13th International Conference on Extending Database Technology*, pages 697–700, 2010.
[27] C. Welty, R. Fikes, and S. Makarios. A reusable ontology for fluents in owl. *Frontiers in Artificial Intelligence and Applications*, 150:226, 2006.