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 (KIT)
{harth,Steffen.Stadtmueller}@kit.edu

Abstract. An increasing amount of data is 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.

1 Introduction

The problem of managing temporal information has been deeply studied in the field of temporal databases [18] and has been more recently addressed in the World Wide Web domain [9,1]. In fact, most data-driven and Web applications need to manage temporal information in order to capture, model, explore, retrieve, and summarize information changing over time. Moreover, the amount of rapidly changing data is likely to grow in the next future with the increasing publication of sensor data, which explicitly represents real-time data of evolving phenomenon over time [19,25,27]. As the information on the Web can change rapidly [4], also Linked Data on the Web[1] cannot be assumed to be static, with RDF statements frequently added to and removed from published datasets [29].

[1] http://lod-cloud.net/

P. Cudré-Mauroux et al. (Eds.): ISWC 2012, Part I, LNCS 7649, pp. 492–507, 2012.

As a consequence, change management and temporal information are receiving an increasing attention in the Linked Data domain. In particular, a number of significant issues have been investigated: a resource versioning mechanism for Linked Data, which allows for publishing time-series of descriptions changing over time [7]; a method to monitor the published datasets, successfully applied to several sources [17]; the maintenance of links over evolving datasets [24].

The capability of managing temporal information plays also a crucial role in several applications and research areas. In *Semantic Data Integration*, temporal information can be used to favor the most up-to-date information when fusing data [22,23]. The analysis of temporal information can also support entity resolution in some complex scenarios where the values of the attributes considered in the matching process change over time [21]. In *Temporal Query Answering and Search*, temporal information can be used to filter out the data of interest given some temporal constraint, or to rank the results of a search engine on a temporal basis. Timelines associated with data can improve the *User Experience* by presenting information in a time-dependent order [30,1].

The capability of designing effective solutions depends on the availability of temporal information and the possibility to collect and process this information across heterogeneous datasets. For example, the modification date associated with RDF documents and extracted via HTTP protocol analysis has been used to fuse data coming from different DBpedia datasets [22]; however, this information is not available in many datasets. Understanding the current status of temporal information published as Linked Data is fundamental for the development of applications able to deal with the dynamism in the data.

In this paper we investigate temporal information published in Linked Data on the Web by analysing its availability and characterisation both from a quantitative and qualitative perspective. To the best of our knowledge, despite the proposal of several approaches to model and query temporal information in RDF [11,5,30,19], support versioning for Linked Data [24], and monitor changes [29,17], a systematic and large scale analysis in this field is still missing. Based on a more precise definition 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 to several application domains because it associates RDF statements and graphs with information about their creation, modification and validity. Since the analysis of the whole LOD cloud is unfeasible, we use the large Billion Triple Challenge[2] (BTC) dataset for our investigation. In particular, we focus on the characterization and availability of temporal meta-information, reviewing the proposed models in the literature for modelling such information and analysing their usage in the BTC.

The analysis of the BTC corpus suggests that the availability of temporal information is still scarce, with negative consequences on the design 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

[2] `http://km.aifb.kit.edu/projects/btc-2011/`

far. Based on the results of our empirical analysis, we provide some guidelines to data publishers and consumers in order to take advantage of the representation approaches proposed so far.

The paper is organized as follows: Section 2 introduces the preliminary definitions we adopt in this paper; in Section 3 we introduce the notion of temporal information and we investigate their availability in the BTC, analysing the more frequent temporal properties 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 and discuss their adoption in well-known datasets. In Section 5 we conduct experiments to quantitatively investigate the adoption of these models in the LOD cloud using the BTC dataset and we discuss our findings. In 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 the 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*. We call statements and graphs also *truth-valuable RDF elements*, as they can be associated with a truth value, under an interpretation function [10].

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^p, and *time intervals*, represented by the standard notation $[t^b; t^e]$, where t^b and t^e represent the time points respectively beginning and ending the interval and $t^b \leq t^e$ (in this paper we do not consider representations of time where intervals are not bound by time points).

Concrete Representation of Time Points on the Web. According to well-accepted best practices, time points are represented on the Web by means of *date formats*. RFC 2616 defines three different date formats that are used in the HTTP protocol[3]. 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 [15]. The third datetime format, e.g., `Sun Sep 7 08:49:37 2007`, is defined by ANSI C's *asctime* format. ISO 8601 defines a numerical date format [16]; 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 [8]. The primitive types, `date`, `dateTime`,

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

`gYearMonth`, `gYear`, `gMonthDay`, `gDay` and `gMonth` defined by these specifications are usually used in RDF data. An alternative representation of time for Linked Data, which denotes temporal entities with URIs and makes use of the OWL Time ontology [12] has also been proposed [5].

RDF statements and documents. Some URIs occurring in RDF statements denote 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, but differently from other documents, they also contain truth-valuable RDF elements (statements and graphs). In other words, a description about an RDF document can provide a meta-description about the content of the RDF document[4].

3 Temporal Information and Temporal Properties

In this section, we first propose an abstract definition of temporal information by introducing the concept of temporal meta-information. Then we analyse the availability of temporal information in Linked Data and the properties that are used more often to represent such information.

- **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 property 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 interpreted as a temporal annotation associated with the element x, the terms temporal information and temporal annotation will be used interchangeably, depending on the context.
- **Temporal meta-information.** We observe that, according to the above definition, truth valuable and non truth valuable RDF entities can be associated with temporal information. Therefore, we introduce a new concept that specifically refers 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 RDF element. The concept of temporal meta-information, which is defined according to semantic criteria, allows distinguishing between temporal information 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) and temporal information associated with truth-valuable RDF elements (e.g, the temporal validity of statement, or the last update of an RDF document).

[4] An increasing number of RDF descriptions are also available in the RDFa syntax from plain HTML and XHTML documents; however, in this paper we focus only descriptions available in RDF/XML documents because the crawled data of the BTC corpus, which we use in our analysis, do not include data extracted from RDFa sources.

3.1 Dataset and Experimental Setup

To give more insights about the usage of temporal information in Linked Data cloud, we analyse the latest release of the 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 which represents only a part of all available Linked Data on the Web, contains over 2.1 bn statements in N-Quads[5] format with over 47 K unique predicates, collected from 7.4 M RDF documents. However, our corpus constitutes a large collection of documents sampled from a wide variety of Linked Data publishers. 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 also observe that the corpus is static, and it samples only RDF/XML, not covering data in other syntaxes like RDFa. We expect these aspects not to have any negative effects on the findings of our analysis, which still targets specifically prominent and well interlinked part of the LOD cloud.

Considering the size of the corpus, we use Apache Hadoop[6] 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[7] Hadoop cluster. 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 which implements a class named SimpleDateFormat[8] in Java. We are confident about the correctness of the collected data because the time parser is well-known and used by a large community.

We assume that if temporal information is present, it is contained in the object position of quads. Thus, we use regular expressions to identify temporal information in the object of every quad in the BTC. However, it has been recently shown that the best practices used to publish data on the Web [3] are not always followed by publishers [13].

We notice that often RDF publishers do not use the date formats defined by standards such as RFC 822, ISO 8601 or XML Schema. In order to collect all temporal information that is represented in the BTC but is not fully compliant to standard date formats, we consider variations of the standards. The variations of the standard date formats are expressed by regular expressions based on the following patterns: `(EEE), dd MMM yy (HH:mm:(ss) (Z|z))` and `yyyy-MM-(dd('T'HH:mm:(ss).(s)(Z|z)))` respectively[9]. We extract

[5] `http://sw.deri.org/2008/07/n-quads/`
[6] `http://hadoop.apache.org/`
[7] `https://opencirrus.org/`
[8] `http://docs.oracle.com/javase/6/docs/api/java/text/SimpleDateFormat.html`
[9] The value in the parentheses is optional.

Table 1. Top twenty PLDs with respect to temporal quads

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 2. Top twenty temporal properties wrt. 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

12,863,547 *temporal quads*, i.e., quads containing a temporal entity, and 1,670 unique temporal properties from the corpus.

Furthermore, to provide a deeper analysis of the distribution of temporal information within the dataset, we extract all the pay-level domains (PLDs) occurring in the context of the quads. Herein, we use PLDs to distinguish individual data providers [20]. Table 1 lists the top 20 PLDs publishing the largest number of temporal quads. For each PLD we report: the total number of quads (*quad.* in Table 1), the number of temporal quads (*Tquad.*), the number of documents (*doc*) and the number of temporal documents (*Tdoc*).

We can notice that although `scinets.org` is listed on top of the list, it does not provide the highest ratio of temporal quads over the total number of quads compared to other datasets. With respect to the temporal quads, we can notice that `musicbrainz.org` and `blogspace.com` represent the largest number of temporal quads as a proportion of all quads. Similarly for the documents, we notice that `legislation.gov.uk`, `rdfize.com` and `blogspace.com` represent the three PLDs with the largest number of temporal documents as a proportion of all documents.

Table 2 lists the top 20 temporal properties that occur more frequently in the BTC, reporting the number of quads and documents they occur in. We also provide an analysis of the distribution of the top-10 most frequent temporal properties within the most significant PLDs, which is plotted in Figure 1. It can

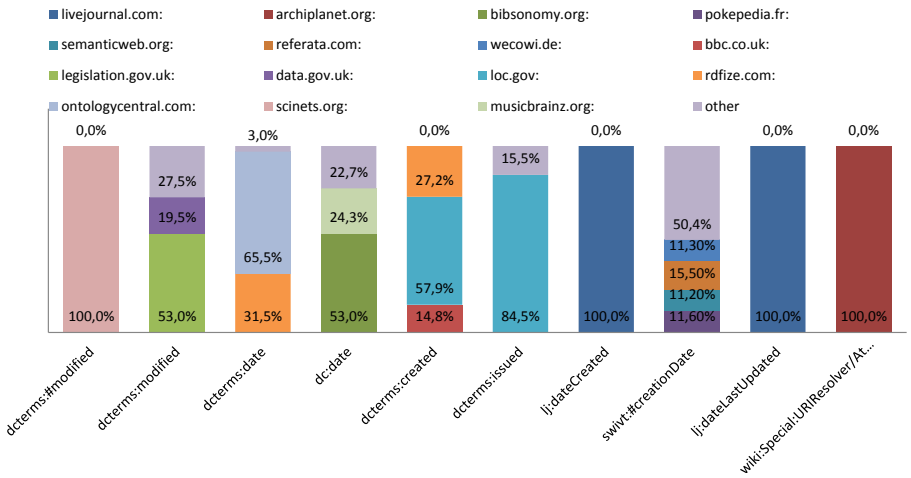

Fig. 1. Distribution of top ten temporal properties with respect to main PLDs

be noticed that not only the properties of the Dublin Core (DC) vocabulary[10] do occur much more frequently than other properties, but they are also used more often across different datasets. Remarkably, the temporal property that occurs more often in the BTC dataset, i.e., `dcterms:#modified`, has a wrong spelling (the correct spelling denotes in fact the second most frequent temporal property in the corpus). As shown in Figure 1, this is also the only temporal property published in the `scinets.org` context, and the spelling is wrong in all the quads having the same context.

4 Temporal Meta-information Description Models

In this section we focus on temporal meta-information, that is temporal information defined as T(x,a,t) where x can be either 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 a temporal entity, requires some sophisticated mechanisms. Several approaches for providing a concrete representation of a temporal annotation have been proposed. We identify three core perspectives that have been adopted for the concrete representation of temporal meta-information in RDF:

- Document-centric Perspective, where time points are associated with RDF documents.
- Fact-centric Perspective, where time points or intervals (usually intervals) are associated with facts; since facts can be represented by one or more statements - we further separate the Fact-centric Perspective into:

[10] `http://www.dublincore.org/documents/dces/`

- Sentence-centric Perspective, which explicitly define the temporal validity of one or more statements annotating them with time points or intervals.
- Relationship-centric Perspective, which encapsulates time points or intervals into objects representing n-ary relations.

In the following we explain in detail the approaches proposed according to the aforementioned 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 the 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 [29].

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, with a Web document G using a predicate a defined in the HTTP protocol according to the schema defined as follows:

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

Metadata-based representation. Let $\langle s, p, o\rangle$ be a statement, u_G a named graph, a_G a temporal property, t^p a time point; the Metadata-based representation associates a temporal meta-information with an RDF document as follows:

$\langle s, p, o, u_G\rangle$
$\langle u_G, a_G, t^p, u_G\rangle$

Examples of datasets providing temporal meta-information to the documents are: Protein knowledge base (UNIPROT) and legislation.gov.uk.

4.2 Fact-Centric Perspective

In the Fact-centric Perspective facts are associated with temporal meta-information that constrain their temporal validity. The first RDF model proposed to formally capture this idea is Temporal RDF [11]. In this model, RDF statements are annotated with time intervals constraining their temporal validity; the

intervals are interpreted over 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 encoding of the above definition into the triple-based RDF data model is not straightforward because RDF can "natively" represents only binary relations. In order to solve this problem, several approaches for encoding the temporal validity of facts into the standard RDF syntax have been proposed. These approaches follow two perspectives that present significant differences: the Sentence-centric Perspective and the Relationship-centric Perspective.

Sentence-Centric Perspective

Two strategies are adopted to represent the temporal validity of fact adopting the Sentence-centric Perspective.

Reification-based representation. Let $\langle s, p, o\rangle$ be a statement, s^{st} an identifier of a statement, a^b_S and a^e_S two temporal properties, and $[t^b{:}t^e]$ a time interval; a Reification-based representation is defined as follows:

$\langle s^{st}, \texttt{rdf:type}, \texttt{rdf:Statement}\rangle$
$\langle s^{st}, \texttt{rdf:subject}, s\rangle$
$\langle s^{st}, \texttt{rdf:predicate}, p\rangle$
$\langle s^{st}, \texttt{rdf:object}, o\rangle$
$\langle s^{st}, a^b_S, t^b\rangle$
$\langle s^{st}, a^e_S, t^e\rangle$

The first four sentences encode the reification of the statement representing the fact using the RDF vocabulary. The temporal properties a^b_S and a^e_S link the statements respectively to the beginning and the ending point of the time interval $[t^b{:}t^e]$ associated with the statement. Notice that a property a_S can have a time point or a time interval as property value. As an example of datasets adopting such approach we mention Timely Yago [30].

In the above approach, every sentence associated with a temporal annotation has to be reified. An alternative approach allows grouping together statements that have the same temporal validity by introducing the concept of *temporal graph* [28]. Temporal graphs are named graphs annotated with timeintervals; each time interval is represented by exactly one temporal graph, where all triples belonging to this graph share the same validity period. Temporal meta-information are collected in a *default graph* which occur as context in the quads.

Applied Temporal RDF-based representation. Let u_{TG} and u_G be the names respectively of a temporal graph and of the default graph, a^b_S and a^e_S two temporal properties, $[t^b{:}t^e]$ a time interval and $\langle s, p, o\rangle$ a statement; the Applied temporal RDF-based representation is defined as follows:

$\langle u_{TG}, a^b_S, t^b, u_G \rangle$
$\langle u_{TG}, a^e_S, t^e, u_G \rangle$
$\langle s, p, o, u_{TG} \rangle$

The temporal properties a^b_S and a^e_S link the temporal graph respectively to the beginning and the ending point of the time interval $[t^b{:}t^e]$. More statements can be associated with the same temporal graph. As an example of dataset that uses such approach is EvOnt [28].

Relationship-Centric Perspective

N-ary Relationship design patterns[11] are introduced to represent RDF relations with arity greater than two. These patterns model an n-ary relation with a set of RDF statements by (i) introducing a specific resource to identify the relation, and (ii) creating links between this resource and the constituents of the relation (resources and literals). These patterns can be used to associate temporal annotations with facts represented by RDF statements to constrain their temporal validity. For example, the fact "Alessandro Del Piero (ADP) plays for Juventus", which is valid within the time interval [1993,2012], can be modelled as a quintuple ⟨ADP, playsFor,Juventus,1993,2012⟩ and represented following the N-ary Relationship pattern. A resource r is introduced to identify the relation and the temporally annotated fact can be represented by the set of RDF statements ⟨ADP,playsFor,r⟩, ⟨r,team,Juventus⟩, ⟨r,from,1993⟩, ⟨r,to,2012⟩. The direction of the links and the strategies adopted for naming the properties can change according to different variants of the pattern [19,25]. However, the temporal annotations are linked to the resources that identify a relation in all the proposed variants. In this paper we define the N-ary Relationship-based representation adopting the variant described in the second use case of the W3C document, the one that occurs more frequently in the BTC corpus.

N-ary-relationship-based representation. Let $\langle s, p, o \rangle$ be an RDF statement, r a new resource, p_1 and p_2 two properties, a^b_R and a^e_R two temporal properties, and $[t^b{:}t^e]$ a time interval; the N-ary-relationship-based representation is defined as follows:

$\langle s, p_1, r \rangle$
$\langle r, p_2, o \rangle$
$\langle r, a^b_R, t^b \rangle$
$\langle r, a^e_R, t^e \rangle$

Although p_1 and p_2 can be two new properties, one of the two is usually equal to p as in the example discussed above. As an example of dataset we mention Freebase[12].

A second approach to model temporal meta-information according to the Fact-centric perspective is based on the concepts of *fluent* and *timeslice* [31]. Fluents

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

are properties that hold at a specific moment in time, i.e., object properties that change over time. The properties representing fluents link two timeslices, i.e., entities that are extended through temporal dimensions.

4D-fluents-based representation. Let $\langle s, p, o\rangle$ be an RDF statement, a_R^b and a_R^e two temporal properties, $[t^b{:}t^e]$ a time interval, and s^t and o^t two timeslices associated respectively with s and o; the 4D-fluents-based representation is defined as follows:

$\langle s^t$, `rdf:type`, `:TimeSlice`$\rangle$
$\langle s$, `:hasTimeslice`, $s^t\rangle$
$\langle s^t, a_R^b, t^b\rangle$
$\langle s^t, a_R^e, t^e\rangle$
$\langle o^t$, `rdf:type`, `:TimeSlice`$\rangle$
$\langle o$, `:hasTimeslice`, $o^t\rangle$
$\langle o^t, a_R^b, t^b\rangle$
$\langle o^t, a_R^e, t^e\rangle$
$\langle s^t, p, o^t\rangle$

Although we could not find any dataset adopting this approach, well-known ontologies like PROTON[13] and DOLCE[14] adopt it.

5 Quantitative and Qualitative Analysis

In this section we analyse and evaluate the adoption of the approaches for representing temporal meta-information. Our quantitative analysis is augmented by a qualitative discussion in Section 5.3, based on both experiments and literature, to highlight the advantages and disadvantages of each approach.

Please observe that some approaches cannot be detected automatically in the data. Therefore, for certain constructs we select a random sample and manually identify the constructs in the sample. We then scale the resulting measure to the entire dataset, which consists of 2.1bn quads in 7.4M documents. Of those, 12.8M were temporal quads (containing a date literal) occurring in 2.5M documents.

Analysing larger samples is infeasible due to the high manual effort involved in checking for constructs in the entire dataset; please note that random sampling is an established method for estimating properties of large populations (e.g., the prediction of election outcomes use small samples and achieve sufficient accuracy [2]). For instance, the error bound for Protocol-based representation is +/- 1.9%. The samples used in the experiments are available online[15].

Not all surveyed approaches are adopted on the web. We did not find any uses of the Applied temporal RDF-based representation and the 4D-fluents-based representation in the data. Table 3 gives an overview of our findings.

[13] `http://proton.semanticweb.org/`
[14] `http://www.loa.istc.cnr.it/DOLCE.html`
[15] `http://people.aifb.kit.edu/sts/data/`

Table 3. Temporal meta-information representation approaches and the respective occurrence compared to i) quads having temporal information; ii) overall quads in the BTC; iii) overall documents in the BTC (n/a = not applicable, - = no occurrence).

Perspective	Approach	Occurrence temp. quads	Occurrence overall quads	Occurrence overall docs
Document	Protocol	n/a	n/a	9.5%
	Metadata	5.1%	0.00019%	0.56%
Fact	Reification	0.02%	0.0000008%	0.006%
	Applied temporal RDF	-	-	-
	N-ary relationship	12.24%	0.0005%	0.6%
	4D-fluents	-	-	-

5.1 Document-Centric Perspective

To identify the use of the *Protocol-based representation* we ascertain how many of the URIs that identified documents in the BTC return date information in the HTTP header. We generate a random sample of 1000 documents (from the context of the quads), and for each document URI in the sample we perform an HTTP lookup to check the last-modified header in the HTTP response. We found that only 95 out of 1000 URIs returned last-modified headers.

To identify the use of the *Metadata-based representation*, we select a sample of 1000 URIs that appear in the subject position of quads with temporal information. We need to ensure that those subject URIs are in fact documents (information resources), as the Metadata-based representation pattern is concerned with documents. Thus, from the sample we exclude URIs containing the # symbol (as URIs with a # per definition do not refer to a document).

For the remaining URIs we send an HTTP request and analyse the response code to determine whether the URI identified a document. We found that 432 (43.2%) identified documents (i.e., directly returned a 200 OK status code). These information resources are not limited to RDF but they also include resources in other formats such as HTML, MP3, XML or PDF. We manually check for RDF documents with only the temporal meta-information such as modified and updated, which resulted in 51 documents.

Of the 51 RDF documents with temporal meta-information in HTTP headers, 43 are also associated with metadata-based dates. Thus, for each of the 43 identified documents we compared protocol-based last-modified and metadata-based last-modified dates. We found that protocol-based last-modified dates are more up-to-date compared to metadata-based dates with an average of almost a year (364 days).

5.2 Fact-Centric Perspective

We analyse the *Reification-based representation* in the BTC by looking for how often reified statements contain temporal information. The pattern first identifies the quads containing predicates that are defined in the RDF reification

vocabulary (i.e., `rdf:subject`, `rdf:predicate`, and `rdf:object`). From the identified cases we extract only those reified statement that have temporal meta-information associated with their subjects. In the entire BTC dataset we found 2,637 reified statements containing temporal meta-information.

To account for *N-ary-relationship-based representation* we again use a combination of sampling of the results of a query over the dataset with manual verification since n-ary relations are impossible to identify just by analysing the graph structure. Hence, we sample and manually identify occurrences.

The following pattern identifies for each document triples of the form $\langle s, p, o \rangle$ and $\langle o, p^*, o^* \rangle$ and furthermore identifies whether o is also associated with a temporal entity. Notice that the possibility to join two triples x and y where $x.object = y.subject$ is a necessary, but not sufficient condition, to identify n-ary relations. All results are contained in a set that we name *scoped set* consisting of 7M temporal quads. Hence, from the scoped set, we select three different random samples of 100 triples and we manually verify if respective documents identify an n-ary relation. Results of such manual analysis show that 10, 10 and 12 out of 100 triples in the samples are used with an n-ary relation.

5.3 Discussion and Recommendations

In the following we discuss the results and provide recommendations for data publishers and consumers.

The approaches that are part of the Document-centric Perspective are more extensively adopted than the approaches of the Fact-centric Perspective. As we hypothesised, the number of temporal meta-information associated with documents is greater than those associated with facts. Still, the use of temporal meta-information for documents (about 10% overall) are not sufficiently high enough to support our outlined use case.

We identify two approaches used for annotating documents with temporal meta-information: the ***Protocol-based representation*** and the ***Metadata-based representation***. We notice that the number of temporal meta-information are much more available in the Protocol-based rather than the Metadata-based representation. The temporal meta-information in the HTTP header, when available, are more up-to-date than the ones in the RDF document itself. *Consumers:* The applications that consume temporal meta-information should first check for temporal meta-information in the Protocol-based representation because they are more up-to-date; in case this information is not available the applications should be able to check in the Metadata-based representation. *Publishers:* Publishers should carefully update the temporal meta-information whenever the data in the document is changed; temporal meta-information in both Protocol- and Metadata-based representation should be consistent.

We identify four approaches used for annotating facts with temporal meta-information, grouped into the Sentence-centric Perspective and the Relationship-centric Perspective. These approaches associate validity expressed as temporal entities to facts.

The use of the ***Reification-based representation*** show a high complexity w.r.t. query processing [14]. The approach appears only in a very small number of quads. *Consumers:* Consumers should be able to evaluate based on the application scenario (e.g., the expected types of queries) if it is possible to either build their applications over such representation or to choose a different, and more efficient approach (e.g. Applied temporal RDF-based representation). *Publishers:* Publishers should be aware that best practices discourage the use of Reification-based representations, as they are cumbersome to use in SPARQL queries [3], even though they may be useful for representing temporal meta-information.

The performance of ***Applied temporal RDF-based representation*** has been reported to have still some efficiency issues [28], especially in the worst case, when the number of graphs (which are associated with temporal annotations) is almost equivalent to the number of triples. *Consumers:* Although we found no usage of the Applied temporal RDF-based representation in the BTC, the 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. *Publishers:* Publishers should take into 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.

The ***N-ary-relationship-based representation*** embeds time in an object that represents a relation. In the BTC, 0.6% of documents contain at least one case of N-ary-relationship-based representation, which is greater than the Reification-based representation but still represents only a small fraction of the overall number of documents. *Consumers:* Consumer applications can evaluate 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 and interpret its meaning. Collecting these temporal meta-information with automatic methods is not straightforward, as shown by the manual efforts required in our analysis to identify this information. *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-relationship-based representation for complex modelling tasks because it allows flexibility on representing temporal meta-information associated with relation.

The ***4D-fluents-based representation*** supports advanced reasoning functionalities, but, probably also because of its complexity, has not been adopted on the Web.

6 Conclusion

The key contribution of this paper is the investigation of temporal information in Linked Data on the Web, which is important for several research and application domains. As time introduces a further dimension to the 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.

As future work, we plan to develop automatic techniques for the assessment of temporal data qualities in Linked Data, such as data currency and timeliness. With the deeper understanding of temporal information gained through our present analysis, we aim to capture and process a large amount of temporal information, overcoming several limitations of preliminary work [26].

Acknowledgements. We thank Basil Ell, Julia Hoxha and Sebastian Rudolph for their valuable comments and acknowledge the support of the EC's Seventh Framework Programme FP7/2007-2013 (PlanetData, Grant 257641).

References

1. Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M.: Temporal Information Retrieval: Challenges and Opportunities. In: 1st Temporal Web Analytics Workshop at WWW, pp. 1–8 (2011)
2. Bartlett, J., Kotrlik, I., Higgins, C.: Organizational Research: Determining Appropriate Sample Size in Survey Research. Information Technology, Learning, and Performance Journal, 43 (2001)
3. Bizer, C., Cyganiak, R., Heath, T.: How to publish Linked Data on the Web. linkeddata.org Tutorial (2008)
4. Cho, J., Garcia-Molina, H.: The Evolution of the Web and Implications for an Incremental Crawler. In: The 26th VLDB, pp. 200–209 (2000)
5. Correndo, G., Salvadores, M., Millard, I., Shadbolt, N.: Linked Timelines: Temporal Representation and Management in Linked Data. In: 1st International Workshop on Consuming Linked Data at ISWC (2010)
6. Crocker, D.H.: Standard for the Format of ARPA Internet Text Messages. RFC 822 (1982)
7. De Sompel, H.V., Sanderson, R., Nelson, M.L., Balakireva, L., Shankar, H., Ainsworth, S.: An HTTP-Based Versioning Mechanism for Linked Data. In: 3rd Linked Data on the Web Workshop at WWW (2010)
8. Fallside, D.C., Walmsley, P.: XML Schema Part 0: Primer Edition, 2nd edn. World Wide Web Consortium (2004)
9. Grandi, F.: Introducing an annotated bibliography on temporal and evolution aspects in the World Wide Web. SIGMOD Record, 84–86 (2004)
10. Gutierrez, C., Hurtado, C., Mendelzon, A.O., Pérez, J.: Foundations of Semantic Web Databases, pp. 520–541 (2011)
11. Gutierrez, C., Hurtado, C.A., Vaisman, A.A.: Temporal RDF. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 93–107. Springer, Heidelberg (2005)

12. Hobbs, J., Pan, F.: An Ontology of Time for the Semantic Web. Processings of the ACM Transactions on Asian Language Information, 66–85 (2004)
13. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the Pedantic Web. In: 3rd Linked Data on the Web Workshop at WWW (2010)
14. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An Empirical Survey of Linked Data Conformance. Web Semantics (2012)
15. Horton, M.R.: Standard for Interchange of USENET Messages. RFC 850, Internet Engineering Task Force (1983)
16. ISO 8601. Data Elements and Interchange Formats-Information Interchange-Representation of Dates and Times (2004)
17. Käfer, T., Umbrich, J., Hogan, A., Polleres, A.: Towards a Dynamic Linked Data Observatory. In: 5th Linked Data on the Web Workshop at WWW (2012)
18. Kline, N.: An Update of the Temporal Database Bibliography. SIGMOD Record, 66–80 (1993)
19. Koubarakis, M., Kyzirakos, K.: Modeling and Querying Metadata in the Semantic Sensor Web: The Model stRDF and the Query Language stSPARQL. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 425–439. Springer, Heidelberg (2010)
20. Lee, H.T., Leonard, D., Wang, X., Loguinov, D.: IRLbot: Scaling to 6 Billion Pages and Beyond. In: The 17th WWW, pp. 427–436 (2008)
21. Li, P., Dong, X.L., Maurino, A., Srivastava, D.: Linking temporal records. The VLDB Endowment (2011)
22. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: Linked Data Quality Assessment and Fusion. In: 2nd International Workshop on Linked Web Data Management at EDBT (2012)
23. Panziera, L., Comerio, M., Palmonari, M., De Paoli, F., Batini, C.: Quality-Driven Extraction, Fusion and Matchmaking of Semantic Web API Descriptions. J. Web Eng. 11(3), 247–268 (2012)
24. Popitsch, N., Haslhofer, B.: DSNotify - A Solution for Event Detection and Link Maintenance in Dynamic Datasets. Web Semantics, 266–283 (2011)
25. Rodrıguez, A., McGrath, R., Liu, Y., Myers, J.: Semantic Management of Streaming Data. In: 2nd International Workshop on Semantic Sensor Networks at ISWC (2009)
26. Rula, A., Palmonari, M., Maurino, A.: Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework. In: 1st International Workshop on Data Quality Management and Semantic Technologies at IEEE ICSC (2012)
27. Sheth, A., Henson, C., Sahoo, S.: Semantic Sensor Web. IEEE Internet Computing 12(4), 78–83 (2008)
28. Tappolet, J., Bernstein, A.: Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 308–322. Springer, Heidelberg (2009)
29. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., Decker, S.: Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: 3rd Linked Data on the Web Workshop at WWW (2010)
30. Wang, Y., Zhu, M., Qu, L., Spaniol, M., Weikum, G.: Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. In: The 13th EDBT, pp. 697–700 (2010)
31. Welty, C., Fikes, R., Makarios, S.: A Reusable Ontology for Fluents in OWL. In: Frontiers in Artificial Intelligence and Applications, p. 226 (2006)