Temporal Knowledge Extraction for Dataset Discovery

Lama Saeeda and Petr Křemen

Czech Technical University in Prague, Czech Republic,
{saeeda.lama, petr.kremen}@fel.cvut.cz

Abstract. Linked data datasets are usually created with different data and metadata quality. This makes the exploration of these datasets a quite difficult task for the users. In this paper, we focus on improving discoverability of datasets based on their temporal characteristics. For this purpose, we identify the typology of temporal knowledge that can be observed inside data. We reuse existing temporal information extraction techniques available, and employ them to create temporal search indices. We present a particular use-case of dataset discovery based on more detailed and completed temporal descriptions for each dataset in the Czech LOD cloud based on the analyzing of the unstructured content in the literals as well as the structured properties, taking into consideration varying data and metadata quality.

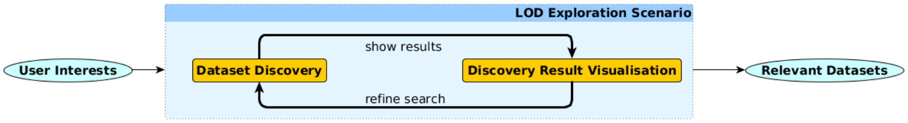
Keywords: temporal information, text processing, linked data, dataset discovery

1 Introduction

Amount of available semantic data in the Linked Open Data (LOD) cloud has increased enormously in the recent years (since 2014 the number of datasets has doubled). However, while the amount of data has been rising, quality of data and metadata is still a serious problem [RMB16]. As a result, the current form of LOD is not ready for direct exploring of the data inside. Datasets in LOD are de-facto black boxes and their content – the actual data – is efficiently hidden during dataset discovery. So, simply speaking – not only data are often hardly interpretable – it is even hard to find them. Currently, real centralized search for which one could offer filters or facets for temporal information does not exist. The same applies for federated search in which e.g. VOiD [BLN11] descriptions could be enriched with a temporal dimension. Thus, possible beneficiaries of our approach involve people exploring LOD (developers, journalists, data analysts, etc.). The aim is to explore datasets which contain potentially relevant data for their applications without the need to search for particular data directly. Figure 1 shows the typical scenario of dataset exploration. Typically, users iterate multiple times through this cycle in order to discover the desired datasets.

To achieve that, many tasks should be performed. Among these, finding temporal context of dataset content and metadata is one of the prominent ones.

Fig. 1. Basic cycle in the exploration process



Temporal context is present in most data across various domains and it is important for both the data and the metadata about data/dataset management. The ultimate goal of dataset exploration is to understand what the dataset is about by minimal information extracted. In the context of this paper, we only consider temporal information.

Varying metadata quality and relevance prevents dataset provision services (CKAN[1] catalogs, etc.) from supporting their users in sufficient search and exploration. For example DCAT [Wor14] start/end dates are often missing or imprecise as we will discuss in the evaluation section. Temporal information is tremendously important in search, question answering, information extraction, data exploration, and more. Techniques for identifying and normalizing temporal expressions work well but they are still poorly used for datasets exploration purposes. The problem is that context is often missing – knowledge evolves over time, facts have associated validity intervals. Therefore, ontologies used for data descriptions should include time as a first-class dimension [WZQ$^+$10].

As structured data found in the LOD typically do not have explicitly defined schema, the notion of dataset summaries emerged in the context of Linked Data. This notion stems from the dataset exploration scenario in which datasets need to be described in order to be discovered. In [BKK16], dataset descriptors have the form of datasets describing other datasets, typically represented as metadata. In this paper, we approach the problem of exploring datasets temporally where we describe the datasets with their temporal information which allow to equip various parts of the dataset with time instants/intervals that the actual content of the dataset speaks about. A temporal value or set of temporal values, which we call the time scope of the statements, is associated with each temporal statement which describes its temporal extent. We compute the temporal coverage of the dataset and formalize it using an integration between two ontologies, the Unified Foundational Ontology (UFO-B) [Gui05] and the OWL Time Ontology [HP04], which we call the **Temporal Descriptor Ontology (TDO)** and finally, we order the temporal statements in the dataset chronologically on a linear timeline to help revealing temporal relationships and reduce users effort while choosing which datasets are aligned with their usage purpose. The main contributions of this work are:

- analysis of temporal dimension of datasets utilizing their actual content;
- systematization of the temporal knowledge extracted with the Temporal Descriptor Ontology (TDO).

[1] `http://ckan.org/`

– provide a baseline formalization to support the representation of the dataset chronologically on a linear timeline for easier dataset exploration process for the future work.

2 State of the Art

Many temporal expression extractors were proposed in literature. **Temporal information extraction systems** usually use rule-based NLP methods, sequence segmentation machine learning algorithms like Hidden Markov Models (HMMs), or Conditional Random Fields (CRF).

[CDJJ14] presented a survey of the existing literature on temporal information retrieval. Also, they categorized the relevant research, described the main contributions, and compared different approaches. A similar survey was presented in [Lim16] which introduced existing methods for temporal information extraction from input texts. Another consideration is the research about temporal information extraction based on the common patterns or knowledge bases. In [KEB07] a supervised machine learning approach for solving temporal information extraction problem (namely CRF) has been adapted. A hybrid system for temporal information extraction from clinical text was proposed in [TWJ$^+$13] using both rule-based and machine learning based approaches to extract events, temporal expressions, and temporal relations from hospital discharge summaries. MedTime [LCB13] is also a system to extract temporal information from clinical narratives. It comprises a cascade of rule-based and machine-learning pattern recognition procedures and achieved a micro-averaged f-measure of 0.88. HeidelTime [SG10], an open source system for temporal expression extraction, is a representative rule-based system that performed well in TempEval2[2] competition. The Stanford temporal tagger (SUTime) [CM12] is one of best currently available temporal taggers with 90.32 F-measure score in TempEval-3 with English Platinum Test set [ULD$^+$13]. SUTime annotations are provided automatically with the StanfordCoreNLP[3] pipeline by including the Named Entity Recognition annotator. SUTime is a rule-based extractor, which means it can normally be configured to use rules specified in external files that are convenient to the data being analyzed.

There is considerable work in natural language processing on **event extraction** in narrative texts, [HKL13] presented a novel automated NLP pipeline for semantic event extraction and annotation (EveSem). In [LCH$^+$13], a method is introduced to effectively identify events by considering the spreading effect of event in the spatio-temporal space. Researchers in [KDF$^+$13] developed and evaluated a system to automatically extract temporal expressions and events from clinical narratives. [DHHN16] address the problem of advanced event and relationship extraction with an event and relationship attribute recognition system. EVITA [SKVP05] is an application for recognizing events in natural language texts. It recognizes events by applying linguistic rules encoded in the

[2] http://www.timeml.org/tempeval2/

[3] nlp.stanford.edu/software/corenlp.shtml

form of patterns. It considers Verb Phrases (NP), Noun Phrases (NP) and Adjectival Phrases (ADJP) as most probable candidates for containing events. It employs different strategies for extracting events from Verbal, Noun and Adjectival Chunks. It make use of POS tagging, lemmatizing, chunking,lexical lookup and contextual parsing to encode event extraction rules.

The markup language TimeML [PCI$^+$03], including the TIMEX3 annotation, has become an ISO standard and the most commonly used guideline for temporal annotations, normalizing the temporal expressions to be convenient to handle programmatically, and resolving relative temporal expressions with respect to a reference date.

Enriching RDF graphs with missing temporal information has also been given attention in publishing historical data as RDF. For instance, authors of [RPNN$^+$14] proposes a generic approach for inserting temporal information to RDF data by gathering time intervals from the Web and knowledge bases. Authors in [MPAGS17] extract temporal timestamps from legacy provenance information. In [BKK16], researchers computed a summary of the used schema terms from the datasets. Then, they enriched the summary by adding time instants and intervals that are used in the dataset. The summary descriptor is reused by other new content-based descriptors, one of which is the temporal descriptor. Researchers focused only on the structured representation of the temporal information to calculate the temporal descriptors.

The Semantic Web provides a suitable environment for representing the temporal data. Web Ontology Language (OWL)[4] through the annotation properties from standard vocabularies, for example, DC[5] gives the ability to incorporate time entities into existing ontologies by representing temporal knowledge and time-based information. Many ontologies were proposed to represent the temporal information in a structured way.

The OWL-Time ontology [HP04] is an OWL-2 DL ontology that provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about durations, and about temporal position including date-time information. Time Intervals[6] is an ontology that extends the OWL-Time ontology. It includes government years and properly handles the transition to the Gregorian calendar within the UK.

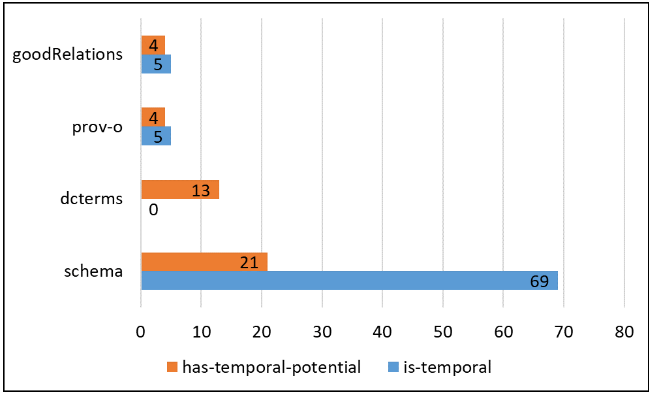
Regarding the temporal information understanding in the current state of the art, there are missing connection between the existing techniques of extraction and the reality of the data in the LOD. Data in LOD use particular formats and knowledge. Extracting the temporal information should not only be about applying NLP techniques but also taking into consideration the RDF knowledge,the connections and relations that already exist, and the typology and th structure of the data within LOD. In this work we approach this matter by contextualizing the current extraction techniques and directing them to be aligned with the ontology.

[4] https://www.w3.org/TR/owl-ref/

[5] http://purl.org/dc/terms/

[6] http://reference.data.gov.uk/def/intervals

Fig. 2. Temporal data-properties in the vocabularies



3 Temporal data analysis

3.1 Temporal data analysis in the Linked Open Vocabularies (LOV)

Vocabularies provide the semantic glue enabling Data to become meaningful Data. Linked Open Vocabularies[7] (LOV) offers several hundred of such vocabularies[8] frequently used in LOD, together with their mutual interconnections. A vocabulary in LOV gathers definitions of a set of classes and properties.

To offer a general solution, extensible towards most properties used across the LOD, we experimented with the most commonly used vocabularies in the datasets according to LOV focusing mainly on the vocabularies that are reused in the Czech Linked Open Data. In order to perform the temporal analysis and to distinguish the temporal properties describing the actual content of the datasets, two annotation properties were created:

is-temporal denotes data properties with explicitly temporal data type expressing time or date, for example but not limited to `xsd:date`, `xsd:time`, `xsd:gYear`, etc..

has-temporal-potential to indicate data properties that potentially contain temporal knowledge in an unstructured form, usually expressed using natural language, for example but not limited to `dcterms:title`[9], `dcterms:description`[10], etc..

[7] http://lov.okfn.org/dataset/lov

[8] currently 603 vocabularies, last accessed 27.07.2017.

[9] http://purl.org/dc/terms

[10] http://purl.org/dc/elements/1.1/description

Table 1. Temporal knowledge in the Czech LOD (Cz LOD)

	is-temporal		has-temporal-potential	
	number of properties	number of triples	number of properties	number of triples
dcterms	12	7955431	2	7734043
schema.org	8	512781	4	149339
goodRelations	6	3110226	0	0
prov-o	2	146115	0	0
CzLOD	36	15268994	8	2597668

We experimented with the following vocabularies as they are most frequently reused in the Czech LOD: Schema[11], dcterms[12], prov-o[13], and goodRelations[14]. An ontology combining these vocabularies augmented with *is-temporal* and *has-temporal-potential* annotaion properties is available[15] for further details. The graph in figure 2 shows the overall number of annotated properties that describe temporal knowledge in the selected common vocabularies. We notice that some properties accept multiple data type ranges, *schema:temporalCoverage* for instance, which accepts either DateTime, Text, or URL. Another observation is that *dcterms* doesn't have any explicitly temporal data type. Even when the property is aimed to express only time or specific date, it is modeled to accept general literal[16] ranges.

3.2 Analysis of the nature of temporal data in Czech LOD

Czech Linked Open Data cloud (Czech LOD)[17] contains dozens of datasets with approx. 1 billion of records. In this work, as an attempt to investigate the nature of the temporal knowledge, we manually experimented with a test set of 10 datasets found in *Czech Linked Data Cloud* involving datasets published by the *OpenData.cz* initiative, to reveal the nature of the temporal knowledge taking into consideration the variety of the topics and the contexts of the datasets. Exploration of the temporal structure in the graph is done through a set of SPARQL queries. We build the graph out of data properties used within the datasets in a structured temporal form or the temporal knowledge found in the strings. After analyzing the nature of the properties used in each dataset, we recognize, in the same manner as the LOV exploration process, two types

[11] `http://schema.org/`

[12] `http://purl.org/dc/terms/`

[13] `http://www.w3.org/ns/prov\#`

[14] `http://purl.org/goodrelations/v1`

[15] available at `https://kbss.felk.cvut.cz/ontologies/dataset-descriptor/temporal-properties-model.ttl`

[16] `http://www.w3.org/2000/01/rdf-schema\#Literal`

[17] `http://linked.opendata.cz/`

of temporal knowledge, *is-temporal* denotes the properties that represent time explicitly through a temporal data property, and *has-temporal-potential* which denotes the properties that contains temporal knowledge in the form of string literals. The latter type is where we will perform the analysis to reveal the hidden temporal knowledge in the datasets. Table 1 shows the number of triples that use these properties inside the datasets. We notice that several temporal properties in the vocabularies are reused across the cloud.

The data inside the Czech LOD are heterogeneous. Triples contain string literals in English as well as in Czech which makes the available temporal analysis tools not efficient enough. Also, temporal information is varying from one dataset to another. In one dataset, usually only structured temporal information or only unstructured temporal information can be found. However, most of the datasets have both types of temporal information.

Based on the analysis of the temporal information in the datasets, we could recognize several forms. Time is mentioned explicitly or implicitly. *Explicit* time mentions have two possibilities, an *instant*, which is a precise time point and an *interval*, which is a time period with distinct start and end points. However, instant form of temporal information can be understood as an interval with the same start and end time. *Implicit* time information found in the datasets has the form of mentions with a specific common well-known implicit temporal property (Christmas, Mother's Day, Independent Day). This information can still differ according to the spatial information though.

3.3 Applying TIE tools on the unstructured temporal knowledge in Czech LOD datasets

Temporal information extractors (namely *SUTime* and *Heidel Time*) are used in order to extract temporal knowledge from the string literals in the datasets. In the string literals, which are expressed in the natural language form, temporal information does not follow any specific structure. For example, the temporal expression "the 20 century" is expressed in many other variants even within the same dataset. We could find it in the written forms, "Twenty century", "20 c", and also "20th-century". Also, the date usually doesn't follow one rule. For example, dates are mentioned in different formats, "DD MONTH YYYY", "DD MM YY", "MM DD YY", and so on. The granularity of the temporal information found in the literals is varying from the very general information "Century" and going down to describe the fractions in specific "minutes". For example, "during the 20th century" and "7,00 till 8,30 and from 10,30 till 15,30". However, the granularity of time information is poorly represented in the structured time schema properties.

Another observation while analyzing the data is the problem of the relative temporal information. For example, in the temporal mention found in one of the string literals describing some event *last year*, the temporal data here has uncertain implicit information, "Which year". It may be related to the year that this dataset piece of information is speaking about, or it can be referring to the creation date of the dataset itself. This information can only be clarified by understanding the whole context of the dataset and making use of the structured data and the data that was already detected in its graph.

4 Improving temporal knowledge representation of datasets

4.1 Temporal information extraction and event extraction

In [BKK16], researchers describe an extension of the basic summary descriptor that they call a temporal descriptor. It extracts temporal references from the dataset and incorporate them into the s-p-o summary. They used only part of the temporal knowledge extracted from some structured temporal data types. They created a simple temporal descriptor that extracts xsd:date literals from the dataset. A huge amount of other structured as well as unstructured temporal information can be found in the cloud. If this information is not considered, accuracy of temporal representation of the dataset will be reduced.

Though *SUTime* and *Heidel time* tools have many advantages and can detect wide spectrum of temporal knowledge, we observed many lacks applying these tools on the string literals. These lacks stem directly from the nature of the temporal data within the datasets, as well as lacking the understanding of the context around. Also, the data heterogeneity regarding the language used inside the datasets. Samples of these lacks are mentioned in Table 2.

Table 2. Samples of the temporal information extraction lacking

Temporal Data Samples	Timex3 value	Lacking
from 1991 till 2006	1991&2006	the range
during 24 h	No temporal expressions	no detection
in the period of 17th - 20th centuries	20th centuries	the range
since its establishment in 2000 until today	2000	the range
30. September 2002	September 2002	day granularity loss
2010 and further	2010	the range
in the 90s of the 20th c	No temporal expressions	no detection
until 1945	1945	the range

We overcome some of these limits by extending the rules in the rule files of the extraction tools with more definitions to detect the temporal expressions that were not possible to be detected by the default settings of the tools. This way, we were able to increase the recall for the retrieved temporal knowledge. Following, respectively, are samples of the rules that we created to spot expressions like {

05/11/1999},{ *05-11-1999*}, and even the special cases to extract the year from the data like {*253/1992 Sb*}.

$$\{ruleType : "time", pattern : /dd/MM/yyyy/\}$$

$$\{ruleType : "time", pattern : /dd - MM - yyyy/\}$$

$$\{ruleType : "time", pattern : /[0-9]\{1,4\}.?yyyy'Sb'/\}$$

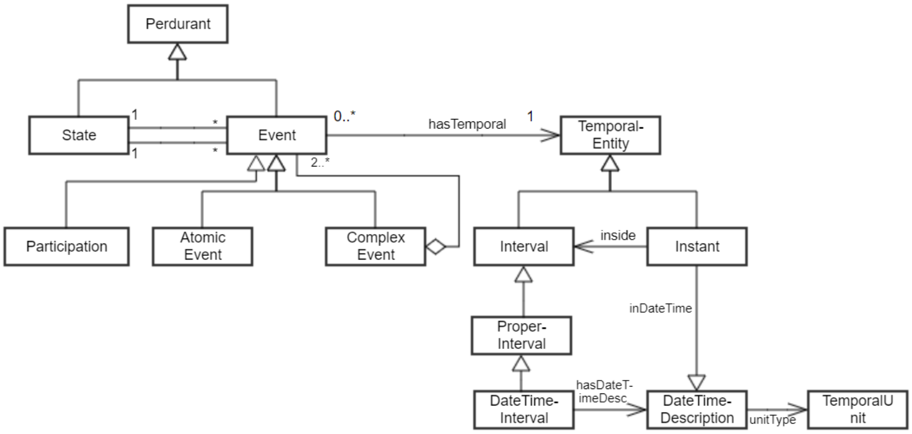
The analysis of the data inside the datasets showed which features are needed to describe the nature of the time information. For example, if this temporal information is connected to a person, it means that it has a relationship, for example (date of birth, date of death, etc.). This information is needed to understand the nature of the temporal knowledge, so not only detecting abstract time intervals and instances is needed, but also the context of this temporal information and what is the meaning of it which is an important information to consider. This approach is designed to provide a better representation and hence better exploration experience of the datasets temporally. To provide the context to the temporal information extracted, we applied a knowledge-driven event extraction from the sentences that contain temporal knowledge following the approach in [SKVP05]. Finally, we formalize this extraction as a pair (e, t), whereas e refers to the event extracted from the string literals and t_e is the corresponding temporal information related to this event.

4.2 Temporal Descriptors Ontology

Based on the analysis in the previous section, in order to describe datasets based on the temporal knowledge and achieve better representation of its temporal dimension, we want to create an integrating ontology on the top of the common vocabularies. To grasp the reality, Temporal Descriptor Ontology (TDO) is built on the top of the Unified Foundational Ontology (UFO) [Gui05], which is one of the top-level ontologies that has a good modeling language and it is supported by several useful tools. UFO presents a level of abstraction that forms a perfect point to start building the TDO ontology. Namely, UFO-B, which is used to model the extracted events whether they are atomic or complex, the participants of the events, and the time span the events occur within. UFO-B suggests that since events happen in time, they are framed by a **Time Interval** [RdAFBG14] but the provided representation of temporal knowledge is very limited.

At this point, we extend the UFO-B with the captured temporal data by the *OWL Time Ontology* and connect it to the events. This provides a generic framework for extracting the temporal information from the datasets. The full definition of OWL-Time Ontology can be found in [HP04]. Here we present those parts that are essential for capturing temporal information found in the datasets. The basic structure of the ontology is based on an algebra of binary relations on intervals developed by Allen and Ferguson [AF97]. Temporal knowledge is represented by the class *TemporalEntity.* This class has only two subclasses, *:Instant* and *:Interval* which axiomatize fundamental intuitions about timepoints

Fig. 3. Temporal Descriptor Ontology

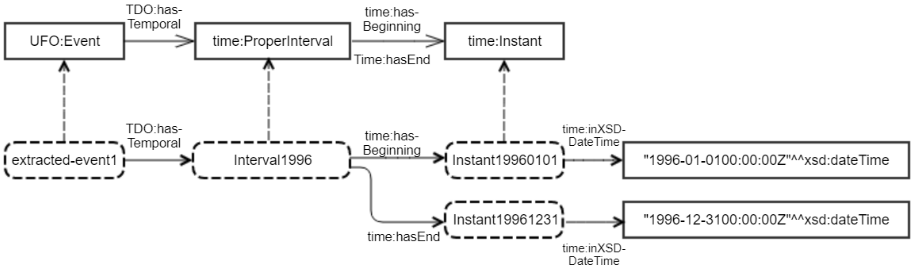


(Instants) and time intervals (Intervals). Core parts of the TDO ontology are shown in fig3.

$$(\forall T)[TemporalEntity(T) \equiv Interval(T) \vee Instant(T)] \tag{1}$$

Time ontology also offers the class *:DateTimeInterval* which may be expressed using a single *DateTimeDescription* class. This class with its related properties are perfect to define a higher level of granularity. For eg. day of the week, or a year, using the *TemporalUnit* class. *ProperInterval* class is used to represent proper intervals, which are intervals whose extremes are different. Example in fig 4 shows the representation of an extracted event and it is temporal interval for the year 1996 using the properties *hasBeginning* and *hasEnd*.

Fig. 4. Temporal Descriptor Ontology



5 Evaluation

Comparison of our approach to DCAT metadata and to the temporal descriptors

DCAT vocabulary provides the property *dct:temporal* to annotate datasets with temporal coverage. According to DCAT specifications, it describes The temporal period that the dataset covers. We used this property, also the temporal descriptors presented in [BKK16] to investigate the quality of the temporal metadata based on comparison with the temporal information extracted from the actual content of the datasets.

Dataset	DCAT startDate	DCAT endDate	TD minDate	TD maxDate	AC minDate	AC maxDate
ds:coi.cz/kontroly	2012-01-01	2014-12-31	2012-01-02	2014-12-31	2012-01-02	2014-12-31
ds:coi.cz/sankce	2012-01-01	2014-12-31	2012-01-01	2014-12-31	2012-01-01	2014-12-31
ds:currency			1999-01-04	2016-04-29	1999-01-04	2016-04-29
ds:drugbank			0002-02-26	2030-12-31	1633-01-01	2030-12-31
ds:legislation/psp.cz			0014-01-01	2099-01-01	0014-01-01	2099-01-01
ds:political-parties-cz			1928-04-12	2013-12-18	1928-04-12	2013-12-18
ds:seznam.gov.cz/agendy			2011-11-16	2030-12-31	1939-01-01	2030-12-31
seznam:objednavky	2014-01-01	2015-02-26	0201-10-07	2020-08-19	2000-01-01	2020-08-19
seznam:plneni			2013-03-26	2015-05-07	2013-03-26	2016-03-31
seznam:smlouvy	2014-01-01	2015-02-26	1992-08-25	2015-05-22	1945-01-01	2019-12-31
ds:vavai/evaluation/2009	2009-01-01	2009-12-31			1034-01-01	2030-12-31
ds:vavai/programmes	2015-03-30	2015-03-30	2003-02-12	2015-01-27	1959-01-01	2022-12-31
ds:vavai/research-plans	2015-03-30	2015-03-30	2003-01-01	2014-06-30	1890-01-01	2014-06-30
ds:vavai/tenders	2015-03-30	2015-03-30	1995-10-02	2016-04-01	1995-10-02	2022-12-31
ds:pravni-vztahy-listiny	2015-02-26	2015-02-26	1993-01-01	2015-07-16	1950-12-31	2015-07-16

Table 3. Comparison of temporal coverage by DCAT metadata, temporal descriptor (TD), and the actual content (AC) temporal representation computed using our approach. *Missing temporal DCAT metadata are indicated by empty cells within 2nd and 3rd column. Empty cells in the 4th and 5th column might indicate either missing data or incomplete descriptor computation procedure. Complete computation of the temporal coverage can be found in the 6th and 7th column.*

We compute the temporal representation of the datasets in the Czech cloud using the approach we presented in the previous section. We compute the temporal scope of the actual content of the triples in the datasets. Table 3[18] shows the overall temporal coverage datasets in the Czech cloud that have both temporal descriptor representation as presented in [BKK16], computed by our approach and compared to the temporal coverage computed by DCAT and the temporal descriptors. Each line represents a dataset in the Czech cloud. Second and third column represent temporal coverage defined by DCAT property dct:temporal[19]. Next two columns represent minimal and maximal computed date by temporal descriptor. Last two columns represent minimal and maximal temporal extraction that we computed from the actual content of the dataset.

[18] Each dataset in the table prefixed with *"ds"* representing URL `http://linked.opendata.cz/resource/dataset/`; Each dataset in the table prefixed with *"seznam"* representing URL `http://linked.opendata.cz/resource/dataset/seznam.gov.cz/rejstriky/`.

[19] `http://purl.org/dc/terms/temporal`

We can notice that for this test pad of 15 datasets, computed temporal representation using our approach is compatible with the temporal descriptors for 46.6% of the cases. For 46.6% of the cases, they differ due to the fact that our approach take the unstructured temporal information into consideration, while the temporal descriptors cares only about the temporal meta-data in the dataset. For the same reason, the dataset *ds:vavai/evaluation/2009* doesn't have any temporal descriptor representation while we are able to compute the temporal coverage for this dataset.

Next, We extend the experiments and we utilize our approach to compute the temporal description of all the datasets available in the SPARQL endpoint of the Czech cloud which contains 76 datasets[20]. We are able to augment the temporal representation for 57.89% of the datasets. The rest of the datasets doesnt have any temporal knowledge in their resources to be extracted. Table 4[21] shows the computed temporal coverage (minimum and maximum date extracted) of each dataset in the cloud compared to DCAT temporal coverage when available.

Dataset	DCAT startDate	DCAT endDate	AC minDate	AC maxDate
ds:ic	2015-01-01	2015-02-26	1972-01-01	2013-12-20
ds:mfcr/ciselniky	2010-01-01	2014-12-31	1900-01-01	9999-12-31
ds:coi.cz/zakazy	2012-01-01	2014-12-31	1986-01-01	2001-01-01
ds:vavai/evaluation/2013	2013-01-01	2013-12-31	1277-01-01	2016-12-31
ds:sukl/drug-prices	2012-01-01	2015-07-29	1990-01-01	2016-04-20
ds:cenia.cz/irz	2015-02-01	2015-03-31	2004-01-01	2012-12-31
ds:vavai/funding-providers	2015-03-30	2015-03-30	1996-12-31	2007-06-01
ds:vavai/evaluation/2011	2011-01-01	2011-12-31	1100-01-01	2050-12-31
ds:vavai/evaluation/2010	2010-01-01	2010-12-31	0900-01-01	2050-12-31
ds:check-actions-law			1945-10-27	2013-12-31
ds:vavai/results	2015-02-26	2015-02-26	1970-01-01	2020-12-31
ds-external:pomocne-ciselniky			1988-01-01	2010-12-31
ds:vavai/evaluation/2012	2012-01-01	2012-12-31	1100-01-01	2050-12-31
ds:vavai/projects	2015-03-30	2015-03-30	1100-01-01	2050-12-31
ds-external:check-actions			1992-01-01	2016-08-01
ds:it/aifa/drug-prices	2012-01-01	2015-07-31	2015-07-15	2015-07-15
ds:/court/cz			2013-06-17	2013-06-17
ds:nci-thesaurus			2013-03-25	2013-05-15
ds:obce-okresy-kraje			2012-09-05	2012-09-05
ds:cpv-2008			2008-01-01	2008-01-01
ds:spc/ai-interactions			2013-05-15	2013-05-15
ds-external:nuts2008/			2008-01-01	2011-12-31
ds-external:geovoc-nuts			2013-01-04	2013-01-04
ds:dataset/fda/spl			2013-15-05	2013-15-05
ds:vavai/cep			1141-01-01	2020-12-31
ds:regions/momc			2014-02-24	2014-02-24
ds:buyer-profiles/contracts/cz			2000-01-01	2024-12-31
ds:legislation/nsoud.cz			2004-01-27	2014-02-27
ds-external:souhrnn-typy-ovm			1969-01-01	2012-12-31

Table 4. The complementary comparison of the temporal coverage by DCAT metadata, to the *actual content* temporal representation computed using our approach for the **rest** of the datasets in the Czech cloud. *Missing temporal DCAT metadata are indicated by empty cells within 2nd and 3rd column.*

[20] last access 2017-08-05

[21] All the dates are normalized to a day granularity.

The dataset **ds:vavai/evaluation/2011** contains data about events that started during *the 12th century* (eg. Pilgrimage element in crusades with Czech participation in the twelfth century). For that the actual data coverage starts at *1100-01-01*. On the other hand, the dataset **ds:vavai/projects** maximum coverage date is *2050-12-31* (Prediction processing of systems utilizing renewable energy sources in Czech republic till 2050).

6 Conclusion and Future Work

In this work, we proposed an approach to improve the temporal knowledge representation of the datasets. We first experimented with the most commonly reused vocabularies in the Linked Open Data Vocabulary (LOV), where we proposed the notions *is-temporal* and *has-temporal-potential*, then we used the same notions to discover the nature of the temporal data in the Czech linked open data cloud as a use case. We proposed to include the temporal knowledge extracted from the textual representation of some properties in describing datasets temporally. Then, Temporal Descriptor Ontology was built on the top of UFO ontology and OWL Time Ontology, utilizing the notions proposed and the temporally annotated properties to provide a better representation for datasets on the temporal dimension. This approach allows representing datasets based on their temporal knowledge time-line. It also improves the metadata quality, allowing the automatic usage of datasets for searching and integrating tasks.

As a future work, we are planning to experiment with more vocabularies in LOV to widen the choices of the temporal properties aligned with the Temporal Descriptor ontology.

Also, timeline **visualization** is an important tool for sensemaking. It allows analysts to examine information in chronological order and to identify temporal patterns and relationships [NXWW14]. Timelines also reduce the chances of missing information and they facilitate spotting anomalies. Currently, we are working on utilizing the computed indices of each dataset aligned with the TDO representation to represent the datasets' triples chronological on a timeline for easy exploration of the datasets temporally. We assume a discrete representation of time with a single day being the atomic granularity. This single granularity can be grouped into larger units like weeks, months, years, decades, and centuries. The base timeline can be defined on any other atomic time interval, such as an hour (of a day), a specific minute, or even seconds. The representation depends on the characteristics of the triples collection in the dataset to be time annotated and do not have effect on the generality of the proposed approach thanks to the TDO representation.

At its core, this visualization is a typical interactive slide-timeline presentation uses single-frame interactivity, meaning that interaction manipulates items within a single-frame without taking the user to new visual scenes. It is augmented by two important features. First, it allows the user to determine the pace of the presentation by using the provided progress bar. Second, it allows the user to interact with the presentation by hovering areas of interest and by using the slider to explore different time windows.

Acknowledgments. This work was partially supported by grants No. GA 16-09713S Efficient Exploration of Linked Data Cloud of the Grant Agency of the Czech Republic and No.SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague.

References

[AF97] James F. Allen and George Ferguson. Actions and events in interval temporal logic. *Spatial and Temporal Reasoning*, page 205245, 1997.

[BKK16] Miroslav Blaško, Bogdan Kostov, and Petr Křemen. Ontology-based Dataset Exploration – A Temporal Ontology Use-Case. In *INTELLIGENT EXPLORATION OF SEMANTIC DATA (IESD 2016)*, Kode, 2016.

[BLN11] Christoph Bohm, Johannes Lorey, and Felix Naumann. Creating voiD descriptions for Web-scale data. *Journal of Web Semantics*, 9(3):339–345, 2011.

[CDJJ14] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of Temporal Information Retrieval and Related Applications. *ACM Comput. Surv.*, 47(2):15:1—-15:41, 2014.

[CM12] Angel X Chang and Christopher D Manning. SUTime: A library for recognizing and normalizing time expressions. *Lrec*, (iii):3735–3740, 2012.

[DHHN16] Peter David, Timothy Hawes, Nichole Hansen, and James J Nolan. Considering context: reliable entity networks through contextual relationship extraction. In *SPIE Defense+ Security*, pages 985107–985107. International Society for Optics and Photonics, 2016.

[Gui05] Giancarlo Guizzardi. *Ontological Foundations for Structural Conceptual Model*, volume 015. 2005.

[HKL13] Siaw Nyuk Hiong, Narayanan Kulathuramaiyer, and Jane Labadin. NATURAL LANGUAGE SEMANTIC EVENT EXTRACTION PIPELINE. (063):333–339, 2013.

[HP04] Jerry R Hobbs and Feng Pan. An Ontology of Time for the Semantic Web. *ACM Transactions on Asian Language Information Processing*, 3(1):66–85, 2004.

[KDF$^+$13] Aleksandar Kova, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. pages 859–866, 2013.

[KEB07] Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. A Supervised Machine Learning Approach for Temporal Information Extraction 3 Conditional Random Field Based Approach. pages 447–454, 2007.

[LCB13] Yu Kai Lin, Hsinchun Chen, and Randall A. Brown. MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, 2013.

[LCH$^+$13] Xuefei Li, Hongyun Cai, Zi Huang, Yang Yang, and Xiaofang Zhou. *Spatio-temporal Event Modeling and Ranking*, pages 361–374. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[Lim16] Chae Gyun Lim. A survey of temporal information extraction and language independent features. *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, pages 447–449, 2016.

[MPAGS17] Albert Meroño-Peñuela, Ashkan Ashkpour, Christophe Guéret, and Stefan Schlobach. Cedar: the dutch historical censuses as linked open data. *Semantic Web*, 8(2):297–310, 2017.

[NXWW14] Phong H Nguyen, Kai Xu, Rick Walker, and BL William Wong. Schemaline: timeline visualization for sensemaking. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 225–233. IEEE, 2014.

[PCI$^+$03] James Pustejovsky, Jose Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34, 2003.

[RdAFBG14] Fabiano Borges Ruy, Ricardo de Almeida Falbo, Monalessa Perini Barcellos, and Giancarlo Guizzardi. An ontological analysis of the iso/iec 24744 metamodel. In *FOIS*, pages 330–343, 2014.

[RMB16] Anisa Rula, Andrea Maurino, and Carlo Batini. *Data Quality Issues in Linked Open Data*, page 87112. Springer International Publishing, Cham, 2016.

[RPNN$^+$14] Anisa Rula, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann. *Hybrid Acquisition of Temporal Scopes for RDF Data*, pages 488–503. Springer International Publishing, Cham, 2014.

[SG10] Jannik Strötgen and Michael Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (July):321–324, 2010.

[SKM11] Jason Switzer, Latifur Khan, and Fahad Bin Muhaya. Subjectivity classification and analysis of the asrs corpus. In *IRI*, pages 160–165. IEEE Systems, Man, and Cybernetics Society, 2011.

[SKVP05] Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. Evita: A robust event recognizer for qa systems. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 700–707, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[TWJ$^+$13] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):828–35, 2013.

[ULD$^+$13] Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. Semeval-2013 task 1: {T}empeval-3: Evaluating time expressions, events, and temporal relations. *Second joint conference on lexical and computational semantics (* SEM)*, 2(SemEval):1–9, 2013.

[Wor14] World Wide Web Consortium. Data Catalog Vocabulary (DCAT). *W3C*, (January), 2014.

[WZQ$^+$10] Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 697–700. ACM, 2010.