From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources

Renata Dividino[1], Thomas Gottron[1], Ansgar Scherp[2], and Gerd Gröner[3]

[1]WeST – Institute for Web Science and Technologies
University of Koblenz-Landau, Germany
{dividino,scherp,gottron}@uni-koblenz.de
[2]Kiel University and Leibniz Information Center for Economics, Kiel, Germany
{ansgar}@informatik.uni-mannheim.de
[3]Paluno – University of Duisburg-Essen, Germany
{gerd.groener}@paluno.uni-due.de

Abstract The Linked Open Data (LOD) cloud changes frequently. Recent approaches focus mainly on quantifying the changes that occur in the LOD cloud by comparing two snapshots of a linked dataset captured at two different points in time. These change metrics are able to measure absolute changes between these two snapshots. However, they cannot determine the dynamics of a dataset over a period of time, i.e., the intensity of how the data evolved in this period. In this paper, we present a general framework to analyze the dynamics of linked datasets within a given time interval. We propose a function to measure the dynamics of a LOD dataset, which is defined as the aggregation of absolute, infinitesimal changes, provided by change metrics. Our method can be parametrized to incorporate and make use of existing change metrics. Furthermore, our framework enables the use of different decay functions within the dynamics computation for different weights on changes depending on when they occurred in the observed time interval. We apply our framework to conduct an investigation on the dynamics of selected LOD datasets. We apply our analysis on a large-scale LOD dataset that is obtained from the LOD cloud by weekly crawls over more than a year. Finally, we discuss the benefits and potential applications of our dynamics function in a real world scenario.

1 Introduction

The Linked Open Data (LOD) cloud is a global information space to structurally represent and connect data. The LOD principles provide a flexible publishing paradigm to integrate and interlink any kind of data from arbitrary datasets, published by various data providers. From the time the Linked Open Data principles have been created until now, the LOD cloud has grown significantly and is a place of continuous changes.

Knowledge about these changes and especially about the change behavior of a dataset over time, i.e., the dynamics of a dataset, is important for many purposes and applications involving Linked Data such as data caching [18], indexing of distributed data sources [13], searching in large graph databases [9], optimizing the execution of queries [14] and recommending appropriate vocabularies to Linked Data engineers [16].

For example, as the data changes, caches and indexes that rely on this data need to be updated since they do no longer reflect the current state of the data anymore. Umbrich et al. [18] proposed a hybrid query execution engine that takes into account the knowledge if a dataset is rather static or dynamic in order to automatically decide whether data is retrieved from caches or from the LOD cloud. Caches are created only for data from static datasets. Therefore, knowledge about the level of dynamics for different data sources is vital to make best use of the resources available for computing caches or performing index updates (e. g., network bandwidth for crawling, computation time).

In related work, changes of LOD sources are analyzed w.r.t. their sets of triples, sets of links, sets of entities, or schema signatures. For example, given two snapshots of a dataset captured at two different points in time, the change analysis at the triple level includes which triples from the previous snapshot have been preserved in the later snapshot, which triples have been deleted, or which ones have been added. For instance, in [6] snapshots of a dataset were analyzed with respect to their set of domain entities, i.e. it was verified if the set of entities described in the datasets has changed. In [7], the authors measure changes with respect to usage of the schema information of a dataset. Käfer et al. [12] quantify changes w.r.t the set of triples, set of links, schema signature. While these kinds of analyses are capable to quantify changes of a dataset captured at two different points in time, they do not really grasp the dynamics of a dataset.

The dynamics of a dataset involves a notion of how "fluid" a dataset is, i. e., how it behaves and evolves over a certain period of time. In the context of this paper, we understand a period of time to be a continuous time interval beginning at an initial point in time up to a final one. Therefore, the dynamics of a dataset involves the analysis of its development over more than two points in time. Due to this time-dependence a measure for dynamics should capture the frequency, degree and regularity of the changes of the data. To the best of our knowledge, there is no established method for measuring the dynamics of LOD datasets.

To fill this gap, we present a formal notion of dynamics for LOD datasets. First, we define dynamics as an aggregation of changes, built on top of contemporary change metrics. Then we extend this notion to incorporate the use of different decay functions for stressing or weakening periods within a time interval. Finally, we analyze the dynamics of different LOD datasets obtained via weekly crawls from the period between May, 2012 and November, 2013. We compute the dynamics of these datasets and compare their change behavior over time. The notion of dynamics has benefits and potential impact in different real world scenarios, which we discuss before concluding the paper.

2 Motivating Scenario

Let us introduce a toy example to illustrate the differences between change and dynamics analysis on LOD datasets. In this example, we describe three snapshots of a dataset captured at three distinct points in time t_1, t_2, and t_3. We are using the FOAF vocabulary for describing persons working at the University of Koblenz-Landau, the University of Kiel, and the University of Duisburg, all of them located in Germany. In addition, we describe relations between persons and their associations to different projects. Besides the FOAF vocabulary, we use domain-specific LOD vocabularies under the domain of uni-

Table 1. Scenario: Example dataset at time t_1.

@prefix	uni-koblenz:	<http://www.uni-koblenz.de/> .
@prefix	uni-duis:	<http://www.uni-duisburg.de/> .
@prefix	rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix	foaf:	<http://xmlns.com/foaf/0.1/> .
uni-duis:GerdGroener	rdf:type	foaf:Person .
uni-duis:GerdGroener	foaf:knows	uni-koblenz:RenataDividino .
uni-koblenz:RenataDividino	foaf:name	"Renata Dividino".
uni-koblenz:RenataDividino	foaf:knows	uni-duis:GerdGroener .
uni-koblenz:ThomasGottron	uni-koblenz:worksFor	uni-koblenz:Robust .
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:RenataDividino.

Table 2. Scenario: Example dataset at time t_2.

@prefix	uni-koblenz:	<http://www.uni-koblenz.de/> .
@prefix	uni-duis:	<http://www.uni-duisburg.de/> .
@prefix	uni-kiel:	<http://www.uni-kiel.de/> .
@prefix	rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix	foaf:	<http://xmlns.com/foaf/0.1/> .
uni-duis:GerdGroener	rdf:type	foaf:Person .
uni-duis:GerdGroener	foaf:knows	uni-koblenz:RenataDividino .
uni-duis:GerdGroener	**foaf:knows**	**uni-kiel:AnsgarScherp.**
uni-kiel:AnsgarScherp	**rdf:type**	**foaf:Person .**
uni-kiel:AnsgarScherp	**foaf:name**	**"Ansgar Scherp" .**
uni-koblenz:RenataDividino	foaf:name	"Renata Dividino".
uni-koblenz:RenataDividino	foaf:knows	uni-duis:GerdGroener .
uni-koblenz:RenataDividino	**foaf:knows**	**uni-koblenz:ThomasGottron.**
uni-koblenz:ThomasGottron	**foaf:mbox**	**mailto:Gottron@uni-koblenz.de.**
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:RenataDividino.
uni-koblenz:ThomasGottron	**foaf:knows**	**uni-kiel:AnsgarScherp.**
uni-koblenz:ThomasGottron	uni-koblenz:worksFor	uni-koblenz:Robust .

Table 3. Scenario: Example dataset at time t_3.

@prefix	uni-koblenz:	<http://www.uni-koblenz.de/> .
@prefix	uni-duis:	<http://www.uni-duisburg.de/> .
@prefix	rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix	foaf:	<http://xmlns.com/foaf/0.1/> .
uni-duis:GerdGroener	rdf:type	foaf:Person .
uni-duis:GerdGroener	foaf:knows	uni-koblenz:RenataDividino .
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:RenataDividino.
uni-koblenz:ThomasGottron	uni-koblenz:worksFor	uni-koblenz:Robust .
uni-koblenz:RenataDividino	foaf:name	"Renata Dividino".
uni-koblenz:RenataDividino	foaf:knows	uni-duis:GerdGroener .

koblenz.de, uni-kiel.de and uni-duis.de for modeling persons and projects. For instance, there are entities like uni-koblenz:ThomasGottron and uni-koblenz:RenataDividino that are connected via a foaf:knows property. Table 1 summarizes the RDF triples published in the first snapshot of the dataset at time t_1.

At time t_2, the same data is visited again. Table 2 shows the RDF triples of this new snapshot. We can directly observe changes in the triples between these two snapshots. In the second snapshot, we observe six new RDF triples (see highlighted triples in Table 2). Table 3 summarizes the RDF triples of the third and last snapshot at time t_3. This snapshot contains the same set of triples as the first one.

Existing metrics proposed in the literature are able to quantify changes for every pair of snapshots of a dataset. For the sake of simplicity, we apply a very simple metric in this example, which only counts the additions, deletions and changes between the set of triples from the first and the second snapshot. In this case, there are six new triples over the total of all triples. The same amount of changes is observed when comparing the second and third snapshot. However, since the first and the third snapshot contain the same set of triples, we cannot observe any changes under the considered metric. The direct comparison of the first and third snapshot of the dataset ignores the changes in the second snapshot.

In this paper, we argue that the consideration of the changes in the second snapshot is of great importance to the analysis of the dataset dynamics in the time period ranging from t_1 to the t_3. Otherwise, when ignoring this evolution the true dynamic character of the dataset is neglected. In the following sections, we systematically introduce metrics of changes and present a formalization of how to incorporate them into our notion of dataset dynamics.

3 LOD Change Analysis

In the literature, many change metrics have been proposed for analyzing RDF data of LOD [12,7,6,8]. These metrics essentially quantify the changes that occurred in a dataset by comparing two snapshots of this dataset. Our goal is to re-use such metrics and to incorporate them as parameter in our framework for measuring dynamics of a LOD dataset (introduced in the subsequent Sec. 4).

We will denote a change metric as a function Δ. Basically, such a Δ-function is a metric that quantifies changes between two datasets, i.e., it is a function that determines the difference (or distance) between two datasets. Without loss of generality, in this paper, we restrict Δ-metrics to determine the difference between two RDF datasets. For instance, changes between two datasets can be measured by the number of differences between the set of triples of these datasets (such as additions and deletions of RDF triples). Please note that our framework for measuring the datasets dynamics can be parametrized to make use of any existing change metrics that satisfies the formal requirements listed below.

Definition 1. *Let $\mathcal{S}$ be the set of all possible RDF datasets. A change metric is a function $\Delta : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ that maps two RDF datasets to a real number and satisfies the following conditions (for X_1, X_2 and X_3 being RDF datasets).*

(i) positivity: $\Delta(X_1, X_2) \geq 0$
(ii) symmetry: $\Delta(X_1, X_2) = \Delta(X_2, X_1)$
(iii) identity of indiscernibles: $\Delta(X_1, X_1) = 0$ *and*
(iv) triangle inequality: $\Delta(X_1, X_3) \leq \Delta(X_1, X_2) + \Delta(X_2, X_3)$

Example 1 (Jaccard distance as change metric). The Jaccard Distance $\Delta_{Jaccard}$ between two RDF sets satisfies the requirements of Definition 1. Let X_1 and X_2 be the two RDF datasets presented in Table 1 and Table 2, then the Jaccard distance between the set of triples of X_1 and X_2 evaluates to:

$$\begin{aligned} \Delta_{Jaccard}(X_1, X_2) =& 1 - \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \\ =& 1 - (13/19) \\ =& 0.32 \end{aligned} \tag{1}$$

4 LOD Dynamics Analysis

In this section, we introduce a formal specification of *dynamics* and the *dynamics function* for LOD datasets.

4.1 Dynamics Function

The dynamics function aims at quantifying the evolution of a dataset over a specific period of time and takes into consideration the changes occurring in this period.

For the sake of simplicity, we model time as a real value. We are looking for a function $\Theta : S \times \mathbb{R} \to \mathbb{R}$, which assigns each dynamic RDF dataset $\mathcal{X}$ consisting of a concrete set of triples $X \in S$ at any point in time t a value which models the quantity of evolution it has undergone[1]. Θ is a monotone, non-negative function. This implies that there cannot be negative evolution. To measure the dynamics as the amount of evolution a dataset exhibited in a given time interval $[t_1, t_2]$, it is sufficient to compute $\Theta(\mathcal{X}, t_2) - \Theta(\mathcal{X}, t_1) \geq 0$. For ease of notation, we will in the following abbreviate X_t for the dataset $\mathcal{X}$ at time t and $\Theta(\mathcal{X}, t)$ by $\Theta(X_t)$.

While it is difficult to define the function Θ directly to provide meaningful values, we will define it indirectly. To this end, we assume that the change rate of a dataset X_t at time t is given by a function $c(X_t)$. Then, we define the difference for two values of the function Θ to be obtained by accumulating the dataset change rate function over a time interval. This means, we integrate the change rate function of a dataset over a given period of time. More formally, the dynamics of a dataset is given by:

$$\Theta(X_{t_2}) - \Theta(X_{t_1}) = \int_{t_1}^{t_2} c(X_t) \mathrm{d}t. \tag{2}$$

[1] This quantity of evolution is an abstract value but can serve for relative comparisons of datasets.

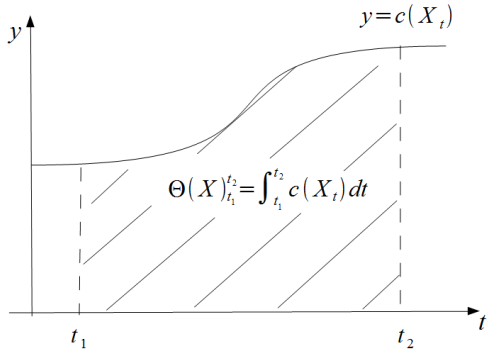


Figure 1. The dynamics of a dataset is obtained by integration over its change rate over time.

The idea of integrating over a change rate function is depicted in Fig. 1 where the area under the curve $c(X_t)$ corresponds to a quantification of a dataset evolution and thus the dynamics of the dataset.

However, also the function c is not explicitly known, and cannot be used for the computation, i.e., it is not possible to determine the change rate of a dataset for a given point in time. Thus, our idea is to use an approximation for $c(X_t)$ based on discrete points in time and the changes between the datasets at these times.

So, we can effectively assume $\mathcal{X} = \{X_{t_1}, \ldots, X_{t_n}\}$ to be a set snapshots of the RDF dataset X at points in time t_i, for $i = 1, \ldots n$. For any two snapshots of a dataset, we can measure changes using a Δ-metric, such as the ones presented in Sec. 3. Our assumption is that for small time intervals (ideally intervals tending towards zero) the change between datasets is a good enough approximation of the change rate. This corresponds to the idea that:

$$\frac{\Delta(X_{t_i}, X_{t_{i-1}})}{t_i - t_{i-1}} \xrightarrow{t_{i-1} \to t_i} c(X_{t_i}) = \frac{\mathrm{d}}{\mathrm{d}t}\Theta(X_{t_i}) \tag{3}$$

Therefore, instead of using the change rate function c, we can use its approximation, namely the Δ-metric of (ideally) small time intervals of each pair X_{t_i} and $X_{t_{i-1}} \in \mathcal{X}$. This corresponds to approximating the change rate function using a step function as depicted in Fig. 2. Computing the integral given in Eq. 2 over this approximated function corresponds to:

$$\Theta(\mathcal{X})_{t_1}^{t_n} \approx \sum_{i=2}^{n} \Delta(X_{t_i}, X_{t_{i-1}}). \tag{4}$$

In the following example, we illustrate the computation of the dynamics of the dataset presented in our motivation scenario.

Example 2 (Computing dynamics based on a Jaccard distance change metric). Let $\mathcal{X} = \{X_{t_1}, X_{t_2}, X_{t_3}\}$ be a dataset, X_{t_1}, X_{t_2}, and X_{t_3} be the snapshots at three distinct points in time presented in Table 1, Table 2, and Table 3, respectively. Then the

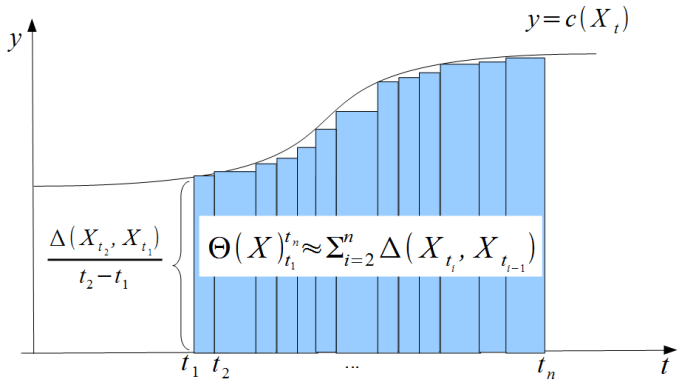


Figure 2. Dataset dynamics defined as aggregation of absolute, infinitesimal Δ-metrics changes.

Jaccard distance between the set of triples from X_{t_1} and X_{t_2}, and from X_{t_2} and X_{t_3} are given as follows:

$$\Delta_{Jaccard}(X_{t_2}, X_{t_1}) = 0.32, \quad \Delta_{Jaccard}(X_{t_3}, X_{t_2}) = 0.32 \tag{5}$$

Then the dynamics of $\mathcal{X}$ is:

$$\begin{aligned} \Theta(\mathcal{X})_{t_1}^{t_n} &= \sum_{i=2}^{n} \Delta_{Jaccard}(X_{t_i}, X_{t_{i-1}}) \\ &= \Delta_{Jaccard}(X_{t_2}, X_{t_1}) + \Delta_{Jaccard}(X_{t_3}, X_{t_2}) \\ &= 0.64 \end{aligned} \tag{6}$$

4.2 Decay Function

So far, we proposed a general framework in which we can compute the dynamics or evolution of any RDF dataset over a period of time and which incorporates any change metric Δ that follows the requirements given in Sec. 3. Applications such as caching and index maintenance benefit from the analysis of the change history of datasets, since update strategies can incorporate the evolution of a dataset in their computation, instead of only the quantification of changes w.r.t. the last two snapshots.

However, for such applications changes tend to be less or more important as time passes, e.g., if a dataset used to change much but does not anymore, its index update strategy may need to be adapted (e.g., it should be less aggressive than it used to be). Therefore, it may be important that changes that took place a longer time ago are weakened and that recent ones are strengthened, or the other way around, i.e., changes should be stressed or weakened relative to how long ago they have happened. For this purpose, the dynamics function should be flexible to incorporate such requirements.

Accordingly, we extend the dynamics function with a decay function $f(t)$. Fig 3 illustrates the influence of the decay function when combined with the dynamics function. In the upper left side of the figure, the dynamics function is shown. In the lower left

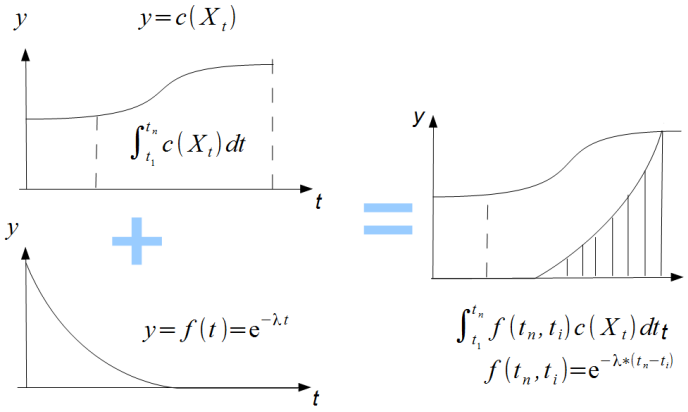


Figure 3. Dynamics function with decay to strengthen the recent changes.

side the decay function (in this example, the exponential decay function) is presented. In the right side, the combination of both is depicted. Please note, that in this example we want to weaken older changes and strengthen the recent ones.

Based on these considerations, we introduce the following modifications to the definition of dynamics: Let $\mathcal{X}$ be a dynamic RDF dataset and $c(X_t)$ be a function which measures the change rate of a dataset at time t, and $f(t)$ be a decay function. Then the decayed dynamics function of $\mathcal{X}$ is defined as:

$$\Theta_{decay}(X_{t_2}) - \Theta_{decay}(X_{t_1}) = \int_{t_1}^{t_2} f(t) \cdot c(X_t) \mathrm{d}t. \tag{7}$$

Consequently, the discretization is given by:

$$\Theta_{decay}(\mathcal{X})_{t_1}^{t_n} \approx \sum_{i=2}^{n} f(t_i) \Delta(X_{t_i}, X_{t_{i-1}}). \tag{8}$$

Example 3 (Dynamics involving a decay function). We continue our Example 2 where we compute the Jaccard distances $\Delta_{Jaccard}(X_{t_2}, X_{t_1})$ and $\Delta_{Jaccard}(X_{t_3}, X_{t_2})$. We want to compute the dynamics of $\mathcal{X}$ using the dynamics with a decay function. In this example, we chose the exponential decay function $f(t_i) = e^{-\lambda \cdot (t_n - t_i)}$ to be our example decay function. For sake of simplicity, we set the parameter λ to 1. Then $\Theta_{decay}(\mathcal{X})$ is:

$$\begin{aligned}
&\sum_{i=2}^{n} f(t_i) \Delta_{Jaccard}(X_{t_i}, X_{t_{i-1}}) \\
=&e^{-\lambda \cdot (t_3 - t_2)} \cdot \Delta_{Jaccard}(X_{t_2}, X_{t_1}) + e^{-\lambda \cdot (t_3 - t_3)} \cdot \Delta_{Jaccard}(X_{t_3}, X_{t_2}) \\
=&0.38 \cdot 0.25 + 1 \cdot 0.25 \\
=&0.345
\end{aligned} \tag{9}$$

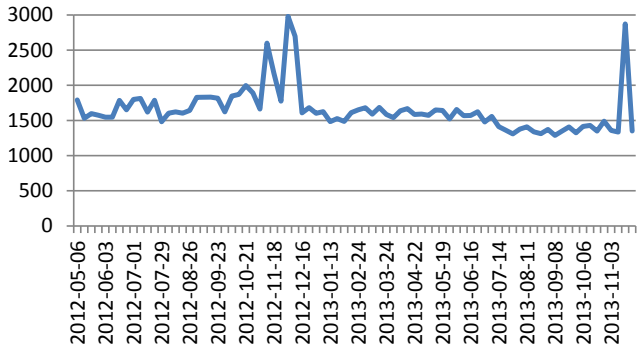


Figure 4. Total number of LOD sources per snapshot

5 Experiments

In this section, we apply our dynamics function to real world LOD data sources. By comparing the evolution of well-known LOD sources to the figures provided by our dynamics function, we illustrate how the proposed function works and relate its results to temporal change patterns. The main goal of the experiments is to show that our approach of quantifying dynamics of LOD sources reflects the broad intuition of historic change events analysis.

For this purpose, we work with data from the Dynamic Linked Data Observatory (DyLDO) [12]. The DyLDO dataset has been created to monitor a fixed set of Linked Data documents (and their neighborhood) on a weekly basis. For sake of consistency, we use only the kernel seeds of LOD documents[2]. Our test dataset is composed of 84 snapshots corresponding to a period of more than one year (from May, 2012 until November, 2013). Furthermore, the DyLDO dataset contains (parts of) various well known and large LOD sources, e.g., dbpedia.com, musicbrainz.com, and bbc.co.uk. We will analyze each of these sources separately by splitting up the DyLDO dataset. For more detailed information about the DyLDO dataset, we refer the reader to [12].

5.1 Analysis of DyLDO Sources

Each snapshot of the DyLDO dataset consists of a set of RDF triples retrieved from different LOD sources. As presented in Fig. 4, the number of LOD sources crawled during this period ranged from 1,287 (Sep. 08, 2013) to 2,980 (Dec. 02, 2012). Given the heterogeneity of the dataset, we expect a wide range of different dynamics behavior, i.e., some of them evolved a lot during this period, others less. Table 4 shows 13 LOD sources from the DyLDO dataset which we selected for our analyses. These sources vary a lot in size, e. g., dbpedia.org is a very large source with more than 4 millions triples per snapshot, and the oreilly.com source is a very small one with around 10.000

[2] The DyLDO crawler uses a deterministic mechanism for labeling blank-nodes such that, for a given document, the labels of blank nodes are consistent.

#	Dataset	Description	Avg. # triples per snapshot	$\Theta(\mathcal{X})$	$\Theta_{decay}(\mathcal{X})$
1	dbpedia.org	DBpedia Project	4,080,910	55.71	23.42
2	neuinfo.org	Neuroscience Information	2,065,028	13.62	5.23
3	linkedct.org	LinkedCT Databrowser	1,782,886	51.75	25.03
4	dbtropes.org	TVTropes.org wrapper	1,729,455	66.96	28.99
5	dbtune.org	DBTune Server	1,427,361	20.90	8.33
6	identi.ca	Open Source social platform	1,341,045	58.45	18.48
7	opera.com	Opera Browser	1,297,630	68.03	30.15
8	ontologycentral.com	Ontology Central	1.146,171	62.63	25.49
9	bnf.fr	Bibliotheque nationale de France	1,181,134	25.06	10.87
10	berkeleybop.org	Berkeley Bioinformatics Projects	903,318	12.82	5.78
11	uriburner.com	Virtuoso LD Middleware	202,529	29.34	13.67
12	iu.edu	Indiana University	112,517	53.23	22.43
13	oreilly.com	O'Reilly Media	17,188	79.35	33.84

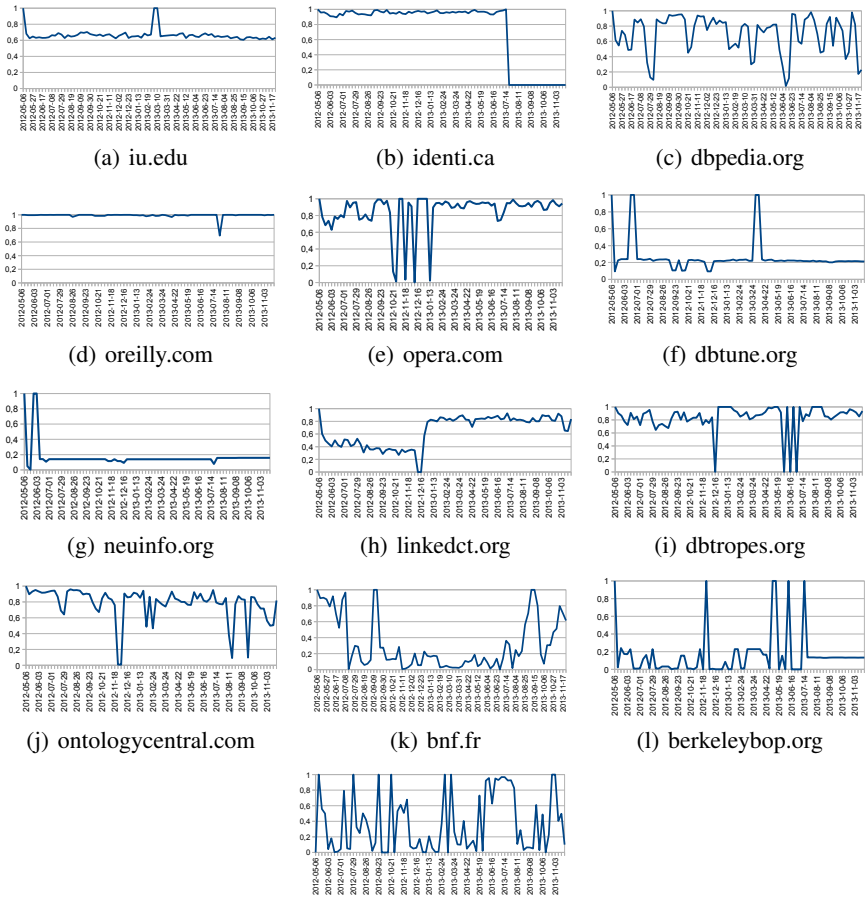
Table 4. LOD sources from the DyDLO dataset

triples per snapshot. Overall they constitute the data sources contributing most volume to the DyLDO data set, on average ca. 80% of the triples. In the following, we look into more detail in each of these data sources.

From the 84 snapshots available in the dataset, we took each pair of subsequent snapshots and computed their Jaccard distance (see Sec. 3). This means, we compute the distance between the sets of triples of each pair, i.e., the more changes are detected (the deletion or addition of triples), the more distant these snapshots are.

Figures 5(a) to 5(m) show the Jaccard distance for all pairs for the 13 chosen DyLDO sources. Recall that the dynamics of a dataset can be approximated by the area under the change rate's curve. Both sources iu.edu (in Fig. 5(a)) and oreilly.com (in Fig. 5(d)) show constantly high change rates over the considered period. Less constant but with equally high change rates are the sources dbtropes.org (in Fig. 5(i)), dbpedia.org (in Fig. 5(c)) and ontologycentral.com (in Fig. 5(j)). Different is the change behavior of the sources bnf.fr (in Fig. 5(k)), berkeleybop.org (in Fig. 5(l)) and uriburner.com (in Fig. 5(m)), where we observe a low change rate over time, but with regular peaks. For the sources dbtune.org (in Fig. 5(f)) and neuinfo.org (in Fig. 5(g)), we observe only very few peaks. In particular, the behavior of the sources linkedct.org (in Fig. 5(h)), opera.com (in Fig. 5(e)) and identi.ca (in Fig. 5(b)) is interesting. Linkedct.org has in the earlier period a decreasing change rate, which later turns into increase. Opera.com has peaks in the earlier period, however it shows intensive and constant change rates at the later weeks. Finally, the identi.ca has high change rates in the first weeks, but no changes are observed at the later weeks.

Based on the Jaccard curves presented in Figures 5(a) to 5(m), we now compute the dynamics for each of these sources. As decay function, we employed again $f(t_i) = e^{-\lambda \cdot (t_n - t_i)}$ and set the decay parameter λ to 0.025 since we would like to consider older changes of almost half a year ago still with a weight of approximately 0.5. Table 4 shows the dynamics without decay function ($\Theta(\mathcal{X})$) and the dynamics when making



(a) iu.edu (b) identi.ca (c) dbpedia.org

(d) oreilly.com (e) opera.com (f) dbtune.org

(g) neuinfo.org (h) linkedct.org (i) dbtropes.org

(j) ontologycentral.com (k) bnf.fr (l) berkeleybop.org

(m) uriburner.com

Figure 5. Jaccard distance plots for the DyLDO sources

use of a decay function ($\Theta_{decay}(\mathcal{X})$) of the DyLDO sources in its last columns. The analysis of the Θ values of each source is straightforward: sources which have mostly high change rates are the most dynamics one. Therefore, in the following we consider the Θ_{decay} of these data sources (see last column of Table 4).

Highly dynamic are the sources oreilly.com, dbtropes.org, ontologycentral.com, dbpedia.org, and iu.edu. This reflects their high change rates over the entire period (with a few lows). Furthermore, the sources opera.com and linkedct.org are also highly dynamic. Looking at their Jaccard curve in Fig. 5(h) and Fig. 5(e), we see that they do not contain high change rates over the entire period. High change rates occur in the latest points in time in the period. Their high dynamic is a consequence of the use of a decay function which strengthens such kind of change behavior. Also due to the decay function, the dynamics of the source identi.ca (in Fig. 5(b)) is very low (please note that its

non-decayed dynamics is one of the highest). This data source presents no changes in the last weeks, and therefore, its dynamics drops off.

The less dynamic sources presented in Table 4 include the sources bnf.fr, berkeleybop.org, uriburner.com, dbtune.org, and neuinfo.org. These sources have a low change rate over the period studied. However, their degree of dynamics differs. The more peaks are observed (especially late in the period), the more dynamic they are.

6 Discussion

Which are the advantages of the continuous Θ-function and how to interpret it? The notion of dynamics on the basis of an integral over a continuous change rate function has several theoretic and practical reasons:

1. it allows for a natural incorporation of the continuous decay functions,
2. it allows to elegantly deal with time intervals of different sizes and,
3. is flexible to incorporate more sophisticated approximations for the change rate function.

The dynamics functions delivers a real number which can be interpreted as the *degree* of evolution of a dataset in a period of time.

What kind of change analysis is applied? In the experiments, we use the Jaccard distance to compute the changes of each snapshot pairs. In this work, we restrict change analysis to quantify changes on the set of triples (deletions and additions of triples). Without modifying the framework, it would be also possible to verify changes under others aspects such as changes on the set of entities, links, and or schema signature. We do not verify how these changes affect the dataset over time w.r.t. consistency, comprehensiveness, expressivity, etc. Instead, we verify if changes are presented. We consider such approaches orthogonal to ours.

What is the difference between the Θ function and the Δ metrics in the literature? Change metrics proposed in the literature [6,7,12] are limited to quantify changes between two datasets. In our framework, we define dynamics as an aggregation of changes, built on top of such change metrics.

Which applications can benefit from the Θ function? The dynamics function delivers a new kind of information to LOD applications, i. e., how much a dataset evolved over a given period of time. Such information can be used by methods and approaches such as updating indexes and caches of data sources [8]. For these approaches, it is important to analyse the historic change events in order to predict the most suitable point in time to update such indexes and caches. There, dynamics serve as feature in the decision process to determine which sources need to be updated with highest priority, i.e., the information of the dataset evolution is used as an input for prediction algorithms. In a following-up work, we investigate the use of our dynamics metrics in index updates scenarios. For instance, in our experiments, the maintenance of indexes relying on the sources dbtropes.org (Fig. 5(i)) and ontologycentral.com (Fig. 5(j)) will certainly need more updates than the sources berkeleybop.org (Fig. 5(l)) and neuinfo.org (Fig. 5(g)).

Which decay function should I use? There are many decay functions proposed in the literature [19]. The decision which decay function is the best depends on the application requirements For instance, for index update scenarios, it may be important to weight recent changes higher than older ones. For the scenario of accessibility of LOD documents and network traffic evolution analysis (for instance, restricting to the period of day (24h) in 30m interval slots), it may be important to strengthen the older changes instead of the recent ones.

The use of a decay function within the dynamics function is, however, not mandatory. We choose in the experiment to use the decay function $f(t_i) = e^{-\lambda \cdot (t_n - t_i)}$ in order to weaken the influence of changes that took place longer time ago. The sources linkedct.org (Fig. 5(h)) and identi.ca (Fig. 5(b)) show similar change rates, but in different periods. The source linkedct.org shows intensive changes in the later period, while the source identi.ca has intensive changes in the earlier period. Due to the decay function, the source linkedct.org has a higher dynamic degree than the source identi.ca (25.03 > 18.48). Without the decay function, this correlation would be changed, i.e., identi.ca would be more dynamic than linkedct.org (58.45 > 51.75).

7 Related Work

Various related work has investigated the characteristics of the LOD cloud. Some works conducted structural analysis of the LOD cloud such as [2,10,1] in order to obtain statistical insights into the characteristics of the data. In addition, there is related work on analyzing the LOD cloud in order to verify its compliance with established guidelines and best practices how to model and publish data as Linked Data [11]. Other works like [14] analyze LOD in order to obtain statistics about, e. g., its distribution in the network. The goal is to apply these statistics for the purpose of query recommendation. Although these works provide interesting insights into the characteristics of LOD, they typically do not consider the dynamics of the cloud, i. e., the way how it changes.

Among those works that are dedicated on the study of the Linked Data dynamics is Ding and Finin [6]. The authors have crawled about 300 million triples from different so-called Semantic Web documents (SWDs) in 2006. The authors have conducted different analyses such as extracting the age of the SWDs based on the last-modified time information contained in the HTTP response header of the SWDs. The data exhibits an exponential distribution, which indicates that many new SWDs have been added or that many old ones are actively modified. Overall, their analysis also shows that the volume of the Semantic Web documents available on the web is growing, an observation which is well consistent with and well known from other sources like the LOD cloud web site[3]. However, it remains unknown at which point in the time the different snapshots of the SWDs have been captured, i. e., also the time span starting from the initial to the final snapshot is unknown.

Umbrich et al. [17] measure the dynamics of Linked Data and the dynamics of Linked datasets with HTML documents on the Web. Their change detection uses (i) HTTP metadata monitoring, (ii) content monitoring and (iii) active notification of datasets.

[3] http://www.lod-cloud.net/, last accessed: 23 March, 2013

These three detection mechanisms are compared by several aspects like costs, reliability, and scalability of the mechanism. The content monitoring applies a syntactic comparison of the dataset content, i.e., a comparison of RDF triples ignoring inference.

The Dynamic Linked Data Observatory is a monitoring framework to analyze dynamics of Linked Data [12]. Snapshots of the Web of data are regularly collected and then compared in order to detect and categorize changes. Using these snapshots, the authors study the availability of documents, the links being added to the documents, and the schema signature of documents involving predicates and values for rdf:type and determine their change rate. Motivated by this work, in [7] Dividino et al. analyzed the changes on the usage of the vocabulary terms. They also make use of the DyLDO dataset in their experiments. The authors show that the vocabulary terms which appear in the description of a LOD document changes a lot.

An analysis of temporal information in Linked Open Data is presented in [15], i. e., temporal information available in document headers and in triples. The experiments on the BTC 2012 dataset show that only 10% of all triples explicitly provide temporal information. Thus, we have decided to apply our analysis not on this dataset but on the DyLDO dataset that provides weekly snapshots of a selected set of crawled resources. The evolution of the Web has been observed in [4] in order to obtain implications of changes on incremental Web crawlers. Incremental crawlers update local data collections if they recognize influencing changes. Likewise, the dynamics of the Web pages is empirically analyzed in [3] with a dedicated focus on the update frequencies of search engine indices. Estimations for changes of data items and elements are proposed in [5]. Such estimations are used if the history of changes is incomplete, e.g., it is known that a Web page has changed but it is not known how often it has changed in a certain period.

8 Conclusions

In this paper, we have presented a general and flexible framework for analyzing data dynamics on the LOD cloud. Different from quantifying changes of datasets, the dynamics capture the evolution of a dataset over time. We propose a function to measure the dynamics of a LOD dataset, which is defined as the aggregation of absolute, infinitesimal changes, where such changes may be quantified by the different existing change metrics in the literature. Furthermore, our method can be parametrized to make use of different decay functions for stressing or weakening changes as time passes. We apply our notion of dynamics by analyzing a large-scale LOD dataset that is obtained from the cloud by weekly crawls over more than a year in order to show the use of our dynamics function in a real world dataset. In future work, we intend to approximate the change rate function based on piecewise linear functions, polynomial interpolation and cubic splines over the observations of changes at discrete points in time. However, the benefit of such more sophisticated approximations needs to be evaluated in different real world scenarios. In particularly, we aim to analyze the benefits of the dynamics function for optimizing methods for updating indexes.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013), REVEAL (Grant agree number 610928).

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In: LDOW. Madrid, Spain (2009)
2. Auer, S., Demter, J., Martin, M., Lehmann, J.: Lodstats – an extensible framework for high-performance dataset analytics. In: EKAW, LNCS, vol. 7603, pp. 353–362 (2012)
3. Brewington, B.E., Cybenko, G.: How dynamic is the Web? Computer Networks 33(1-6), 257–276 (2000)
4. Cho, J., Garcia-Molina, H.: The Evolution of the Web and Implications for an Incremental Crawler. In: VLDB. pp. 200–209 (2000)
5. Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Trans. Internet Technol. 3(3), 256–290 (Aug 2003)
6. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: ISWC. pp. 242–257. Springer-Verlag, Berlin, Heidelberg (2006)
7. Dividino, R.Q., Scherp, A., Gröner, G., Gottron, T.: Change-a-lod: Does the schema on the linked data cloud change or not? In: COLD. CEUR Workshop Proceedings, vol. 1034 (2013)
8. Gottron, T., Gottron, C.: Perplexity of Index Models over Evolving Linked Data. In: ESWC'14: Proceedings of the Extended Semantic Web Conference (2014), (to appear)
9. Gottron, T., Scherp, A., Krayer, B., Peters, A.: Lodatio: using a schema-level index to support users infinding relevant sources of linked data. In: KCAP. pp. 105–108. ACM (2013)
10. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: Scovo: Using statistics on the web of data. In: ESWC. LNCS, vol. 5554, pp. 708–722. Springer (2009)
11. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. J. Web Sem. 14, 14–44 (Jul 2012)
12. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing linked data dynamics. In: ESWC. pp. 213–227 (2013)
13. Konrath, M., Gottron, T., Staab, S., Scherp, A.: Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Sem. 16, 52–58 (2012)
14. Neumann, T., Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins. In: ICDE. pp. 984–994 (2011)
15. Rula, A., Palmonari, M., Harth, A., Stadtmüller, S., Maurino, A.: On the diversity and availability of temporal information in linked open data. In: ISWC. pp. 492–507. Springer-Verlag, Berlin, Heidelberg (2012)
16. Schaible, J., Gottron, T., Scheglmann, S., Scherp, A.: Lover: support for modeling data using linked open vocabularies. In: EDBT/ICDT 2013 Workshops. pp. 89–92. EDBT, ACM, New York, NY, USA (2013)
17. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., Decker, S.: Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: LDOW (2010)
18. Umbrich, J., Karnstedt, M., Hogan, A., Parreira, J.X.: Hybrid sparql queries: Fresh vs. fast results. In: ISWC. pp. 608–624 (2012)
19. Yu, L., Placide, M.: Information decay in building predictive models using temporal data. JSW 7(2), 479–484 (2012)