Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not?

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

[1]WeST – Institute for Web Science and Technologies
University of Koblenz-Landau
56070 Koblenz, Germany
{dividino,scherp,groener,gottron}@uni-koblenz.de
[2]DWS – Research Group on Data and Web Science
University of Mannheim, Germany
{ansgar}@informatik.uni-mannheim.de

Abstract Recent work analyzing changes on the Linked Open Data (LOD) cloud on fine-grained weekly snapshots shows that vocabularies published on the cloud are highly static. While this result is quite expected, there is another kind of schematic information that can be observed on the LOD cloud: the use of the vocabularies in the cloud. With use, we mean the combinations of sets of properties and sets of types to describe the resources in a specific domain. Current literature does not tackle this question sufficiently. In order to gain insight into how the use of vocabularies on the LOD cloud changes over time, we present illustrating examples and a formalization of the research question. Subsequently, we present early results of experiments applied on weekly snapshots that show that the use of vocabularies indeed changes quite a lot over time.

1 Introduction

Applications that access and process Linked Open Data (LOD) are susceptible to changes of the data. The changes may affect applications to various degrees: from irrelevant effects, which have no influence on the data processing in an application, to rather critical changes that make data processing impossible without system adaptations. A first step towards coping with these changes is to understand the kind and degree of changes (and their potential effects). Consequently, the question of in which way and how much does LOD change over time has been subject to different work in the past. Quite often, the analysis are motivated by a concrete problem and focus on investigating certain "patterns" of changes over time. For example, Käfer et al. [4] addressed mainly the problem of how data dynamics affect data synchronization, smart caching, and link maintenance in hybrid architectures. Therefore, they investigated the (un-)availability of documents, quantify how many documents change, and the kinds of changes that occurred. Similarly, Ding and Finin [1] investigated changes of structured data and their influence on methods for harvesting data. These investigations mainly address changes of entities represented by unique subject URIs as well as changes on an (RDF) document level.

However, not only the changes of entities have an influence. Changes on the schema level of the data can have a much higher impact on applications consuming Linked

Data. Schema information over Linked Data is used for various purposes such as indexing distributed data sources [5], searching in large graph databases [2], optimizing the execution of queries [6], or recommending appropriate vocabularies to Linked Data engineers [8]. So far, investigations on the schema level have been relatively coarse-grained. Käfer et al. [4] consider only changes in the schema signature of documents, which involves the set of RDF predicates and object values for rdf:type. At this level, changes have been observed to occur very rarely and even if, then to a very low degree. However, such a coarse analysis does not reveal all the changes. For instance, it does not capture how the elements in the schema signature are composed to describe entities, neither, how the description of individual entities changes with respect to their schema. While this kind of changes may be less frequent than the changes of the data itself, they would have a high impact on applications that rely on schema information. Schema-level indices or summaries, for instance, must be re-computed or at least updated.

Hence, in this paper we investigate the dynamics of the schema w.r.t. its usage. In a more abstract way, the URI representing some entity is described by a set of properties P and a set of types T. Adding, removing, or exchanging a property or type will change the schema-level description of this entity, and thus result in a change of the use of vocabularies in the Linked Data cloud. Even if vocabularies such as Dublin Core, FOAF, etc. do not change frequently, we assume that the different observable combinations of properties in P and types in T used to describe a the resources actually change a lot.

In order to investigate schema dynamics on the LOD cloud, we make two contributions in this paper: First, we present a formal framework that defines what we understand by schema dynamics in terms of changes in the use of vocabulary properties and types. Second, we present the results of an early investigation of different metrics applied on the schema information computed from weekly snapshots of the Dynamic Linked Data Observatory (DyLDO) dataset[1]. This dataset has already been used for the analysis of LOD dynamics by Käfer et al. [4].

2 Scenario: Changes in the Use of LOD Vocabularies

To illustrate our notion of schema dynamics, let us introduce a toy example. We are using the FOAF vocabulary for describing persons working at the University of Koblenz-Landau in Koblenz, Germany. In addition, we describe relations between persons and their association to different projects. Besides the FOAF vocabulary, we use a domain-specific LOD vocabulary under the domain of uni-koblenz.de for modeling projects. For instance, there are individuals like uni-koblenz:ThomasGottron and uni-koblenz:Renata-Dividino that are connected via a foaf:knows property. Thomas Gottron works for the uni-koblenz:Robust project and Renata Dividino for the uni-koblenz:Media project. Table 1 summarizes the statements published on the university web site on July 2, 2013.

On July 3, 2013, a crawl of the same data, i. e., the university website is taken. Table 2 shows an excerpt of this new snapshot. Based on the statements shown in Tables 1 and 2 we can directly observe changes in the data, such as the introduction of new individuals (uni-koblenz:AnsgarScherp). But, let us take a closer look at how the

[1] http://swse.deri.org/dyldo, last accessed: July 19, 2013

Table 1. Scenario: Schema excerpt from July 2, 2013.

@prefix	uni-koblenz:	<http://www.uni-koblenz.de/> .
@prefix	rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix	foaf:	<http://xmlns.com/foaf/0.1/> .
uni-koblenz:GerdGroener	rdf:type	foaf:Person .
uni-koblenz:GerdGroener	foaf:knows	uni-koblenz:RenataDividino .
uni-koblenz:ThomasGottron	foaf:name	"Thomas Gottron".
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:RenataDividino.
uni-koblenz:RenataDividino	foaf:name	"Renata Dividino".
uni-koblenz:RenataDividino	foaf:knows	uni-koblenz:GerdGroener .
uni-koblenz:RenataDividino	foaf:mbox	mailto:dividino@uni-koblenz.de .
uni-koblenz:Robust	rdf:type	uni-koblenz:Project.
uni-koblenz:Robust	foaf:homepage	uni-koblenz/Project/Robust .
uni-koblenz:Robust	foaf:seeAlso	uni-koblenz:Projects.
uni-koblenz:ThomasGottron	uni-koblenz:worksFor	uni-koblenz:Robust .
uni-koblenz:Media	rdf:type	uni-koblenz:Project .
uni-koblenz:Media	foaf:homepage	uni-koblenz/Project/Media .
uni-koblenz:Media	foaf:seeAlso	uni-koblenz:Projects .
uni-koblenz:RenataDividino	uni-koblenz:worksFor	uni-koblenz:Media .

Table 2. Scenario: Schema excerpt from July 3, 2013.

@prefix	uni-koblenz:	<http://www.uni-koblenz.de/> .
@prefix	rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix	foaf:	<http://xmlns.com/foaf/0.1/> .
uni-koblenz:GerdGroener	rdf:type	foaf:Person .
uni-koblenz:GerdGroener	foaf:knows	uni-koblenz:RenataDividino .
uni-koblenz:GerdGroener	foaf:knows	uni-koblenz:AnsgarSherp .
uni-koblenz:AnsgarScherp	rdf:type	foaf:Person .
uni-koblenz:AngarScherp	foaf:name	"Ansgar Scherp" .
uni-koblenz:ThomasGottron	foaf:name	"Thomas Gottron".
uni-koblenz:ThomasGottron	foaf:mbox	mailto:Gottron@uni-koblenz.com.
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:RenataDividino.
uni-koblenz:ThomasGottron	foaf:knows	uni-koblenz:AnsgarScherp.
uni-koblenz:RenataDividino	foaf:name	"Renata Dividino".
uni-koblenz:RenataDividino	foaf:mbox	mailto:dividino@uni-koblenz.de
uni-koblenz:RenataDividino	foaf:knows	uni-koblenz:GerdGroener .
uni-koblenz:RenataDividino	foaf:knows	uni-koblenz:ThomasGottron .
uni-koblenz:Robust	rdf:type	uni-koblenz:ExternProjects .
uni-koblenz:Robust	foaf:homepage	uni-koblenz/Project/Robust .
uni-koblenz:Robust	foaf:seeAlso	uni-koblenz:Projects .
uni-koblenz:ThomasGottron	uni-koblenz:worksFor	uni-koblenz:Robust .
uni-koblenz:Media	rdf:type	uni-koblenz:InternProjects .
uni-koblenz:Media	foaf:homepage	uni-koblenz/Project/Media .
uni-koblenz:Media	foaf:seeAlso	uni-koblenz:Projects.
uni-koblenz:RenataDividino	uni-koblenz:worksFor	uni-koblenz:Media .

vocabularies are *used* to describe the individuals in the dataset. While the FOAF vocabulary and the domain ontology of the university themselves did not change (not shown in the tables), we observe changes in how the terms of these vocabularies are combined. For example, in the first snapshot, the property foaf:name is used in combination with foaf:knows and uni-koblenz:worksFor. This combination describe the individual uni-koblenz:ThomasGottron. In the later snapshot this combination does not occur anymore.

Furthermore, instead of the type uni-koblenz:Project, now the types uni-koblenz:ExternProjects and uni-koblenz:InternProjects are used. This implies that all combinations of vocabulary terms including the type uni-koblenz:Project do not occur anymore. An example is combination of the types uni-koblenz:Project, foaf:homepage and foaf:seeAlso in the first snapshot which does not occur any more in the later snapshot.

The examples given above, demonstrate a change in the use of the vocabulary terms for describing groups of individuals. Nevertheless, there are also combinations which remain unchanged. For instance, in both snapshots we can observe the combination of the type foaf:Person and foaf:knows as well as the property foaf:name being used in combination with foaf:knows, foaf:mbox and uni-koblenz:worksFor. Please note that adding further foaf:knows edges to, e. g., uni-koblenz:GerdGroener and uni-koblenz:ThomasGottron to connect them with uni-koblenz:AnsgarScherp does not change the use of the vocabulary in our notion since the foaf:knows property has already been used for describing the aforementioned individuals in the first snapshot.

In summary, while we do not observe any change in the vocabularies describing our scenario data, we recognize that the actual use of these vocabularies changes. This change is reflected in the different combinations of *types* and *properties* that can be observed in the data. In the next section, we systematically introduce the question of changes in the use of LOD vocabularies and present a formalization of our notions.

3 Data Levels for Observing Schema Changes in Linked Data

We distinguish two levels of how schema information is provided for data on the LOD cloud: the *abstract schema level* and the *entity mapping level*. On the so-called *abstract schema* level, we are interested in the vocabulary terms that are used in the dataset, i. e., the distinct properties and types defined by vocabularies. Moreover, the abstract schema describes which combinations of vocabulary terms are used. For instance, in the scenario we saw that the properties foaf:knows, foaf:mbox, foaf:name, and uni-koblenz:worksFor are always used together. A change at this level implies a change of the combinations of properties and types. This is understood as schema change without considering the corresponding underlying data, i. e., the individuals exhibiting the combinations of properties and types. In contrast, the *entity mapping* level associates individuals with the term combinations observed on the abstract schema level. Naturally, sets of individuals expose the same combination of properties and types of the abstract schema. In the example in the scenario (see Table 2), both individuals uni-koblenz:ThomasGottron and uni-koblenz:RenataDividino are mapped to the property set foaf:name, foaf:mbox, foaf:knows, and uni-koblenz:worksFor. In contrast, the individual uni-koblenz:GerdGroener is mapped to the set of the RDF types foaf:Person and property foaf:knows.

Formal Definition of Data Levels. Based on this notion of two levels of schema and entity in LOD, we present our concept in a more formal manner. Our formalization is based on the idea of *Characteristic Sets* (CS) proposed by Neumann and Moerkotte [6], which is used for selectivity estimation of RDF queries with multiple joins. The characteristic set of an individual s in an RDF data set G is the set of all properties P that

are used to describe s. Following previous analytics of the schema level on LOD [3] we extend this notion to consider not only the properties but also the types (classes) used to describe individuals in a dataset.

Definition 1 (Extended Characteristic Set). *Let G be an RDF dataset, P be the set of properties in G, T be the set of types in G, and we assume $P \cap T = \emptyset$. Then an extended characteristic set, ecs, is an element of the powerset over P and T, thus ecs $\in \mathcal{P}(P \cup T)$.*

Definition 2 (Extended Characteristic Set Assignment). *Let G be an RDF dataset containing* triples (s, p, o)*,where s is called the subject, p the predicate and o the object. Let S be the set of all subjects defined in triples of G. Then we define $\Lambda : S \rightarrow \mathcal{P}(P \cup T)$ as the extended characteristic set assignment, and set $\Lambda(s)$ to be the extended characteristic set assignment of a subject s. This means $\Lambda(s)$ contains all predicates and types used to describe s.*

Using the notion of extended characteristic set (ECS), we can now define the abstract schema. The abstract schema of a dataset is represented by the set of ECSs observed in an RDF dataset. In essence, the abstract schema is the result of the different combinations of vocabulary terms used to describe the individuals in the dataset.

Definition 3 (Abstract Schema). *The abstract schema AS is a subset of all possible extended characteristic sets, $AS \subseteq \mathcal{P}(P \cup T)$, and it is defined via:*

$$AS(G) = \{\Lambda(s) | (s, p, o) \in G\}$$

Informally, the abstract schema is a set of combinations of properties and classes, where each property and each class is a term from an RDF vocabulary and it is observed at least once to describe at least one individual in this dataset. Each ECS can be seen as a partition of the dataset, i. e., each individual is member of exactly one partition.

Changes in the Extended Characteristic Sets over Time. The abstract schema computed from an RDF dataset in a specific time represents a snapshot of the vocabulary terms used at this point in time. Consequently, when we analyze changes of ECSs over time, we may observe that new sets appear, existing ones are split up/merged, or disappear. The addition of new ECS to the abstract schema means that vocabulary terms are used in a combination that has not been observed before. The deletion of an ECS from the abstract schema means that a specific combination of vocabulary terms is not used any longer. When two ECSs are merged, we observe that there is a formal agreement on the semantics of individuals now being described by the same properties and types. Similarly, when a set splits into two ECS, we observe that the individuals do not agree anymore w.r.t. their semantics.

Any change at this level means that the use of the vocabulary terms to describe the individuals has changed. Thus, the intended semantics of (some of) the individuals have changed. The abstract schema captures these dynamics of the individuals' semantics.

Please note, changes on the abstract schema level reflect the changes of the actual use of vocabulary terms for describing individuals in an RDF dataset. However, it does not characterize changes in the vocabularies themselves. While the ECSs observed in

Table 3. Changes in the Extended Characteristic Sets of the Abstract Schema Level

$AS(G)$ - 2 July 2013	$AS(G)$ - 3 July 2013	Status
ecs_1 = {foaf:Person, foaf:knows}	ecs_1 = {foaf:Person, foaf:knows}	Unchanged
ecs_2 = {foaf:name, foaf:knows, uni-koblenz:worksFor}		Deleted
ecs_3 = {foaf:name, foaf:knows, foaf:mbox, uni-koblenz:worksFor}	ecs_3 = {foaf:name, foaf:knows,foaf:mbox, uni-koblenz:worksFor}	Unchanged
ecs_4 = {foaf:Project, foaf:homepage, foaf:seeAlso}		Deleted
	ecs_{4_a} = {foaf:ExternProject, foaf:homepage,foaf:seeAlso}	New
	ecs_{4_b} = {foaf:InternProject, foaf:homepage,foaf:seeAlso}	New
	ecs_5 = {foaf:Person, foaf:name}	New

an RDF dataset may change for snapshots taken at different points in time, the actual definition of the vocabularies themselves is typically stable. With other words, vocabularies like FOAF, Dublin Core, SKOS, etc. are hardly changed and updated even over a longer period of time.

Example. To illustrate possible changes on the abstract schema, let us consider the changes in our scenario described in Sec. 2. Table 3 shows these changes. The abstract schema of the dataset presented in Table 1 is described by ecs_1 ={foaf:Person, foaf:knows}, ecs_2 ={foaf:name, foaf:knows, uni-koblenz:worksFor}, ecs_3 = {foaf:name, foaf:knows, foaf:mbox, uni-koblenz:worksFor}, and ecs_4 = {foaf:Project, foaf:homepage, foaf:seeAlso}. Likewise, the abstract schema of the dataset presented in Table 2 is described by ecs_1 = {foaf:Person, foaf:knows}, ecs_3 = {foaf:name, foaf:knows, foaf:mbox, uni-koblenz:worksFor} , ecs_{4_a} = {foaf:ExternProject, foaf:homepage, foaf:seeAlso}, ecs_{4_b} = {foaf:InternProject, foaf:homepage, foaf:seeAlso}, and ecs_5 = {foaf:Person, foaf:name}. The ECS ecs_1, and ecs_3 remain unchanged across the two snapshots of our example. The set ecs_2 is deleted in the later version. Additionally, we observe with ecs_5 the use of a new combination of types and properties {foaf:Person, foaf:name}. Lastly, instead of using references to the class uni-koblenz:Project, the later version of the dataset refers to uni-koblenz:ExternProject and uni-koblenz:InternProject. Thus, the set ecs_4 is no longer used and instead the sets of ecs_{4_a} and ecs_{4_b} are applied.

Mapping Individuals and ECSs. Having analyzed the changes of the ECSs on the abstract schema level, we now look at the associations of the individuals to the ECSs on the mapping level. As said above, sets of individuals contained in an RDF dataset are described by ECSs defined on the abstract schema level.

A entity mapping set (EMS) is the group of individuals associated to one ECS.

Definition 4 (Entity Mapping Set). *Let $AS(G)$ be the abstract schema of a given RDF data set G, and ecs $\in$ $AS(G)$ be an extended characteristic set. We define the entity*

Table 4. Entity mapping level changes

2 July 2013	3 July 2013	Status
$EMS(ecs_1)$ = {uni-koblenz:GerdGroener}	$EMS(ecs_1)$ = {uni-koblenz:GerdGroener}	Unchanged
$EMS(ecs_2)$ ={uni-koblenz:ThomasGottron}		Deleted
$EMS(ecs_3)$ ={uni-koblenz:RenataDividino}	$EMS(ecs_3)$ ={uni-koblenz:RenataDividino, uni-koblenz:ThomasGottron}	Changed
$EMS(ecs_4)$ ={uni-koblenz:Robust}		Deleted
	$EMS(ecs_{4_a})$ = {uni-koblenz:Robust}	New
	$EMS(ecs_{4_b})$ = {uni-koblenz:Media}	New
	$EMS(ecs_5)$ = {uni-koblenz:AnsgarScherp}	New

mapping set EMS of ecs as:

$$EMS(ecs) \;=\; \{s \,|\Lambda(s) = ecs \;\wedge\; (s,p,o) \in G\}$$

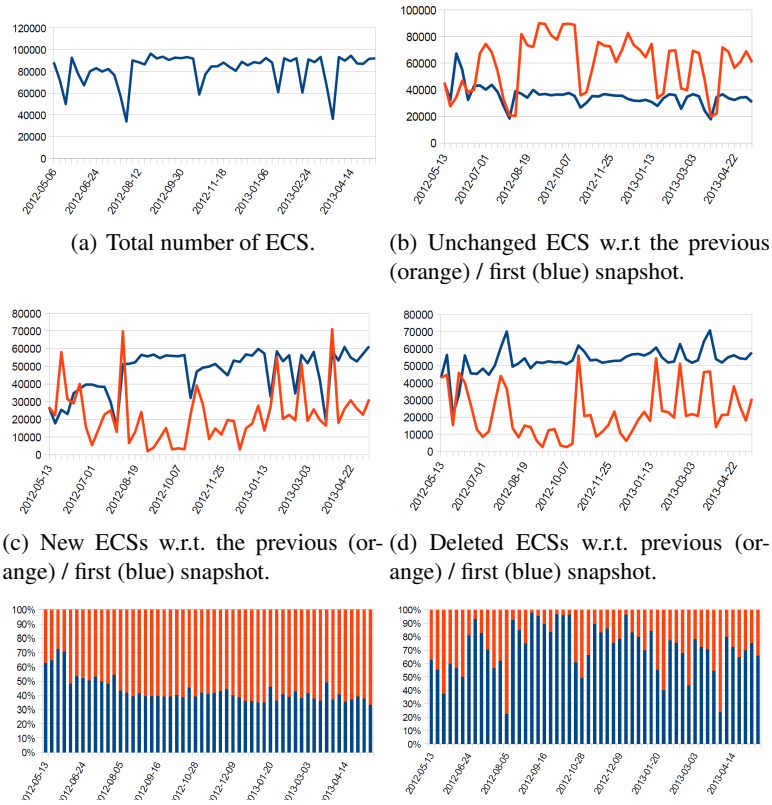
Changes in the Entity Mapping Sets over Time. Changes of the EMSs at the mapping level are characterized by changes in the association of individuals to the ECSs. These changes on the mapping level include: (1) an individual is moved to a different EMS, (2) a new individual is added to an EMS, or (3) an individual is not mapped to an EMS anymore as it has been deleted from the RDF dataset.

Example. Regarding our scenario, Table 4 summarizes the changes on the entity mapping level. Each row represents the mapping of a specific individual. The columns represent the mapping of the individuals at a specific point in time. For example, in the first row the individual uni-koblenz:GerdGroener is mapped to an EMS (identified by ecs_1) in the dataset from July 2 (first column). In the dataset from July 3 (second column), the individual uni-koblenz:GerdGroener remains in the same EMS. The individual uni-koblenz:ThomasGottron has moved from the EMS identified by ecs_2 to ecs_3 since the set ecs_2 is merged into ecs_3. Thus, on the later snapshot, ecs_2 does not identify any group of individuals anymore. A new individual uni-koblenz:AnsgarScherp is added to the July 3 dataset, and the new EMS identified by ecs_5 is used for describing this individual. Finally, uni-koblenz:Robust and uni-koblenz:Media are split into two EMSs (identified by ecs_{4_a} and ecs_{4_b}).

4 Analysis of the DyLDO Schema Dynamics

We analyzed schema changes at the two introduced levels on the DyLDO data set. We consider 53 snapshots corresponding to a period of one year (from Mai 13, 2012 until Mai 12, 2013). For more detailed information about the DyLDO dataset, we refer to [4].

Analysis of Abstract Schema Level. Fig.1(a) shows the number of ECS in the abstract schema per snapshot. We observe that the size of the abstract schema remains relatively stable over time. This implies that individuals in the dataset are mainly described



(a) Total number of ECS.

(b) Unchanged ECS w.r.t the previous (orange) / first (blue) snapshot.

(c) New ECSs w.r.t. the previous (orange) / first (blue) snapshot.

(d) Deleted ECSs w.r.t. previous (orange) / first (blue) snapshot.

(e) Total percentage of unchanged (blue), and new (orange) ECS w.r.t. the first snapshot.

(f) Total percentage of unchanged (blue), and new (orange) ECS w.r.t. the previous snapshot.

Figure 1. Abstract schema changes for the different snapshots of the DyLDO dataset.

by a nearly constant number of semantic sets (i. e., the individuals of our dataset are distributed over the existing ECSs and each ECS describes the intended semantics of a group of individuals). In our experiments not more than 96.369 variations of ECSs were observed in the snapshots. On average, 83.898 sets were used.

In addition, we observe in Fig. 1(a) that at six distinct points in time, the size of the abstract schema clearly decreases (on May 21, 2012, July 29, 2012, Oct. 21, 2012, Jan. 20, 2013, Feb. 19, 2013, and March 24, 2013). In theory, the more ECSs exist the larger is the semantic diversity among the individuals. Likewise, the less ECSs exist, the less diversity exists. Thus, at the observed points in time, we observe a reduction on the semantic diversity among the individuals.

Fig. 1(b) shows how far the number of ECS remains unchanged w.r.t. their previous version (orange line) and to the first version of the dataset (blue line). For any two abstract schemas, AS_1 and AS_2, the unchanged ECS are defined by the intersection of

these abstract schemas, $(AS_1 \cap AS_2)$. Comparing the first and the last snapshot, 35% of the ECS remain unchanged. For those long-term ECSs, we can conclude that they are established term combinations used to describe individuals and that there is a global agreement among the domains about this description.

In average, each version keeps 73% of the ECSs from the previous version. Nevertheless, ECS deletions and additions occurs. For any two abstract schemas, AS_1 and AS_2, the deleted ECS of AS_1 in AS_2 are defined by the set difference of these abstract schemas, $(AS_1 \setminus AS_2)$. Accordingly, the news ECS in AS_2 w.r.t. AS_1 are defined by the set difference of these abstract schemas, $(AS_2 \setminus AS_1)$. On average, 27% of the ECSs in a snapshot are new, and 29% of the ECS from the previous have been deleted.

However, at the six points in time where the number of ECS decreases (see Fig. 1(a)), we observe high peaks of ECS deletions (see Fig. 1(d)), and additions (see Fig. 1(c)). For instance, if we take the snapshot from May 21, 2012, we can see that 55% (27.656) of the ECS are unchanged (see Fig. 1(d)), and 45% (22.338) are new (see Fig. 1(c)). Further, 50% (44.694) of the ECS are from its previous snapshot (May 13, 2012) do not occurs in the snapshot from May 21, 2012 (deleted ECS). Fig 1(e) and Fig 1(f) show (in percentage) the total number of new ECSs and unchanged ECSs in a snapshot compared to the first/previous snapshots.

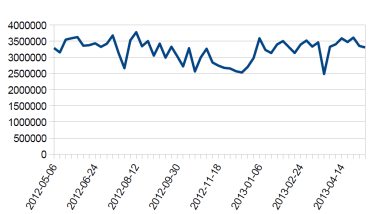
To sum up, even though we observe that the number of ECS (combinations of vocabulary terms) used in the snapshots is quite stable, and that, on average, most of the ECSs characterizing a group of individuals remain the same w.r.t. its previous version, there clearly is a frequent change in the use and combination of vocabulary terms in LOD. Please note that the DyLDO dataset contains only a one-year history of a part of the LOD cloud. Even for such a short period of time, we could show that only 35% of the combinations of vocabulary terms compared w.r.t. to the first snapshot remained the same. Thus, we can conclude that the abstract schema is highly dynamic.

Analysis of Entity Mapping Level. We turn now to the analysis of the changes on the entity mapping level. Fig. 2(a) shows the total number of individuals (subjects of the triples in the dataset) per snapshot. Similar to the abstract schema, we observe that the size remains mainly stable over time (about 3 million individuals).

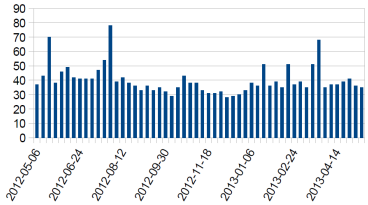
In accordance to the six points in time where we observe intensive reduction on the ECSs size (see Fig. 1(a)), the amount of individuals also decreases (on March 24. 2013, the size reaches its smallest value). In the follow-up snapshots, we observe a high increase of the number of individuals (the size reaches its maximum on Aug. 12, 2012).

Fig. 2(b) shows that, on average, there are 41 individuals per EMS. On the six points in time, where the total numbers of EMS drastically decreases (as well as the total number of individuals), we observe that the average size of EMS increases. For instance, on July 29, 2012 the average number of individuals per EMS increases to 78 (its maximum value). This means that these few remaining EMSs are more dense.

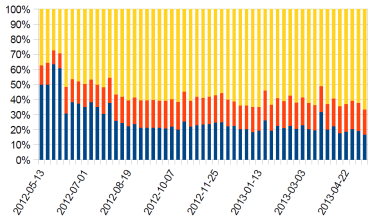
In correspondence to the analysis of ECS, the total number of ECSs is equivalent with the total number of EMSs since each EMS correlates with an ECS. We consider the EMS of new ECS, to be a new EMS. For any unchanged ECS, we check if their EMS has changed or not. Unchanged EMSs are the sets that contain the same individuals in the current version w.r.t. the previous / first version of the dataset. For instance, for any two datasets G_1 and G_2, two abstract schemas $AS_1 \in G_1$, $AS_2 \in G_2$, and given an



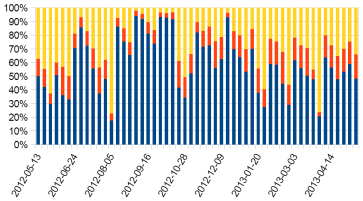
(a) Total numbers of individuals (subjects) per snapshot.



(b) Average of individuals per EMS.



(c) Total percentage of changed (blue), unchanged (orange), and new EMS (yellow) w.r.t the 1st version.



(d) Total percentage of changed (blue), unchanged (orange), and new EMS (yellow) w.r.t the previous version.

Figure 2. Mapping changes of the data from the different snapshots of the DyLDO dataset.

unchanged ECS ($ecs \in (AS_1 \cap AS_2)$), then we check if $EMS(ecs) \in G_1$ = $EMS(ecs) \in G_2$. Changed EMSs are sets that contains not the same group of individuals w.r.t. the previous / first version. This means, some individuals may be deleted and others may be added to this set. For instance, for any two datasets G_1 and G_2, two abstract schemas $AS_1 \in G_1$, $AS_2 \in G_2$, and given an unchanged ECS ($ecs \in (AS_1 \cap AS_2)$), then we check if $EMS(ecs) \in G_1 \neq EMS(ecs) \in G_2$.

Fig. 2(c) and Fig. 2(d) show the number of EMSs (in percentage) that remain unchanged, the ones that change, and the new ones w.r.t. to the previous snapshot and first snapshot. Please note that the Figures 2(c) and 2(d) showing the dynamics of EMSs correlate to the Figures 1(e) and 1(f) presenting the dynamics of ECSs. Considering only the unchanged ECS, on average, 20% of all EMSs changes w.r.t. the previous version. Taking the first and last snapshot, 51% of the EMSs changed. This implies that 17% of all EMSs remains unchanged w.r.t. the vocabulary terms and the set of individuals they are composed to. For those long-term EMSs, we can conclude that they are established terms used to describe individuals and that there is a global agreement among the domains about this description, and that the set of individuals it describes is also well-defined. These sets characterizes the static partition of the dataset.

In conclusion, the entity mapping level changes in one order of magnitude more than the abstract schema (this is obviously due to their size). Still, their dynamics highly correlates and cannot be considered separately.

5 Related Work

There exists many approaches dedicated on the study of the Linked Data dynamics. Ding and Finin [1] have crawled about 300 million triples from different so-called Semantic Web documents (SWDs) in 2006. This dataset is referred to as the SW06MAY dataset. The authors conclude that there has been a more active ontology development in the earlier time period that transitioned into more (re-)use oriented activities (i. e., use of the ontologies in the above mentioned SWDs). 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[2]. Likewise, we also show that the volume of the data in the LOD cloud tends to grow. Due to the refactoring phase, the volume stays at a constant rate.

Umbrich et al. [9] measure the dynamics of Linked Data and the dynamics of Linked Data sources with HTML documents on the Web. Their change detection uses (i) HTTP metadata monitoring (HTTP headers including timestamps and ETags), (ii) content monitoring and (iii) active notification of data sources. 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 data source 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 [4]. 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 and determined their change rate. Only 25% of the documents change frequently and they contain a balance of documents with additions and deletions. Moreover, regarding the types of changes occurring on an RDF-triple level, the authors conclude that the schema signature of documents involving predicates and values for rdf:type changed very infrequently. Motivated by this statement, we decide to study the dynamics of the schema information w.r.t another perspective. Finally, they showed that the rate of fresh links being added to the documents is very low. An analysis of temporal information in Linked Open Data is presented in [7], i. e., temporal information available in document headers and in triples. The experiments on the BTC 2012 dataset shows the use of temporal information (about 10% overall) are not sufficiently high enough to support our outlined use case. In our approach, we do not verify the agreement between the changes and the temporal information.

6 Conclusions

In this paper, we have investigated schema dynamics on the LOD cloud from a new point of view. Instead of looking how the vocabularies change over time, we verify how the *use* of vocabularies changes. With use, we mean how the properties and types, defined in the vocabularies, are applied to describe individuals in the LOD cloud. We formalize the notion of *abstract schema level* and *entity mapping level*. The abstract schema level corresponds to the set of combinations of vocabulary terms extracted

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

from a dataset. Each combination identifies a distinct group of individuals. The entity mapping level corresponds to all sets of such individuals' groups. We study schema dynamics w.r.t. these two levels.

Additionally, we provide a quantitative analysis on the schema information of the DyLDO dataset. The observation of weekly snapshots over a one-year period shows that only 35% of the combination of vocabulary terms from the first snapshot remain the same. All the others have been changed (e. g., merged, split, deleted). This implies that the data and the usage of vocabularies in the LOD cloud are in a continuous changing process. Moreover, we could also observe that during the monitoring period, six intense change events have taken place. In these phases, the amount of schema and data information has been strongly reduced and in the follow-up snapshots these information increased again.

We plan to proceed this research into three directions: (1) investigate the reasons for the peaks occurring in the plots, e.g., check the impact of the (un-) availability of documents, (2) conduct an evaluation at the pay-level domain. We assume that the schema changes of the domains in the LOD cloud behave differently and each of them influences in a certain degree the aggregated behavior of the cloud, (3) verify what kind of schema changes occur and extract patterns of changes to use as indicators for predictions.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST.

References

1. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: ISWC 2006. LNCS, vol. 4273, pp. 242–257. Springer (2006)
2. Gottron, T., Scherp, A., Krayer, B., Peters, A.: Lodatio: Using a schema-level index to support users in finding relevant sources of linked data. In: K-CAP. pp. 105–108 (2013)
3. Gottron, T., Knauf, M., Scheglmann, S., Scherp, A.: A Systematic Investigation of Explicit and Implicit Schema Information on the Linked Open Data Cloud. In: ESWC'13. pp. 228–242 (2013)
4. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing linked data dynamics. In: ESWC. pp. 213–227 (2013)
5. 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)
6. Neumann, T., Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins. ICDE 0, 984–994 (2011)
7. Rula, A., Palmonari, M., Harth, A., Stadtmüller, S., Maurino, A.: On the Diversity and Availability of Temporal Information in Linked Open Data. In: ISWC 2012. LNCS, vol. 7649, pp. 492–507. Springer (2012)
8. Schaible, J., Gottron, T., Scheglmann, S., Scherp, A.: LOVER: Support for Modeling Data Using Linked Open Vocabularies. In: LWDM'13 (2013)
9. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., Decker, S.: Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: LDOW (2010)