Understanding Ontology Evolution Beyond Deltas

Georgia Troullinou
ICS-FORTH
Heraklion, Greece
troulin@ics.forth.gr

Giannis Roussakis
ICS-FORTH
Heraklion, Greece
rousakis@ics.forth.gr

Haridimos Kondylakis
ICS-FORTH
Heraklion, Greece
kondylak@ics.forth.gr

Kostas Stefanidis
ICS-FORTH
Heraklion, Greece
kstef@ics.forth.gr

Giorgos Flouris
ICS-FORTH
Heraklion, Greece
fgeo@ics.forth.gr

ABSTRACT

The dynamic nature of the data on the Web gives rise to a multitude of problems related to the description and analysis of the evolution of such data. Traditional approaches for identifying and analyzing changes are descriptive, focusing on the provision of a "delta" that describes the changes and often overwhelming the user with loads of information. Here, we take an alternative approach which aims at giving a high-level overview of the change process and at identifying the most important changes in the ontology. For doing so, we consider different metrics of "change intensity", taking into account the changes that affected each class and its neighborhood, as well as ontological information related to the importance and connectivity of each class in the different versions. We argue that this approach will allow a better understanding of the intent (rather than the actions) of the editor, and a better focusing of the curator analyzing the changes; traditional delta-based approaches can subsequently be used for a more fine-grained analysis.

1. INTRODUCTION

With the growing complexity of the Web, we face a completely different way of creating, disseminating and consuming big volumes of information. The recent explosion of the Data Web and the associated Linked Open Data (LOD) initiative has led several large-scale corporate, government, or even user-generated data from different domains to be published online and become available to a wide spectrum of users [1]. Dynamicity is an indispensable part of LOD; LOD datasets are constantly evolving for several reasons, such as the inclusion of new experimental evidence or observations, or the correction of erroneous conceptualizations [11].

Understanding this evolution using the differences (deltas) between versions of datasets (schema and instances) has been proved to play a crucial role in various curation tasks, like the synchronization of autonomously developed dataset versions [2], the visualization of the evolution history of a dataset [5, 4], the need for accessing previous versions of a dataset to support historical or cross-snapshot queries [8], and the integration [3] and synchronization [6] of interconnected LOD datasets. Towards this, various approaches have been used for formally describing those deltas, ranging from low-level deltas (describing simple additions and deletions, see, e.g., [12]), to high-level ones (describing complex updates, such as, for instance, different change patterns in the subsumption hierarchy [6]).

However, both types of deltas aim at providing a descriptive analysis of the changes, but not at providing an overview of the changes, or the areas of the ontology that were mostly affected by the change process. Identifying the most affected areas would require a significant amount of analysis on behalf of the curator, given that the number of changes recorded are often in the range of several hundreds (or even thousands) [6]. Moreover, deltas, even high-level ones, do not provide a supervisory overview of the changes, and cannot be easily used to observe trends in the changes. Further, not all changes have the same effect on the ontology, as some may be minor (e.g., changing the label of a class), whereas others (or their combination) may significantly affect the structure or the focus of the ontology as a whole, or of particular areas in the ontology (e.g., the deletion of properties that would disconnect previously connected areas in the ontology, thereby changing its topology and focus).

Our objective in this paper is to *identify the classes that were mostly affected by the evolution process*, thereby properly directing the focus of the curator. To do that, we rely on a set of *assessment measures* that allow quantifying the "intensity" of the changes that each class underwent, based on various *assessment dimensions*. These dimensions are related to the number of changes affecting that class or its neighborhood, as well as the effect of these changes on the centrality and relevance [10] of the class in the considered versions. These measures, and their combination, will allow appropriate ranking of the classes in the ontology, in terms of "change intensity", under various different (complementary) viewpoints. Our approach is demonstrated using experiments from the CIDOC-CRM ontology, confirming the feasibility of our approach and the considerable insights gained on the evolution process.

In Section 2 we describe the assessment dimensions, which are employed in Section 3 for defining the assessment measures that quantify the importance of the changes related to each class, whereas Section 4 concludes.

2. ASSESSMENT DIMENSIONS

In this section, we present different dimensions that can be combined in order to study various aspects of ontology evolution. Each of these dimensions captures a characteristic that is arguably important in order to quantify the intensity of the changes that a class underwent.

We consider four such dimensions. The first two are related to the amount (number) of changes that this specific class (or its neighborhood) underwent during the evolution process and are defined more precisely in Subsections 2.1, 2.2. The other two are based on the idea that the amount of interest exhibited by the curator related to a class is also related to how important this class is in his ontology, and how this importance changed during the evolution process. To capture the notion of importance, we use two metrics that have been proposed in [10], namely *centrality* (Subsection 2.3) and *relevance* (Subsection 2.4).

Note that our approach is class-centric, i.e., is concerned with identifying the classes that are of most interest to the curator. However, we can easily extend our approach to properties as well.

2.1 Number of Changes

Consider the evolution of a dataset from a version V_1 to a version V_2. In principle, as presented in [6], low-level deltas are used to describe the set of triples which were added ($\delta^+_{V_1,V_2}$) along with the set of triples triples which were deleted ($\delta^-_{V_1,V_2}$) during the evolution from V_1 to V_2. The number of detected changes over this evolution is the size of their low-level delta δ_{V_1,V_2}, i.e., $|\delta_{V_1,V_2}| = |\delta^+_{V_1,V_2}| + |\delta^-_{V_1,V_2}|$.

In our case, we are interested in the changes related to a specific class only, so we use the following definition:

DEFINITION 1. *Assume the low-level delta $\delta_{V_1,V_2} = \langle \delta^+_{V_1,V_2}, \delta^-_{V_1,V_2} \rangle$ between two dataset versions V_1 and V_2, and a class n. We define the low-level delta of n between V_1 and V_2 as $\delta_{V_1,V_2}(n) = \langle \delta^+_{V_1,V_2}(n), \delta^-_{V_1,V_2}(n) \rangle$, where $\delta^+_{V_1,V_2}(n) = V_2^n \setminus V_1^n$, $\delta^-_{V_1,V_2}(n) = V_1^n \setminus V_2^n$ and $V_i^n = \{t \in V_i | t = (u_1, u_2, u_3), n = u_j, \text{for some } j\}$.*

Then, the number of changes in which n appears is defined as $|\delta_{V_1,V_2}(n)| = |\delta^+_{V_1,V_2}(n)| + |\delta^-_{V_1,V_2}(n)|$.

2.2 Number of Changes in Neighborhood

Apart from the number of changes over a specific class n, another interesting dimension is the number of changes in the classes "around" n; this allows determining whether the topology of the ontology changed in a particular area. More specifically, we define the *neighborhood* of a class n for two dataset versions V_1, V_2 (denoted by $N_{V_1,V_2}(n)$) as the set of classes that are either related to n via a subsumption relationship, or are connected with n via a property (through the property's domain/range), in either of V_1, V_2. Then:

DEFINITION 2. *Consider two ontology versions V_1 and V_2, and a class n with neighborhood $N_{V_1,V_2}(n)$. We define the number of changes in $N_{V_1,V_2}(n)$ as:*
$|\delta^N_{V_1,V_2}(n)| = \sum_{c \in N_{V_1,V_2}(n)} |\delta_{V_1,V_2}(c)|$.

2.3 Centrality

The notion of centrality [10] is used to quantify how central is a specific class in a specific dataset version. To identify the centrality of a class n in a dataset version V_j, we initially consider the instances it contains by calculating its *relative cardinality*. The relative cardinality $RC_{V_j}(e(n, n_i))$ of a property $e(n, n_i)$, which connects the classes n and n_i, is defined as the number of the specific instance connections between these two classes divided by the total number of the connections of the instances that the two classes have. Then, we combine the data distribution with the number of the incoming/outgoing properties of this class. As such, the in/out-centrality (C^{in}/C^{out}) is defined as the sum of the weighted relative cardinalities of the incoming/outgoing properties:

DEFINITION 3. *Assume the class n that appears in an ontology version V_j. The in-centrality $C^{in}_{V_j}(n)$ (respectively, the out-centrality $C^{out}_{V_j}(n)$) of n is defined as the sum of the weighted relative cardinality of the properties $e(n_i, n)$ (respectively, $e(n, n_i)$):*

$$C^{in}_{V_j}(n) = \sum_{i=1}^{m} RC_{V_j}(n_i, n) \cdot w_{(n_i,n)}$$

$$C^{out}_{V_j}(n) = \sum_{i=1}^{m} RC_{V_j}(n, n_i) \cdot w_{(n,n_i)}$$

where m represents the number of the incoming (respectively, outgoing) properties in the schema.

The weights $w(n_i, n_j)$ in the above formula have been experimentally defined and vary depending on whether the property is user-defined or RDF/S, giving higher importance to user-defined ones. This is partly because the user-defined properties correlate classes, each exposing the connectivity of the entire schema, in contrast to the hierarchical or other kinds (e.g., rdfs:comment) of RDF/S properties. More details on the notion of centrality can be found in [10].

2.4 Relevance

The notion of relevance [10] has been proposed as adequate for quantifying the importance of a class within a dataset. Relevance is based on the idea that the importance of a class should describe how well the class could represent its neighborhood. Intuitively, classes with many connections with other classes in the ontology should have a higher importance than classes with fewer connections. Thus, the relevance of a class is affected by the centrality of the class itself, as well as by the centrality of its neighboring classes. Moreover, since the dataset might contain huge amounts of data, the actual data instances of the class should also be considered when trying to estimate its importance (relevance). Formally, relevance is defined as follows:

DEFINITION 4. *Assume a class n that appears in an ontology version V_1. Assume also that the numbers of the incoming and outgoing properties that connect n with other classes are np_{in} and np_{out}, respectively. The relevance of n in V_1, $Rel_{V_1}(n)$, is the sum of the in- and out-centrality of n multiplied by the corresponding number of these classes, divided by the sum of out-centrality of the incoming classes n_i and the in-centrality of the outgoing classes n_j:*

$$Rel_{V_1}(n) = \frac{C^{in}_{V_1}(n) * np_{in} + C^{out}_{V_1}(n) * np_{out}}{\sum_{i=1}^{np_{in}} C^{out}_{V_1}(n_i) + \sum_{j=1}^{np_{out}} C^{in}_{V_1}(n_j)}$$

3. ASSESSMENT MEASURES

In this section, based on the assessment dimensions defined above, we provide different ways for combining them, resulting in useful conclusions regarding the dataset/ontology evolution. As a running example, we consider two versions of the CIDOC-CRM[1] ontology; clearly, our work could be generalized for any two versions of any given ontology. Specifically, we use versions v3.2.1 and v3.3.2, that we will subsequently call V_1 and V_2, which were released in February and October of 2002, respectively.

Table 1 illustrates a sample of classes that appear in both versions with their values of the assessment dimensions. The values of the assessment measures introduced in this section appear in Table 2.

3.1 Deltas Change

The most obvious metric is related to the number of changes that affected a particular class n. An important additional concern in this respect is that the changes on the neighborhood may be relevant as well; even if a class itself did not change much, many changes in the class' neighborhood may indicate that this class is interesting from the evolution analysis perspective. Thus, we define the delta change of a class as the weighted sum of the number of changes involving n and the number of changes involving neighbors of n:

$$\Delta Change_{V_1,V_2}(n) = \alpha \cdot |\delta_{V_1,V_2}(n)| + \beta \cdot |\delta^{\mathcal{N}}_{V_1,V_2}(n)|$$

The parameters α, β can be used by the curator to fine-tune this metric, by assigning a different weight on the number of changes in the class or the number of changes in the neighborhood, depending on the application. For instance, if the curator is not interested in the changes of the neighborhood, he could consider setting $\beta = 0$; for this case, the class *E7.Activity* would have a score of 28 (see column $|\delta_{V_1,V_2}(n)|$ in Table 1). If, on the other hand, the curator wants to consider both types (class and neighborhood) as equally important, then he could set, e.g., $\alpha = \beta = 1$; this is the case for Table 2, where the score $\Delta Change$ of *E7.Activity* is 119.

3.2 Change of Centrality and Relevance

An indirect way of measuring the effects of a change on a class is by determining how much the importance of a class changed by means of the change in its centrality or relevance. This is, in many cases, superior to the simple counting of changes, because it shows the cumulative effect of these changes on the class; and not all changes have the same effect. For example, we notice that, even thought the classes *E7.Activity* and *E2.Temporal_Enity* have the same number of changes (28 – see column $|\delta_{V_1,V_2}|$), the changes of the *E2.Temporal_Enity* have had a greater impact on its centrality. A closer look at the changes affecting these classes will reveal that the latter is involved in the addition of 27 new relationships, which increase its connectivity.

To capture this idea, we provide metrics that compute the absolute difference of the (in/out) centrality before and after the change; we also provide a metric that combines in and out centrality. Formally:

$$\Delta C^{in}_{V_1,V_2}(n) = |C^{in}_{V_2}(n) - C^{in}_{V_1}(n)|$$

$$\Delta C^{out}_{V_1,V_2}(n) = |C^{out}_{V_2}(n) - C^{out}_{V_1}(n)|$$

$$\Delta C^{overall}_{V_1,V_2}(n) = \Delta C^{in}_{V_1,V_2}(n) + \Delta C^{out}_{V_1,V_2}(n)$$

A step further of the centrality, the relevance indicates the representative power of a class, as regards its area/neighbourhood. The absolute difference of the values of this measurement in versions V_1 and V_2 reflects the change in the importance of a class during the evolution from V_1 to V_2:

$$\Delta Rel_{V_1,V_2}(n) = |Rel_{V_2}(n) - Rel_{V_1}(n)|$$

In our running example (Tables 1, 2), the class *E55.Type* has greater $\Delta C^{overall}_{V_1,V_2}$ than the class *E3.Condition_State*, however relevance exhibits the opposite behavior, that is, *E3.Condition_State* has higher $\Delta Rel_{V_1,V_2}$ than *E55.Type*. This observation indicates that even though the change of *E3.Condition_State*'s connectivity is imperceptible, the importance of the class is influenced to a great extent by the changes of its neigbourhood. As such, the focus of the specific area has probably been shifted.

Clearly, the information about the changes of each neighbourhood provides an overview of the evolution of each area (and its classes). Although the $\Delta Change_{V_1,V_2}$ takes into consideration the changes on the neighbourhood, the $\Delta Rel_{V_1,V_2}$ offers a more intuitive outcome. Taking as example the classes *E3.Condition_State* and *E7.Activity*, the latter has the greatest value (119) of $\Delta Change_{V_1,V_2}$ in the specific sample, while the corresponding value for *E3.Condition_State* is only 34. On the other hand, the difference of relevance values from V_1 to V_2 for the class *E3.Condition_State* (see column $\Delta Rel_{V_1,V_2}$) is one of the highest in Table 2, meaning that the effect of said changes in the importance of *E3.Condition_State* were significant. Overall, we conclude that $\Delta Change_{V_1,V_2}$ considers the changes (on the class itself or its neigbourhood) equally important, whereas $\Delta Rel_{V_1,V_2}$ reflects the evolution of class importance according to the trend in the changes in its surround area.

3.3 Combined Change

Note that, in some cases, different changes on the same class may "cancel out" the net effect on the importance of the class. In this case, the metrics of Subsection 3.2 may not be fully adequate. In this section, we provide a combining function that considers both the change of importance (in particular, of relevance) of a class, and the changes that this class (and its neighbourhood) underwent, essentially balancing between the two alternatives above:

$$CC_{V_1,V_2}(n) = \Delta Change_{V_1,V_2}(n) \cdot \Delta Rel_{V_1,V_2}(n)$$

Our objective here is to compute an aggregated score for the class based on its behaviour on a pure quantitative measure that cares only for numbers of changes and a more sophisticated measure that additionally regards the evolution of the connectivity of the class in the ontology. For example, in Table 2, we can see that although the classes *E7.Activity* and *E28.Conceptual_Object* have similar values for $\Delta Rel_{V_1,V_2}$, their CC_{V_1,V_2} differs significantly, due to the different number of changes affecting those classes (see column $\Delta Change_{V_1,V_2}$).

4. CONCLUSIONS

In this paper, we proposed measures that can be used to assess the evolution intensity of each class in a given dataset version. This is intended as an aid for the curator, allowing him to quickly get an overview of the most important

[1] `http://www.cidoc-crm.org/official_release_cidoc.html`

Table 1: Assessment Dimensions

CIDOC-CRM Class	$\lvert\delta_{V_1,V_2}\rvert$	$\lvert\delta^N_{V_1,V_2}\rvert$	Rel_{V_1}	$C^{in}_{V_1}$	$C^{out}_{V_1}$	Rel_{V_2}	$C^{in}_{V_2}$	$C^{out}_{V_2}$
E7.Activity	28	91	0,94	0,28	0,51	1,25	0,33	0,56
E2.Temporal_Entity	28	34	0,15	0,04	0,32	2,29	0,35	0,63
E28.Conceptual_Object	13	39	0,68	0,16	0,44	0,22	0,054	0,32
E55.Type	24	26	0,83	0,25	0,21	1,47	0,31	0,59
E3.Condition_State	6	28	1,70	0,14	0,44	0,38	0,09	0,39

Table 2: Assessment Measures

CIDOC-CRM Class	$\Delta Change_{V_1,V_2}$	$\Delta C^{in}_{V_1,V_2}$	$\Delta C^{out}_{V_1,V_2}$	$\Delta C^{overall}_{V_1,V_2}$	$\Delta Rel_{V_1,V_2}$	CC_{V_1,V_2}
E7.Activity	119	0,04	0,04	0,09	0,30	36,7
E2.Temporal_Entity	62	0,31	0,31	0,62	2,14	132,7
E28.Conceptual_Object	52	0,11	0,12	0,23	0,45	23,74
E55.Type	50	0,063	0,37	0,43	0,63	31,63
E3.Condition_State	34	0,04	0,04	0,09	1,31	44,72

changes, and properly focusing on the truly important ones. Our work is motivated by the fact that deltas are not suitable for providing an overview of the evolution process, as they are too descriptive (at a very detailed level), and often large and overwhelming for the curator. Towards this aim, we considered several dimensions that capture information on the number of changes, the centrality and the relevance of the ontology classes between different dataset versions and studied how to combine them to provide measures and useful insights on the ontology evolution.

As a future work, we plan to consider alternative assessment measures, e.g., measures that rank higher the classes that are important for the ontology, even if the changes that they underwent are minor; we may also consider alternative combinations of assessment dimensions and measures. To evaluate our approach, we plan to perform a user study; our usability experiments will consider the overhead imposed to the users for understanding the degree of change between two ontologies versus how well the proposed measures conceive the changes.

Moreover, given our set of assessment dimensions or measures, possibly augmented by additional ones, we envision to offer a number of different, user-defined, ways for combining them in order to determine complex assessment measures that provide helpful deductions regarding an ontology evolution and take explicitly into account the information needs of particular users. For example, a user may choose to combine the assessment dimensions based on the notion of *attitude*. We consider two types of attitude: the *overriding* and the *combinatory* one. Assume, for instance, the simple case in which we have only two assessment dimensions, namely $\mathcal{A}$ and $\mathcal{B}$. Then, in the overriding attitude, one of the dimensions, say $\mathcal{A}$, is given priority over the other, meaning that $\mathcal{B}$ is somehow applicable only when $\mathcal{A}$ is not (e.g., consider a user that cares for changes in the neighborhood of a class only when there are no changes on the class itself). In the combinatory attitude, both $\mathcal{A}$ and $\mathcal{B}$ contribute to the final assessment measure. For example, assume a user that considers equally important the number of changes of a class with the change of the relevance of the class between these two versions. These are similar to the standard preference methods for prioritized and pareto composition [9].

Interestingly, we believe that the issue of allowing users to define their own assessment measures arises not only when aggregating different choices of a single person, but also when aggregating choices of different people (modeling the needs of groups of users). In the former case, there is a need to combine different criteria that are posed by a single user, whereas in the latter case, we seek to reach a consensus among the members of the group. As a short-term goal, we plan to integrate such a capability to our system for visualizing the evolution of an ontology [7].

5. ACKNOWLEDGMENTS

This work was partially supported by the EU projects, DIACHRON (FP7-601043), MyHealthAvatar (FP7-600929) and iManageCancer (H2020-643529).

6. REFERENCES

[1] V. Christophides, V. Efthymiou, and K. Stefanidis. *Entity Resolution in the Web of Data*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, 2015.

[2] R. Cloran and B. Irvin. Transmitting RDF graph deltas for a cheaper semantic Web. In *SATNAC*, 2005.

[3] H. Kondylakis and D. Plexousakis. Ontology evolution without tears. *J. Web Sem.*, 19:42–58, 2013.

[4] H. Kondylakis and D. Plexousakis. Exploring RDF/S evolution using provenance queries. In *Workshops of EDBT/ICDT*, 2014.

[5] N. F. Noy, A. Chugh, W. Liu, and M. A. Musen. A framework for ontology evolution in collaborative environments. In *ISWC*, 2006.

[6] V. Papavasileiou, G. Flouris, I. Fundulaki, D. Kotzinos, and V. Christophides. High-level change detection in RDF(S) kbs. *ACM Trans. Dat. Syst.*, 38(1):1, 2013.

[7] Y. Roussakis, I. Chrysakis, K. Stefanidis, and G. Flouris. D2V: A tool for defining, detecting and visualizing changes on the data web. In *ISWC*, 2015.

[8] K. Stefanidis, I. Chrysakis, and G. Flouris. On designing archiving policies for evolving RDF datasets on the web. In *ER*, 2014.

[9] K. Stefanidis, G. Koutrika, and E. Pitoura. A survey on representation, composition and application of preferences in database systems. *ACM Trans. Dat. Syst.*, 36(3):19, 2011.

[10] G. Troullinou, H. Kondylakis, E. Daskalaki, and D. Plexousakis. RDF digest: Efficient summarization of RDF/S kbs. In *ESWC*, 2015.

[11] F. Zablith, G. Antoniou, M. d'Aquin, G. Flouris, H. Kondylakis, E. Motta, D. Plexousakis, and M. Sabou. Ontology evolution: a process-centric survey. *Knowledge Eng. Review*, 30(1):45–75, 2015.

[12] D. Zeginis, Y. Tzitzikas, and V. Christophides. On computing deltas of RDF/S knowledge bases. *ACM Transactions on the Web*, 5(3):14:1–14:36, July 2011.