SemaDrift: A Protégé Plugin for Measuring Semantic Drift in Ontologies

Thanos G. Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, Marina Riga, Panagiotis Mitzias, Ioannis Kompatsiaris

Centre for Research & Technology Hellas
6th Km Charilaou - Thermi
57001 Thessaloniki, Greece
(+30) 2311 257738

{athstavr, andreadisst, skontopo, mriga, pmitzias, ikom}@iti.gr

Abstract. Semantic drift is an active research field, which aims to identify and measure changes in ontologies across time and versions. Yet, only few practical methods have emerged that are directly applicable to Semantic Web constructs, while the lack of relevant applications and tools is even greater. This paper presents a novel software tool developed in the context of the PERICLES FP7 project that integrates currently investigated methods, such as text and structural similarity, into the popular ontology authoring platform, Protégé. The graphical user interface provides knowledge engineers and domain experts with access to methods and results without prior programming knowledge. Its applicability and usefulness are validated through two proof-of-concept scenarios in the domains of Web Services and Digital Preservation; especially the latter is a field where such long-term insights are crucial.

Keywords: Semantic drift; concept drift; semantic change; ontologies; Protégé.

1 Introduction

Evolving semantics, also referred to as *semantic change*, is an active and growing area of research that observes and measures the phenomenon of change in the meaning of concepts within knowledge representation models, along with their potential replacement by other meanings over time. In the Semantic Web (also known as Web 3.0), the representation of the underlying knowledge is typically assumed by ontologies. Thus, it can be easily perceived that semantic change can have drastic consequences on the use of ontologies in Semantic Web and Linked Data applications. In this setting, semantic change, i.e. the structural difference of the same concept in two ontologies [1], relates to various lines of research. Such examples are *concept* and *topic shift* [2], *concept change* [3], *semantic decay* [4], *ontology versioning* [5] and *evolution* [6]. A brief disambiguation of these terms can be found in [7].

This paper focuses on *semantic drift*, i.e. the phenomenon of ontology concepts gradually changing as knowledge evolves, obtaining possibly different meanings, as interpreted by various user communities or in a different context, risking their rhetorical, descriptive and applicative power [8]. *Concept drift* can refer to this language-related phenomenon, but also in abrupt parameter value changes in data mining [9].

We present the SemaDrift plugin for the Protégé platform[1], aimed at assisting a wider audience to monitor and manage concept drift. The plugin was developed in the context of the PERICLES FP7 project[2], integrating and extending existing studies [2] and previously developed open, reusable methods [7]. A graphical user interface (GUI) makes the tool more attractive for a wider audience, including non-experts, towards accessing methods for monitoring evolving semantics, as a vehicle to measure and manage ontology change. The tool is validated through two realistic real-world applications, in Digital Preservation and Web Services, demonstrating its applicability and usefulness.

The rest of the paper is structured as follows: Section 2 presents related work in metrics and tools for measuring drift. Section 3 presents the proposed framework consisting of the drift metrics and the tools functionality. Section 4 presents proof-of-concept applications, while conclusions and future work are listed in the final section.

2 Related Work

Measures of semantic richness of Linked Data concepts have been investigated in [4], proving that increasing reuse of concepts decreases its semantic richness. Other studies have examined change detection between two ontologies at a structural or content level [1]. Concept drift has been measured either by clustering while populating ontologies [10] or by applying linguistic techniques on textual concept descriptions [11]. A vector space model by random indexing has been utilized to track changes of an evolving text collection [8]. A strategy to represent change has been based on ontology evolution [6]. However, most of these techniques are not directly applicable to Semantic Web constructs or present limited statistical data.

An appealing solution transfers the notions of *label*, *extension* and *intension* from machine learning concept drift to semantic drift, further defining them in ontology terms [2]. Much philosophical debate examines how and by which properties a concept can be identified across time and appropriate formalization [12]. Some have utilized the notions of perdurance and endurance [13], so as to seek identity, by defining rigid properties that have to be persistent across instances and, thus, can identify entities [9]. Further works have followed, focusing on the extensional drift aspect of statistical data [14]. In this work we adopt, implement and integrate the methods in [2] into a familiar application for knowledge engineers, targeting not only the lack of reproducible cross-domain metrics for semantic drift, but also the lack of similar graphical user interfaces.

[1] The Protégé Ontology Editor: http://protege.stanford.edu
[2] PERICLES FP7 project: www.pericles-project.eu

3 The SemaDrift Protégé Plugin

SemaDrift aims to bring novel semantic drift measuring capabilities into a popular ontology platform, Protégé. Protégé offers many advantages to be chosen as the tool to integrate with. Traditionally as a desktop application, and recently also as a web application, it provides a user-friendly graphical interface for authoring ontologies and included entities, and naturally constitutes a more flexible alternative to plain text or RDF/OWL, especially for the unfamiliarized users.

Additionally, Protégé also integrates a variety of add-ons developed by its highly active community of users, like e.g. reasoners and third-party plugins, such as query tools and rich graph visualizations. The SemaDrift plugin fits perfectly into this multi-purpose environment, allowing users to interleave drift measurement, ontology authoring, reasoning, querying and visualization.

Both the plugin and its underlying drift metrics library are available online[3] under Apache V2 license. The metrics library was developed in Java and is based on the OWL API[4] for parsing ontologies and on Simmetrics[5] for implementing text similarity algorithms. The plugin is written in Java Swing[6], as required by Protégé.

3.1 Semantic Drift Metrics

This section presents a brief summary on the definition of the adopted metrics, as they were initially defined in [7]. Three aspects (types) of change are considered: (a) *Label* refers to the description of a concept via its name or title, thus equivalent to its $rdfs{:}label$. Label drift employs string similarity, using Monge-Elkan [15]; (b) *Intension* refers to the concept's characteristics implied via its properties, thus equivalent to the set of OWL datatype and object property triples where the concept participates either as subject or object. Intension drift uses Jaccard similarity between sets of triples; (c) *Extension* refers to the set of things a concept extends to, thus its instances. Extension drift employs Jaccard similarity between sets of instances. Total or *Whole* drift for a concept is defined as the average drift for the three aspects.

Meanwhile, the correspondence of a concept across versions can be either known or unknown. In the *identity-based* approach, concept A in ontology O_1 is known to have evolved into concept B in ontology O_2. On the contrary, in the *morphing-based* approach, concept A's identity correspondence to a single concept in the latter ontology is unknown, as concepts constantly morph into new ones. To preserve the general applicability of the tool without requiring any further, detailed and domain-dependent user input, we follow the latter approach, measuring concept drift in comparison to every other concept of an evolved ontology.

[3] SemaDrift Library API and Protégé Plugin online: http://mklab.iti.gr/project/semadrift-measure-semantic-drift-ontologies, hosted at MKLab tools: http://mklab.iti.gr/results/tools

[4] OWL API: http://owlapi.sourceforge.net

[5] https://github.com/Simmetrics/simmetrics

[6] https://docs.oracle.com/javase/7/docs/api/javax/swing/package-summary.html

3.2 Functionality

A comprehensive look at the SemaDrift plugin functionality is shown in Fig. 1. The tool provides a subset of the basic functions of the underlying SemaDrift API, in a graphical manner. For that purpose, it exposes some of its functions and accommodates the outcomes in suitable user controls using the Java Swing library. This edition of the plugin focuses on ontology pairs, i.e. two versions of the same ontology, in order to provide more insight into them and their differences, fitting also into the Protégé workspace philosophy. Usually, the users work on a single ontology at a time, which is always displayed as a tree hierarchy of classes at the left pane. Then, plugins occupy the right pane, which is free to accommodate their functions (Fig. 1).

As a first step the user has to select the pair of ontologies for which to measure drift. To take advantage of the environment, the plugin assumes that the first selected ontology is the one currently loaded in Protégé, allowing also its in-depth visualization, reasoning and query execution. The second ontology can be selected from the SemaDrift pane using the "Browse" button to look through local or remote storage.

After both ontologies are available, pressing on the "*Measure Drift*" button will display the SemaDrift metric results. Stability, as a measure of drift, is shown in two sections: overall average stability per aspect and concept pair stability for all aspects. The first section constitutes the most generic, abstract measure of drift. It displays a table with the average drift of all concepts from the former ontology to the latter, per each of the four aspects: label, intension, extension and whole. Naturally, the measurements are derived using the metrics and algorithms for each aspect described in the previous section, yielding a value from zero (no similarity) to one (full similarity).

The second section of results is displayed in respective tables. Each table row corresponds to a concept of the former ontology and each column to a concept of the latter. Consequently, each cell holds the similarity metric (i.e. concept stability) between each pair of concepts. These similarity values between pairs can further be utilized by users for different purposes; examples are given in the next section.

Concluding, the GUI in its current form is in essence a first step towards measuring semantic drift in a graphical manner. Its many possible extensions considered are given in the final section of this paper.

4 Use Case Scenarios

This section validates the applicability and usefulness of the proposed SemaDrift tools through two proof-of-concept scenarios presented below in the domains of Digital Preservation and Web Services.

4.1 Semantic Drift in Digital Preservation

The field of Digital Preservation shows much need for change detection across time and versions. The realistic scenario presented here serves as a means for validating the applicability of the framework in real-world conditions, while showcasing the usability of the SemaDrift Protégé plugin.

For the scenario a dataset was synthesized based on a ten-year period (2003-2013) acquisition log of software-based artworks by Tate Galleries, London[7]. A set of ontologies were developed, one for each year in the decade, modelling the respective domain concepts based on the Software-Based Art ontology found in [16] and [17]. A key problem we wanted to address was to investigate whether the terms used for indexing the artworks (i.e. "*Computer-based Art*", "*Mixed-Media Art*" and "*Software-based Art*") refer to semantically similar or different notions. Thus, the ontologies of the dataset were loaded in SemaDrift and the outcomes of the proposed methods were visualized, yielding otherwise inaccessible insights about semantic drift across time.

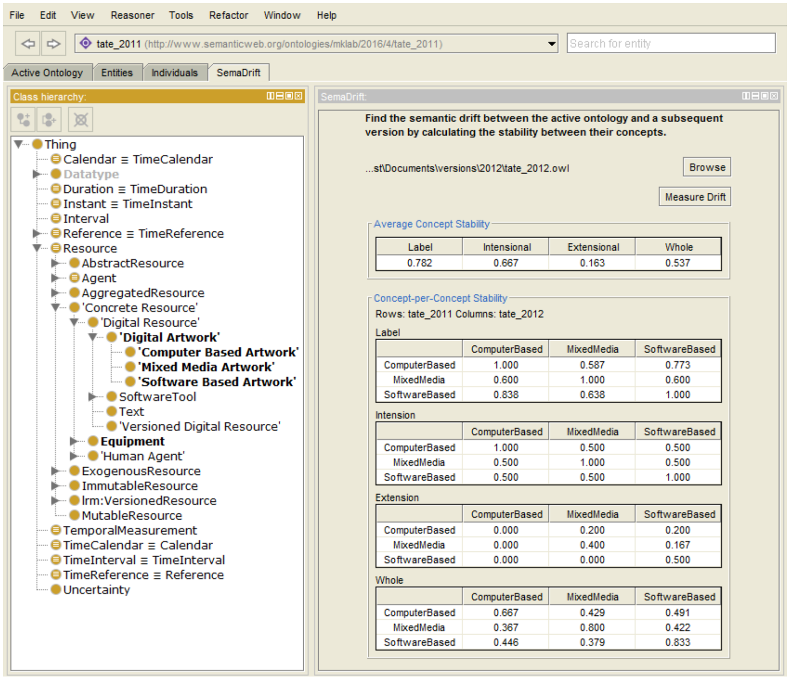


Fig. 1. SemaDrift Protégé plugin: The native tree hierarchy of the open ontology is shown on the left, while the plugin-provided content resides on the right, showing a second ontology to compare to, accompanied by the respective measurements.

Each pair of ontologies, either temporally consecutive or not, can be loaded and examined in SemaDrift. After examining all pairs in our scenario, here we showcase for simplicity only three concepts from the 2011 and 2012 versions on Fig. 1. The minimum stability is noted in the Extensional aspect, by its low Average Concept Stability. Investigating further in Concept-per-Concept Stability, instances of *ComputerBased* art in 2011 are shared between *MixedMedia* and *SoftwareBased* in 2012, while some *MixedMedia* instances are now categorized as *SoftwareBased*.

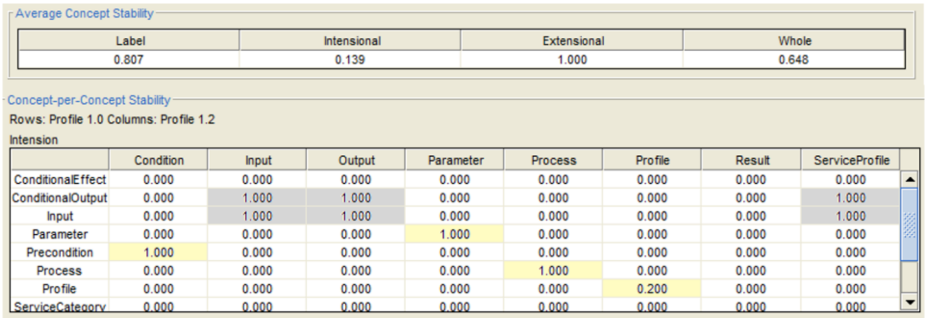
[7] Partnership with TATE within the context of the PERICLES FP7 project provided realistic knowledge for the generated models.

The other aspects are in fact stable, bearing high values. Labels are unchanged across the matrix diagonal. The other values actually represent cross-concept similarity, which can be misinterpreted as drift; an issue that future identity-based methods can tackle. The same holds for Intension: properties are retained across versions, but also all three concepts are similar, as they share half of their properties (yielding 0.5 cross-concept similarity and an overall 0.667 average).

All in all, after inspecting the results together with the domain experts from Tate, we concluded that the proposed tool and methodology indeed capture the underlying terminology change, as certain results coincided with official Tate policies (e.g. total abandonment of using the term “Computer-based” after 2012).

4.2 Semantic Drift in the Web Services Domain

The second scenario uses versions 1.0 and 1.2 of the OWL-S ontology[8] for semantic markup of Web Services, in order to demonstrate the tool’s scalability as well as its ability to quickly pinpoint semantic drifts in ontologies. In OWL-S, each service has a *Profile*, a *Grounding* and a *Process Model*. A critical piece of *Profile* metadata are operation IOPEs, defining its Input and Output information (e.g. credit card number and total price), Preconditions required to proceed with it (e.g. credit card clearance) and its Effects (e.g. transferring ownership of goods or granting access). In this scenario, the *Profile* ontology changes are immediately apparent in the SemaDrift plugin.


Average Concept Stability

Label	Intensional	Extensional	Whole
0.807	0.139	1.000	0.648

Concept-per-Concept Stability
Rows: Profile 1.0 Columns: Profile 1.2
Intension

	Condition	Input	Output	Parameter	Process	Profile	Result	ServiceProfile
ConditionalEffect	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ConditionalOutput	0.000	1.000	1.000	0.000	0.000	0.000	0.000	1.000
Input	0.000	1.000	1.000	0.000	0.000	0.000	0.000	1.000
Parameter	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Precondition	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Process	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
Profile	0.000	0.000	0.000	0.000	0.000	0.200	0.000	0.000
ServiceCategory	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Fig. 2. Average concept and concept-per-concept intension stability for OWL-S, 1.0 vs 1.2.

As Fig. 2 shows, average concept drift originates from intension, which is investigated further. No instances exist for measuring extension, and labels changed only slightly. Some concepts vanished (e.g. *ConditionalEffect*, *ServiceCategory*) and some stayed the same (symmetrical concepts *Process*, *Parameter*). However, changes were detected in *Profile*, which bears altered properties and *Precondition*, which migrated to *Condition*. Other concepts present full stability simply because they bear no properties (marked as gray, while the remaining non-zero entries are marked in yellow).

---

[8] OWL-S ontology: https://www.w3.org/Submission/OWL-S/

5 Conclusions and Future Work

This paper presented a Protégé plugin for measuring semantic change in terms of concept drift. Based on state-of-the-art notions, methods for measuring label, intensional, extensional and whole (total) drift have been adapted, optimized and implemented in the SemaDrift open source software library. The proposed domain-independent, cross-platform software tool was integrated with the popular Protégé platform, enriching its multi-purpose knowledge engineering environment with semantic drift measurement capabilities, as showcased through two proof-of-concept scenarios, in Digital Preservation and semantic markup for Web Services.

SemaDrift shows much room for future improvement. The chain of ontology versions to compare to, which is currently only limited to two, will be increased using more GUI controls. Combined with visualization capabilities, the user will be able to view entire morphing chains effortlessly, targeting long-term investigation. While now the method does not require further input to pinpoint identities, users could do so in the future, yielding a series of identity-based metrics which could be more valuable in certain cases. Finally, a standalone desktop application is planned to allow this level of flexibility at the GUI level as well as to appeal to a wider audience.

6 Acknowledgements

This research received funding by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES.

References

1. Tury, M., Bieliková, M.: An approach to detection ontology changes. In: Workshop proceedings of the sixth international conference on Web engineering - ICWE ’06. p. 14. ACM (2006).
2. Wang, S., Schlobach, S., Klein, M.: Concept drift and how to identify it. J. Web Semant. 9, 247–265 (2011).
3. Uschold, M.: Creating, integrating and maintaining local and global ontologies. In: Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000), Berling, Germany.
4. Pareti, P., Klein, E., Barker, A.: A Linked Data Scalability Challenge: Concept Reuse Leads to Semantic Decay. In: Proceedings of the ACM Web Science Conference. ACM Press-Association for Computing Machinery (2016).
5. Yildiz, B.: Ontology Evolution and Versioning. Tech. Report, TU Vienna. (2006).
6. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven ontology evolution management. Knowl. Eng. Knowl. Manag. Ontol. Semant. Web. 133–

140 (2002).
7. Stavropoulos, T.G., Andreadis, S., Riga, M., Kontopoulos, E., Mitzias, P., Kompatsiaris, I.: A Framework for Measuring Semantic Drift in Ontologies. In: 1st Int. Workshop on Semantic Change & Evolving Semantics (SuCCESS'16). CEUR Workshop Proceedings, Leipzig, Germany (2016).
8. Wittek, P., Darányi, S., Kontopoulos, E., Moysiadis, T., Kompatsiaris, I.: Monitoring term drift based on semantic consistency in an evolving vector field. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–8. IEEE (2015).
9. Meroño-Peñuela, A., Hoekstra, R.: What is Linked Historical Data? In: Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014). pp. 282–287. Springer International Publishing (2014).
10. Fanizzi, N., Amato, C., Esposito, F.: Conceptual Clustering: Concept Formation, Drift and Novelty Detection. In: The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings. pp. 318–332. Springer (2008).
11. Gulla, J., Solskinnsbakk, G., Myrseth, P.: Semantic Drift in Ontologies. In: Proceedings of 6th International Conference on Web Information Systems and Technologies (WEBIST), Valencia, Spain. pp. 13–20 (2010).
12. Guarino, N., Welty, C.: A Formal Ontology of Properties. In: Knowledge Engineering and Knowledge Management Methods, Models, and Tools. pp. 97–112. Springer Berlin Heidelberg (2000).
13. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 166–181. Springer Berlin Heidelberg (2002).
14. Mitzias, P., Riga, M., Waddington, S., Kontopoulos, E., Meditskos, G., Laurenson, P., Kompatsiaris, I.: An ontology design pattern for digital video. In: Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015) (2015).
15. Monge, A.E., Elkan, C.: The Field Matching Problem: Algorithms and Applications. In: 2nd Intl. Conf. Knowledge Discovery and Data Mining (KDD). pp. 267–270 (1996).
16. Lagos, N., Kontopoulos, E., Riga, M., Mitzias, P., Meditskos, G., Waddington, S., Laurenson, P.: Designing for inconsistency–the dependency-based PERICLES approach. In: East European Conference on Advances in Databases and Information Systems. pp. 458–467. Springer International Publishing (2015).
17. Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T., Lagos, N., Vion-Dury, J.-Y., Meditskos, G., Falcão, P., Laurenson, P., Kompatsiaris, I.: Ontology-based Representation of Context of Use in Digital Preservation. In: 1st Workshop on Humanities in the Semantic Web - WHiSe, co-located with the 13th Extended Semantic Web Conference (ESWC 2016), At Heraklion, Crete, Greece. p. CEUR Workshop Proceedings, Vol-1608, pp. 65–72 (2016).