An overview of Semantic Change: Understanding the Phenomenon, Current Trends and Future Research Roadmap

Community of Practice on Evolving Semantics

A report written by PERICLES Project partners in collaboration with participants of the Community of Practice (CoP) on Evolving Semantics.

This report results from over a year of discussions and knowledge sharing by researchers interested in the research and application of the area of evolving semantics and is a summary of those conversations. The document also touches upon recent research in the area of semantic change, starting with the role of semiotics in identifying semantic drift, to content and community change in the media-art case study, and finally to the study and detection of semantic drift in ontologies.

Introduction and Motivation

Semantic drift (also often referred to as "semantic change" or "evolving semantics") is an active and growing area of research that observes and measures the phenomenon of changes in the meaning of concepts within knowledge representation models, along with their potential replacement by other meanings over time. Understanding this phenomenon and being able to anticipate trends plays a central role in the management and long-term access and reuse of digital collections.

This field touches one of the research interests within the PERICLES project as it addresses the challenge of ensuring that digital content remains understandable and from this perspective accessible and reusable in an environment that is subject to continual change.

Thus, in order to collect a series of perspectives on the theory of evolving semantics and cultural change and their relevance for the practical field, the **Evolving Semantics Community of Practice (CoP)** brought together a group of researchers and academics in the areas of science, linguistics and semantic web. The main aim was the exchange of thoughts, research and perspectives amongst experts working on topics closely related to the change in semantics, to find inspiration for one's own work and help "thinking out of the box".

Members

The CoP was led mainly by Prof. Sándor Darányi from the University of Borås, with contributions by Stratos Kontopoulos from CERTH/ITI and Alastair Gill from King's College London. The rest of the CoP participants included:

- Daniel Galarreta, CNES
- Albert Meroño Peñuela, VU University Amsterdam
- Stephanie Roth, Swedish Council for Higher Education
- Jeremy Barraud, University of the Arts London
- Emma Tonkin, University of Bristol

Approaches

Discussions during the lifetime of this CoP led to distinguishing the most prominent approaches adopted by CoP members for studying semantic change - we note that this work is ongoing and that it is often an interest that has been woven throughout a researcher's career. The CoP and the PERICLES project in general, along with the SuCCESS'16 and Drift-a-LOD'16 workshops (see also next section), have provided an arena for presenting and discussing this work. The research described in the following subsections gives a flavour of the vibrant and contrasting approaches which have been pursued in relation to the study of semantic change. From an evaluation of

semiotics for the study of meaning and knowledge, via the study of Tate galleries' catalogue data and social media community behaviour, to the study of semantic change in ontologies and a tool to measure it, there is evidence of the breadth of this research.

A Semiotic Contribution to the Semantic Drift Issue

Beside the approaches designed to meet the needs of detecting semantic drifts and preserving data semantics, there is room for an epistemological analysis of these issues. What are the epistemological positions that underlie these approaches, that is, what are the notions that are implemented in the expressions of these problems and their solutions? This first approach will primarily focus on two aspects: the notions of meaning and of knowledge.

The question of meaning. Using this notion rests upon a semiotic conception, conscious or not, naïve or intellectual of that term, in so far as semiotics is considered as a "discipline that has been designed to make explicit – in the form of a conceptual construction – the conditions of the grasp and of the production of meaning" [Greimas et al., 1982].

The question of knowledge. As far as knowledge is concerned, it obviously depends upon an epistemological presupposition, since it asks the question of the relation between language and reality either under a positivist, realist or constructivist hypothesis.

Before focusing upon these questions in more detail, we first explain the motivation for this focus given our experiences of applying tools and methods to the detection of semantic drift. In 2004 the Rosetta probe was launched by the European Space Agency (ESA) in order to meet the comet 67P/Churyumov-Gerasimenko ten years later on 6 August 2014. It was the first mission in history to rendezvous with a comet. More than twenty years would take place between the decision and the landing (not to mention the phase of exploiting the collected data!). Studies were conducted in an R&D program between 2002 and 2012 by the French space agency (CNES) in order to bring solutions to the question of detecting possible knowledge loss. At the mid-term of this period a new question was added: exploiting scientific data long after their production or by a new scientific community which was also related to the problem of knowledge variations and their detection. In the first case the goal was to prevent loss of knowledge which could result in a severe dysfunction of the Rosetta lander. In the second case, the community which had been intended to exploit the resulting data "may have lost its familiarity with some terminology, and the definition of the community may have been broadened to include other members with different backgrounds" [CCSDS, 2012].

In order to address these questions, a range of techniques were developed and applied from statistics, knowledge engineering, logic and linguistics, and community mining. Specifically in the first case, a blended approach [Condamines et al., 2003] was proposed to detect evolution within the documents using:

- Statistical distribution of terms within the project documents
- Impact on taxonomies of domains of the project [Rothenburger, 2002]
- Logical relations between "interesting" concepts [Rajman et al., 1998]
- Linguistic cues extracted from documents [Habert et al., 2002]

In the second case links between web pages were used to identify and characterize the virtual communities on the Web using the analysis of links [Kleinberg, 1997] and identification of communities [Flake et al., 2000].

Despite the high quality outcomes of these approaches, and despite the fact that all the variations detected were interpretable, none of the results allowed the prediction of a knowledge loss likely

to lead to a feared incident. Why was there such an outcome? Because from the beginning we made the implicit assumption that it is possible to assimilate the meaning of an observed fact to a piece of knowledge. It is clear that a piece of knowledge is necessarily for conveying a meaning – a meaningless knowledge is simply absurd –, but that does not imply in a constructivist epistemology that the meaning of an observed fact corresponds to a piece of knowledge. Indeed according to that epistemology, an observation is an artefact that results from a particular computational model which corresponds to a particular point of view associated to a discipline. If this observation is a view of something, it does not follow that it is a view of an object which can be recognized as such by other viewpoints that could be called upon in that context.

In any case, because of this relation between meaning and knowledge, a theory of knowledge must include a theory of meaning, namely a semiotic theory. Roughly speaking semiotics can be defined as a theory of signs or as a theory of language. Considered as a theory of signs, it is akin to Aristotle's conception of language: "Spoken words are the symbols of mental experience and written words are the symbols of spoken words" [Aristotle, 2015]. C.S. Peirce (1839-1914) is the most distinguished representative of this trend. He developed a semiotic theory that is at once general, triadic and pragmatic [Everaert-Desmedt, 2011]. There are other triadic theories of signs such as that of Charles Morris (1901-1979) who drew upon behaviourist principles [Falk, 2004]. Considered as a theory of language, semiotics is akin to the linguistic theory of F. de Saussure (1857-1913). Saussurean linguistic is "based upon cardinal oppositions between *langue* and *parole*, synchrony and diachrony, the paradigmatic and the syntagmatic, and the orders of signifier and signified" [Norris, 2004]. Despite the fact that de Saussure is generally considered as "a major fountainhead of semiotics" [Bouissac, 2004], it was essentially Louis Hjelmslev (1899-1965) and Algirdas Greimas (1917-1992), and later on the School of Paris [Greimas et al., 1982], who developed a semiotics in its own right drawing on Saussurean principles.

In order to account for the designing activity of complex systems such as space systems we proposed a semiotic methodology which was based upon Hjelmslev's concepts. In order to clarify how different trades can agree on their different views of a space system and share their knowledge, we introduce a multi-viewpoints semiotic methodology in order to answer that question [Galarreta, 2010] [Galarreta, 2013].

This methodology rests in part on Hjelmslev's theory (known as glossematics) [Hjelmslev, 1961], which partly rephrases positions of F. de Saussure regarding language as a system. Hjelmslev (with Saussure) considered that any true language possesses two planes: a plane of content (Saussure's signified) and a plane of expression (Saussure's signifier), and rejects the idea that a word can convey just by itself a meaning, and even more, that this meaning can be reduced to its signifier.

The outstanding proposal of Hjelmslev in his theory is the statement that a language has two planes and that these two planes have similar functioning. Therefore there is not any possible meaning if one of these planes is missing. Our contribution to the project of proposing a theory of knowledge that includes a theory of meaning took the form of a **multi-viewpoints semiotic methodology**, in which we defined viewpoints in such a way that they played the role of planes.

One is not forced to embrace the epistemological position we adopted, but must be conscious that within that epistemology the above result cannot be ignored. The Estonian semiotician Juri Lotman (1922 – 1993) insists on that idea: "The idea of the possibility for a single ideal language to serve as an optimal mechanism for the representation of reality is an illusion. A minimally functional structure requires the presence of at least two languages and their incapacity, each independently of the other, to embrace the world external to each of them. This incapacity is not a deficiency, but rather a condition of existence, as it dictates the necessity of the other (another person, another language, another culture)" [Lotman, 2009].

In that sense detecting semantic drifts needs to correlate views produced by different viewpoints. Preserving data needs to take into account the risk of losing knowledge of the objects for which these data account. In that situation a semiotic analysis leads to the interpretation of risk not as a sign of ignorance, but on the contrary as what is usually defined as a piece of knowledge [Galarreta, 2007]. We believe that this semiotic approach brings a new perspective in understanding what can be expected from studies of semantic drift as well as how to conceptualise the issues involved in this line of research. In the following, we present two studies which bring their own perspectives on the issues of semantic drift and community change.

Observing Change over Time

The Tate Galleries hold the national collection of British art from 1500 to the present day and international modern and contemporary art. The collection embraces all media, from painting, drawing, sculpture and prints to photography, video and film, installation and performance. The 19th century holdings are dominated by the Turner Bequest with around 30,000 works of art on paper, including watercolors, drawings and 300 oil paintings. The catalogue metadata for the 69,202 artworks that Tate owns or jointly owns with the National Galleries of Scotland are available in JSON format as open data[1]. Out of the above, 53,698 records are timestamped. The artefacts are indexed by Tate's own hierarchical subject index which has three levels, from general to specific index terms. As such, this timestamped data set provides a unique and interesting snapshot of linguistic and cultural changes over a period of 200 years.

Although there are certainly limitations associated with this data set (for example, it is limited to artworks and to a set of artworks with specific coverage), we believe that this collection can provide interesting insights in itself since artworks are a reflection of the society and culture in which they were produced, as well as the fact that it provides an opportunity for exploring relevant methods - we note in particular the open source **Somoclu** tool (Wittek et al. 2015) developed in PERICLES.

In this set of experiments, we were interested in Somoclu's capability to reconstruct Tate's complex index terms from lower level, i.e. single word components in a dynamically changing semantic environment modelled by multivariate statistics. This was important to find out more about the quality of automatic subject index construction as compared with its manual predecessor in a scalable environment. Index term drift detection, measurement and evaluation were based on the analysis of emergent self-organising maps (ESOMs; Ultsch, 2005), leading to drift logs on all indexing levels. Parallel to that, covering every time step of collection development, we also extracted normalized histograms to describe the evolving topical composition of the collection, and respective pie charts to describe the thematic composition of term clusters. Further, to check cluster robustness, hierarchical cluster analysis (HCA) dendrograms were computed for term-term matrices, also compared with those from term-document matrices. On one hand, these gave us a detailed overview of semantic drift in the analyzed periods. On the other hand, the observed dynamics could be modeled on a gravitational force field and its generating potential (more detailed descriptions of these methods can be found in DaránYi et al. 2016). By this physical metaphor, we could shed light on the dynamic origins of semantic drifts as well as other characteristic changes over time; key findings and analytic approaches are reported in the following.

[1] https://github.com/tategallery/collection

Firstly in terms of semantic drifts, content mapping means that term membership for every cluster in every time step is recorded and term positions and dislocations over time with regard to an anchor position are computed, thereby recording the evolving distance structure of indexing terminology. This amounts to drift detection and its exact measurement. Adding a drift log results in extracted lists of index terms on all indexing hierarchy levels plus their percentage contrasted with the totals.

Drifts can be partitioned into splits and merges. In case of a split, two concept labels that used to be mapped on the same grid node in one epoch (or time period) become separated and tag two nodes in the next phase, while for a merge, the opposite holds. From an information retrieval (IR) perspective splits decrease recall and merges decrease precision, limiting the quality of access; from the perspective of long term digital preservation, they indicate at-risk indexing terminology.

Splits and merges were listed by Somoclu for every epoch over both measurement periods. For instance a sample semantic drift log file recorded that due to new entries in the catalogue in 1796-1800, by 1800 on subject index level 2, the term 'art' was separated from 'works', as much as 'scientific' was from 'measuring', whereas 'monuments', 'places and 'workspaces' were merged, i.e. mapped onto the same coordinates. Therefore, based on the same subject index terms, anyone using this tool in 1800 would have been unable to retrieve the same objects as in 1796.

In a vector field, all the terms and their respective semantic tags are in constant flux due to external social pressures, such as new topics over items in the collection due to the composition of donations or fashion. Without data about these pressures quasi embedding and shaping the Tate collection, the correlations between social factors and semantic composition of the collection could not be explicitly computed and named. Still, some trends could be visually recognized over both series of maps, going back to their relatively constant semantic structure where temporary content dislocations did not seriously disturb the relationships between terms, that is, neighbouring labels tended to stick with one another, such as 'towns, cities, villages' vs. 'inland' and 'natural'. In other words, the lexical fields as locally represented by Somoclu remained relatively stable.

The stability of these fields was measured in terms of drift rates which were computed by detecting the splits and merges that happened to the best matching units (BMUs). Specifically, we were not looking at the distance they travelled, rather at the fact that they formed or joined or moved away from a cluster (i.e. a BMU) in between epochs.

Overall, in this particular collection, splits between level 1 concepts took place occasionally, whereas both splits and merges occurred on indexing levels 2-3 on a regular basis. The drift rate was increasingly high: for level 2 index terms, it was 19-22 % in the 1796-1845 period vs. 15-27.5 % in 1960-2009, whereas for level 3 terms it was 29-57 % (1796-1845) vs. 54-61 % (1960-2009). These percentages suggest that the more specific the subject index becomes, the more volatile its terminology, especially with regard to modern art. At the same time, evaluation on all three indexing levels suggested that certain parameter combinations of Somoclu excelled in reconstituting the original Tate index terms, such as 'towns, cities, villages' above.

There is also evidence from this analysis relating to the social tensions shaping the Tate collection, for example by comparing the level 2 indexing vocabularies across different periods. In general, this is where one witnesses the workings of language change, part producing new concepts, part letting certain index terms decay. For example, the focus may shift from a concept to its variant (e.g. 'nation' to 'nationality'), a renaissance of interest in the transcendent beyond traditional notions of religion and the supernatural ('occultism', 'magic', 'tales'), fascination for the new instead of the old, or a loss of interest in 'royalty' and 'rank'. Toys and concepts like 'tradition',

the ‘world’, ‘culture’, ‘education’, ‘films’, ‘games’, ‘electricity’ and ‘appliances’ make a debut in art. A representation of such tendencies of content change combined with manifest tensions is visualized in Figure 1. In the figure, blue basins host content, whereas brown ridges indicate tensions.

Here, the tendency to merge or split means a projected possible, but not necessarily continuous, trend - should the composition of the collection continue to evolve over the next epoch like it used to develop over the past one, the indicated splits and merges would be more probable to form new content agglomerations than random ones. Further investigation relating to content dynamics showed the interplay between semantic similarity and term importance in a social perspective, leading to work in progress. For a more comprehensive overview, we encourage the reader to study selected sections in chapters 3-4 of Deliverable 4.5 “Context-aware content interpretation”.

Fig. 1: Excerpt from the tension vs. content structure changes in the level 2 (intermediate) index term landscape in 1796-1805 (left image shows period 1796-1801; right image shows period 1801-1805). Blue basins host content, brown ridges indicate tensions. Whereas ‘towns’, ‘cities’, ‘villages’ remain merged over both epochs, ‘inland’ and ‘natural’ become merged by 1805.

To summarise, this work has investigated drift detection and measurement using data from the Tate Galleries, has shown that semantic drift is a regular occurrence, and that this becomes more common with greater index term specificity. In addition to surveying the evolving semantic content structure, the analytic tool Somoclu has also mapped the parallel evolution of classification tension structure, a requirement for future modeling and anomaly prediction tasks.

Semantic Change and User Communities

In addition to changes in the meaning and interpretation of words and concepts, a closely related issue, noted previously, is that of different user communities accessing the same resources or the risk of a user communities changing over time. Indeed, this may present an even greater risk of resources becoming inaccessible, since different communities are likely to wish to access information or objects in slightly different ways, for example using different search terms (specific to the different communities), or for a slightly different purpose, perhaps to look for ideas, inspiration and enjoyment rather than simply to identify the provenance of objects.

One way to investigate the role of communities is to use online social networks as a way to empirically study and understand those user communities around a cultural heritage organisation, again the Tate galleries. Our motivation is the elicitation of information that enables us to identify and support these user groups effectively at present and into the future. We do not view ‘user community’ as a single or static entity, and therefore expect that such methods are

well suited to understanding such complex phenomena, as well as providing a tractable monitoring method that could be implemented in future.

Being able to identify change in this community is important to preservation for assessing the social and cultural context of risk, in particular, it is important for the institution (as well as larger cultural and government agencies) to be able to monitor and manage who their audience is for access to the institution and its resources (Schlieder, 2010). We use social media for the monitoring of social context with a view to mitigating risk resulting from changes in social context. In the case of the Tate user community identified using social media data, this is largely self-selecting, and we therefore expect it to be fluid and dynamic; any changes are likely to evolve over time.

By combining social network analysis of the Tumblr network along with topic modelling (Blei, 2012) to the content of the posts of the Tate community on Tumblr, we have identified an example of change in this community relating to the growth in and around 2012: Using 5 and 15 topic model solutions (shown in Figures 2 and 3) to describe the data at different levels of granularity, these two approaches were able to identify change in the content generated by the Tumblr community in relation to Tate; in particular, the adaptation of this social network and its content to meet its new needs. The 5 topic model identified a temporary change in focus from catalogue data to image data, and a greater focus on the Tate Modern and sharing exhibition information. Although the first two topic changes may indicate an exploration with new media, it is the focus on Tate Modern by the community and sharing/promotion of exhibitions which seems to indicate a more substantive shift in community usage of Tumblr.

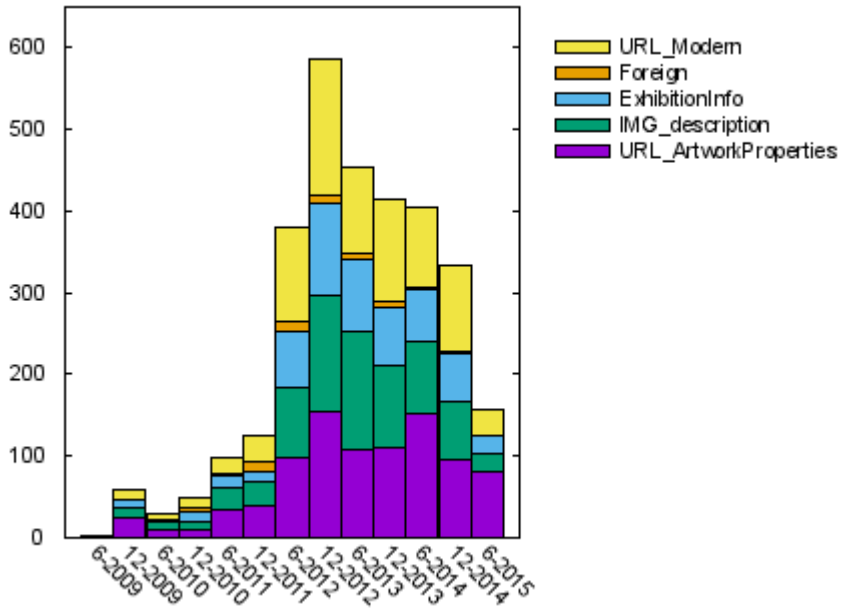

Fig. 2: Tumblr topic frequency over time (2009-2015) - 5 topic model.

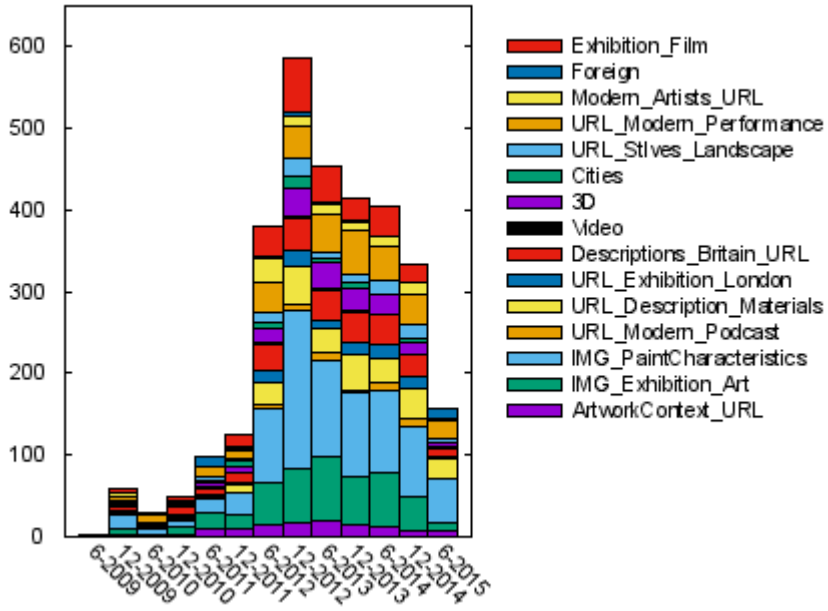

Fig. 3: Tumblr topic frequency over time (2009-2015) - 15 topic model.

For the 15 topic model, although many of the topics are used infrequently and which come and go in usage, in this analysis example we focused on five. From this example analysis, we found that following 2012 there was an increase in the popular describing and critiquing of art objects, along with a temporary focus on Tate Modern artists, and similarly less focus on images relating to exhibitions and performances at Tate Modern. Of these, we note that the change of focus relating to Tate Modern artists rather than exhibitions is interesting, and provides more detail to the general increase in posts relating to Tate Modern identified in the 5 topic model; in contrast, the increase in description and critique of art objects captured by the 15 topic model is only regarded as a temporary change in exploring the use of image descriptions in the 5 topic model. Regardless of these nuances, we view these broad changes as the increase in number of art appreciation posts, as well as an increased interest in the community relating to Tate Modern. Both of these large scale changes of community behaviour are indicative of a social and cultural context, which we expect to be important in understanding the Tate in its broader online and offline community context.

Overall, the results from the two models show similar changes in the Tate Tumblr community (primarily the description of art objects and coverage of Tate Modern), but their different granularity and probabilistic generation mean that they provide detail in different ways, in some cases identifying increase of a topic, and in others the change in use from one topic to a similar one, but with nuanced differences. This would indicate therefore that at least for initial monitoring purposes, it would makes sense to include the topics from both models in this process, thereby allowing the greatest insight into community change processes; the disadvantage to this is that there would be a slightly greater amount of data to consider, but in this case it does not seem to be too arduous, given that this would result in 20 topics in total. As previously noted in relation to the 15 topic model, some of these topics occur with a relatively low frequency in the Tumblr data – this may lead to the possibility that such a model over fits the data, however, given that we propose the inclusion of the 5 topic model, then we expect this risk to be mitigated by the use of the broader topics, and greater coverage that this smaller model provides.

In the following two studies, we turn to the specific application of semantic drift research in the area of ontologies and the Semantic Web: the first presenting an empirical analysis of drift in linked data, and the second for its detection in this data.

Learning Semantic Drift from Past Versions of Linked Data Vocabularies

The Semantic Web brings structure to the content of the Web [Berners-Lee et al., 2001], by combining the Resource Description Framework (RDF), vocabularies and reasoning. This greatly simplifies the integration of heterogeneous data sources in common Web dataspaces. In this design of semantically-enhanced Linked Data on the Web, schemas like *ontologies* and *vocabularies* play a central role, allowing users to semantically describe and link their data. These vocabularies are curated by a number of publishers that regularly release new vocabulary versions. For example, *schema.org* has released 23 different vocabulary versions between 2012 and 2015[2]. Usually, publishers change their vocabularies to "reflect changes in the real world, changes in the user's requirements, and drawbacks in the initial design" [Stojanovic et al., 2002]. The quality of those updates is difficult to estimate, and only manually assessed at best.

The observation of Web data of dubious quality has given rise to various methods of *Web data quality* assessment. Data quality is commonly conceived as "fitness for use by data consumers" [Wang 1996], and metrics for measuring quality of Semantic Web data in various dimensions are being deployed [Zaveri 2016]. However, these metrics focus on dimensions at the *dataset* level, and few are concerned with diachronic -- i.e. developing and evolving over time -- Web vocabularies. Currently, no Web data quality metric quantifies the appropriateness of changes in a *Web vocabulary version chain*, thus leaving the quality of their evolution undetermined.

So, what is the quality of the evolution processes of vocabularies in Linked Data? Are changes introduced in a revision sensible? Are current Linked Data vocabularies evolving in a predictable and coherent way? How can we approach the measurement of such evolution?

To answer these, the approach followed in this work proposes a metric for the *quality of the evolution processes of diachronic Linked Data vocabularies.* This metric is based on the assumption that state-of-the-art machine learning has reached such a level of maturity that it can provide high-quality prediction models for stable chains of datasets. Ontology evolution predictors [Stojanovic 2004] are today well understood, and have proven useful to build high-quality ontology change prediction models [Pesquita 2012].

Based on this, we define a metric to measure the quality of a Linked Data vocabulary over time. To do so, we first find optimal models of change from past versions in a vocabulary chain, using state-of-the-art machine learning tools [Pesquita 2012] and well understood ontology evolution predictors [Stojanovic 2004]. We consequently use the performance of these change models as a quality metric for vocabulary evolution. Note that this measure formalises stability of a diachronic Linked Data vocabulary. Hence our contributions with this work are manifold: (i) we define a Linked Data vocabulary evolution quality metric based on the performance of inferred optimal change models; in order to do this, (ii) we generalize from an existing change learning method for biomedical ontologies into a domain agnostic method applicable to any Linked Data vocabulary; (iii) we investigate characteristics of diachronic Linked Data vocabularies and their impact on assessing evolution quality with our method; and (iv) study the stability (or quality) of evolving

[2] See http://lov.okfn.org/

Linked Data vocabularies for 669 versions organized in 139 version chains retrieved from various Web sources. We find that 39.80% of the evaluated version chains score above 0.9, 36.10% do so between 0.5 and 0.9, and 25.10% display random predictability.

These contributions are useful in 3 different ways: first, our analysis shows that a large number of vocabularies on the Web do not change in expected and smooth ways. The effect of this is that data still described using older versions of vocabularies can have a radically different meaning that the intended one. Secondly, thanks to our quality metrics it is possible to pinpoint precisely to the versions in the chains where radical changes have happened, i.e. where the development of the vocabulary has previously been a bumpy road. This can be useful for engineers to improve the quality of the vocabularies and the data annotated with them. Finally, the models themselves can be useful to support vocabulary engineers to make if not the right modelling choices, at least choices that are coherent with the previous engineering process.

The SemaDrift Suite of Tools to Measure Semantic Drift in Ontologies

One of the applications of semantic drift is in identifying and measuring 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. For this purpose PERICLES partners developed a novel set of tools, namely the **SemaDrift software suite** to measure semantic drift in ontologies across time or versions, using text and structural similarity methods to provide valuable insights. The methods are directly applicable to any ontology originating from any domain of application (Stavropoulos et al., 2016b). The *SemaDrift Library* is a core API that provides the methods for open-source software development. The suite is complemented with two graphical user interfaces that expose these functions for domain experts and non-developers to measure drift in any domain: the *SemaDrift Protégé Plugin* (Stavropoulos et al., 2016a) and the *SemaDrift Fx desktop application*. All software modules are available under Apache License.

- **SemaDrift Library (API)**: This API written in Java is the core library that processes and parses ontology versions to extract drift metrics. It supports an array of multiple ontology versions and leverages the OWL-API library for parsing. It also provides some utilities to clients, such as to obtain the ontology hierarchies in tree-structures, to eliminate the need of re-processing models.
- **SemaDrift Protégé Plugin**: This plugin provides integration with the Protégé popular ontology creation software, providing a GUI to calculate drift. It leverages the Java SemaDrift Library to provide drift metrics for two consecutive ontology versions: one open in Protégé and a second ontology of choice. It requires a 4.* version of the Protégé application.
- **SemaDrift Fx**: This standalone desktop application enables drift measurement between two consecutive ontology versions of choice. It provides a more user-friendly GUI for leveraging the SemaDrift Library API in the JavaFx framework, which allows more visual capabilities to be added in the future (multiple ontologies, dynamic graphs, visual morphing chains).

SemaDrift has been evaluated and validated through two proof-of-concept scenarios in the domains of Web Services and Digital Preservation, these are described as follows:

SuCCESS and Drift-a-LOD Workshops

The fruitful discussions within the CoP led to organizing two workshops. The first one was **SuCCESS'16** (1st Int. Workshop on Semantic Change & Evolving Semantics) and was co-located with the SEMANTiCS conference that took place in Leipzig, Germany on 12-15 September, 2016. More than 20 participants attended the workshop, where 6 papers and a very interesting keynote presentation were featured.

As a continuation of SuCCESS'16, a second relevant workshop was co-organized by CoP member Albert Meroño Peñuela together with Laura Hollink from Centrum Wiskunde & Informatica, The Netherlands, with assistance by CoP members and PERICLES partners Sándor Darányi and Efstratios Kontopoulos. **Drift-a-LOD'16** (Detection, Representation and Management of Concept Drift in Linked Open Data) was co-located with EKAW2016 (20th International Conference on Knowledge Engineering and Knowledge Management) and took place in Bologna, Italy, on 20 November 2016. The workshop was yet another success, featuring 2 keynotes, 6 paper presentations and a special late breaking results session. All in all, the SuCCESS and Drift-a-LOD workshops resulted in 6 papers co-authored by the CoP members, 5 of which reported PERICLES results. Also, 1 of the 5 PERICLES papers was the result of collaboration between Consortium partners and non-PERICLES members of the CoP, indicating the high level of collaboration among CoP members.

Finally, based on the success of these two workshops, the Drift-a-LOD organizing team is planning the 2017 edition of the workshop, currently preparing a workshop proposal submission for ISWC'17 (16th International Semantic Web Conference).

Discussion and Future Directions

In this document we have described ongoing work that represents current research trends in the area of semantic change. In particular, the first section of this report brings together and disambiguates research relevant to semantic change and explores how the pertinent diverse terms relate to each other. We then briefly compared recent theoretical research which has touched upon the area of semantic change, starting with the role of semiotics in identifying semantic drift, with empirical studies of content and community change in relation to Tate Galleries, and finally to the study and detection of semantic drift in ontologies.

One of the applications of this work touched upon in this document is in the area of Digital Preservation (although it is by no means limited to it). For example, one well understood source of erosion in data and cultural heritage is technological changes, with hardware and software obsolescence turning what used to be computer-readable into corrupted bits e.g. due to "bit rot". However, more relevant to the work discussed in the current report is the assertion of the role of language change: with progress, new concepts have to be named and old ones' meanings shift in new directions either in the population as a whole, or in pockets of different uses; a final risk is in terms of social pressures and community changes, which also feeds back into language change [Blank, 1999; Hestrom, 1997; Schlieder, 2010].

Not that we are claiming that the problem of semantics in culture is a recent challenge: it goes back at least two millennia. To ask for the meaning of sentences and words had started in Vedic India, continued in classical Greece, and through scholasticism and different philosophical undercurrents, culminated in a great number of theories of word and sentence semantics by the 20th century. This alone hints at the fact that no single, unified theory of meaning exists to the

current day. Information theory encouraged people to believe that they understand the problem because they can mechanize the communication process, although Claude Shannon made it clear that information theory left out semantics and dealt with communication on a formal ground only [Shannon, 1948]. Worse, the very term “information retrieval” from the sixties onward reinforced misunderstandings and masked the fact that people are interested in meaningful answers by the computer, whereas the nature of semantics is still cryptic.

An early warning to the information retrieval community was the problem of inter-indexer consistency, subject to renewed interest in the World Wide Web environment [Chi, 2015], as follows. Given a test set of documents to be manually assigned keywords to signal their content, human indexers could not agree between themselves what the subject of a certain document might be and indexed it by only partly overlapping keywords. Automatic indexing [Luhn, 1957] replaced these insecurities by statistical assumptions but, as the history of evaluation in information retrieval and text categorization over the past fifty years has shown, a solution to scalability did not answer the original question. In all, inconsistent indexing limits future access to digital objects of any kind, and fluctuations in word meaning due to changes in word use or by a need to redefine concepts posits a constant danger.

All of this demonstrates that the issue of semantic change is widely relevant both in terms of applications and research fields; and this is why we believe that there is not one single approach which can answer the issue of semantic change, and as we have illustrated above, there needs to be an integrated approach bringing together advances in semiotics, computational semantics, community analysis and ontologies plus the Semantic Web. In the following, we draw attention to some fundamental issues which must be addressed if we are to better understand the relationships, problems, implications, and meaning of semantic change in computer science and related areas. We firstly present three challenges.

The first is that the gap between disciplines attempting to answer this research question is still too large. Even though the common interest and study of semantic change brings together researchers and practitioners from many fields, such as databases, ontology-based data access, data mining, knowledge organization systems and linguistics, these communities have different perspectives and approaches for defining and tackling the problem of semantic change, with these often tailored to their specific research field. As we have begun to do in the introduction to this document, we believe that it is necessary to work towards common spaces and terminologies to define the challenge of semantic drift in a broader way.

A second - related - challenge is the lack of (formal) vocabularies by which to describe and discuss semantic change. This is true for traditional library science paradigms (e.g. how to express that a category has split in two at a certain point in time), but also in more formal settings, like the Semantic Web, where to the best of our knowledge there is no ontology to formally describe the semantics of phenomena like conceptual change or semantic drift. We suppose that this could easily result in the future in reasoning services implemented by query engines which transparently take semantic drift into account for users interested in data that inherently and fundamentally changes over time. Indeed, this would be a suitable application of some of the research described in this document.

A third one is the lack of large scale, time-stamped test datasets to study semantic change under laboratory standards. For example, most researchers of the Semantic Web use the Dynamic Linked Data Observatory, CommonCrawl, or data from the Internet Archive to perform their experiments. However, the availability of datasets which combine schema information, instance information, natural language, machine-readable statements, and which also contain rich metadata, such as their collection timestamps and provenance, is still very scarce. Additionally,

the interval at which such crawls must be performed (e.g. once an hour, a day, a week, a year?) depending on what analyses are to be done is still not well understood and leads to further scarcity of appropriate data.

Turning now to opportunities, the first, key opportunity to go forward is the chance to combine semantics with statistics. In the past, this has been tackled from two perspectives, distributional semantics and knowledge representation. To some extent, these can be viewed as embodying the inductive versus deductive reasoning dichotomy found in Artificial Intelligence (such as statistics and machine learning contrasted with logic and symbolic reasoning). However recent advances such as Semantic Statistics (and its RDF Data Cube standard [Cyganiak et al., 2013]) suggest that combining both approaches is possible, and desirable (for the most part). For example, the classic distinction between *intension* and *extension* of concepts (i.e. their formal semantic definition, and their composition in terms of instances, respectively) are better tractable if methods allow for both statistical and symbolic representations of what these mean. We propose that this extends to the study of semantic change in these concepts.

The second, and as far as we can tell, untapped opportunity in this area is in using version control systems for understanding change in data. PROV-O, the W3C standard for describing provenance, is acquiring great momentum in Semantic Web applications that use it to describe, at a very fine grained level, the workflows performed on data. We can see similarities in the use of modern version control systems, such as git; application containers, such as docker; and popular source code repositories as *data* repositories, such as GitHub, and note the potential of these infrastructures to better (more precisely, more semantically, and at a higher scale) describe changes not only in software, but also on individual datasets. We see that these can be readily exploited as a source of knowledge on the semantic drift.

Drawing all this together, in our collaborations to date we have begun to address some of the challenges of semantic change, but our conclusion makes it clear that there are still many opportunities for research and applications. Although there are great challenges in the study of semantic change, there are also great opportunities for its study from statistical techniques and availability of data, which mean that researchers today and in the future may be able to make unprecedented progress in this area. We look forward to learning about, and being a part of, these discoveries.

References

Aristotle. (2015). On Interpretation Translated by E. M. Edghill https://ebooks.adelaide.edu.au/a/aristotle/interpretation/

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), pp. 34-43.

Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. Historical semantics and cognition, 13, pp. 61-89.

Blei, David. (2012). Topic modeling and digital humanities. Journal of Digital Humanities 2 (1), pp. 8-11.

Bouissac, P. (2004). Saussure's legacy in semiotics. In Sanders, C. The Cambridge Companion to Saussure. Cambridge University Press, pp. 240-260.

CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-M-2

Chi, Y. (2015). A Complete Assessment of Tagging Quality: A Consolidated Methodology. JASIST, 66(4), pp. 798--817.

Condamines, A., Galarreta, D., Perrussel, L., Rebeyrolle, J., Rothenburger, B., Viguier-Pla, S. (2003). Tools and methods for knowledge evolution measure in space project, Proceedings of IAF-IAA.6.2.7, Bremen, Germany.

Cyganiak, R., Reynolds, D., Tennison, J. (2013). The RDF Data Cube Vocabulary. World Wide Web Consortium (W3C). http://www.w3.org/TR/vocab-data-cube/.

Darányi, S., Wittek, P., Konstantinidis, K., Papadopoulos, S., and Kontopoulos, E. (2016). Physics as a metaphor to study semantic drift. In Proceedings of SUCCESS'16, 1st International Workshop on Semantic Change & Evolving Semantics, Vol. 1695.

Everaert-Desmedt, N. (2011). Peirce's Semiotics. In Hébert, L. (Ed.), Signo [online], Rimouski (Quebec), http://www.signosemio.com/peirce/semiotics.asp.

Falk, J.S. (2004). Saussure and American linguistics. In Sanders, C. The Cambridge Companion to Saussure. Cambridge University Press, pp. 107-123.

Flake, G. W., Lawrence, S., Giles, C.L. (2000). Efficient Identification of Web Communities. In the proceedings of the 6th ACM SIGKDD International Conference on Knowledge discovery and data mining, Boston, Massachusetts, USA, ACM, pp. 150-160.

Galarreta, D. (2007). A Contribution to a Semiotic Approach of Risk Management. In Project Management and Risk Management. In Charrel PJ & Galarreta D. (editors). Complex Projects. Studies in Organizational Semiotics Springer.

Galarreta, D. (2010). A Semiotic Approach of contexts for Pervasive systems. 12th International Conference on Informatics and Semiotics in Organisations IFIP WG8.1 Working Conference University of Reading, UK.

Galarreta, D. (2013). Are things, objects? A semiotic contribution to the Web of Things. Web of Things, People and Information Systems. 14th International Conference on Informatics and Semiotics in Organisations (ICISO 2013). IFIP WG8.1 Working Conference, Stockholm, Sweden.

Greimas A. J. and Courtés J. (1983). Semiotics and Language: an Analytical Dictionary, Translated by Christ Larry, Patte Daniel, Lee James, McMahon Edward II, Phillips Gary, & Rengstorf Michael. Bloomington, Indiana University Press.

Habert, B. & Zweigenbaum, P. (2002). Contextual acquisition of information categories: what has been done and what can be done automatically? In B. Nevin (Ed.), Volume 2. Computability of language and computer applications. Amsterdam / Philadelphia: John Benjamins.

Hedstrom, M. (1997). Digital preservation: a time bomb for digital libraries. Computers and the Humanities 31, 3, pp. 189–202.

Hjelmslev, L. (1961). Prolegomena to a theory of language. Madison: University of Wisconsin Press.

Kleinberg, J. (1997). Authoritative Sources in a Hyperlinked Environment Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998, and as IBM Research Report RJ 10076, May 1997.

Lotman, J. (2009). The explosion and the Culture. Walter de Gruyter GmbH & Co. KG, D-10785 Berlin.

Luhn, H.P. (1957). A statistical approach to mechanised encoding and searching of library information. IBM Journal of Research and Development, 1, pp. 309-317.

Michael Martin, Martí Cuquet, and Erwin Folmer (Eds.) (2016). Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), Leipzig, Germany, September 12-15, 2016. Available from: http://ceur-ws.org/Vol-1695/ (URN: urn:nbn:de:0074-1695-3)

Norris, C. (2004). Saussure, linguistic theory and philosophy of science. In Sanders, C. The Cambridge Companion to Saussure. Cambridge University Press. pp. 219-239.

Pesquita, C., Couto, F.M. (2012). Predicting the Extension of Biomedical Ontologies. PLoS Computational Biology 8(9), e1002630, doi:10.1371/journal.pcbi.1002630

Rajman, M. & Besançon, R. (1998). Text Mining – Knowledge Extraction from unstructured textual data, Proceedings of 6th Conference of International Federation of Classification Societies (IFCS-98), Roma (Italy), July 1998, pp. 473-480

Rothenburger, B. (2002). A Differential Approach for Knowledge Management, ECAI workshop on Machine Learning and Natural Language Processing for Ontology Engineering, Lyon.

Schlieder, C. (2010). Digital heritage: Semantic challenges of long-term preservation. Semantic Web, 1(1, 2), pp. 143-147.

Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, 27 (July and October), pp. 379–423, 623–656.

Stavropoulos, T.G., Andreadis, S., Kontopoulos, E., Riga, M., Mitzias, P., Kompatsiaris, I. (2016). SemaDrift: A Protégé Plugin for Measuring Semantic Drift in Ontologies. Detection, Representation and Management of Concept Drift in Linked Open Data (Drift-a-LOD) co-located with the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW), At Bologna, Italy.

Stavropoulos, T.G., Andreadis, S., Riga, M., Kontopoulos, E., Mitzias, P., Kompatsiaris, I. (2016). A Framework for Measuring Semantic Drift in Ontologies. 1st Int. Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), co-located with the 12th European Conference on Semantics Systems (SEMANTiCS'16), at Leipzig, Germany, Volume: CEUR Workshop Proceedings Vol-1695.

Stojanovic, L. (2004). Methods and Tools for Ontology Evolution.Ph.D. Thesis, University of Karlsruhe.

Ultsch, A. (2005). Clustering with SOM: U* c. In Proceedings of WSOM-05, 5th Workshop on Self-Organizing Maps, Paris, France, September 2005. pp. 75–82.

Wang, R., Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management of Information Systems, March 1996, 12(4), pp. 5-33.

Wittek, P., Gao, S. C., Lim, I. S., and Zhao, L. (2015). Somoclu: An efficient parallel library for self-organizing maps. At http://arxiv.org/pdf/1305.1422.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. (2016). Quality Assessment for Linked Data: A Survey. Semantic Web – Interoperability, Usability, Applicability, 7(1), pp. 63-93.

This project has received funding from the European Union's Seventh Framework Programme
for research, technological development and demonstration under grant agreement no 601138.