Confidence Management for Learning Ontologies from Dynamic Web Sources

Gerhard Wohlgenannt[1], Albert Weichselbraun[2], Arno Scharl[3] and Marta Sabou[3]
[1]*Vienna University of Economics and Business, Augasse 2-6, 1090 Wien, Austria*
[2]*University of Applied Sciences Chur, Ringstrasse 34, 7004 Chur, Switzerland*
[3]*MODUL University Vienna, Am Kahlenberg 1, 1190 Wien, Austria*

Keywords: Ontology Dynamics, Confidence Management, Ontology Learning, Evidence Integration, Trend Detection.

Abstract: Dynamic environments require effective update mechanisms for ontologies to incorporate new knowledge. In this position paper we present a dynamic framework for ontology learning which integrates automated learning methods with rapid user feedback mechanism to build and extend lightweight domain ontologies at regular intervals. Automated methods collect evidence from a variety of heterogeneous sources and generate an ontology with spreading activation techniques, while crowdsourcing in the form of Games with a Purpose validates the new ontology elements. Special data structures support dynamic confidence management in regards to three major aspects of the ontology: (i) the incoming facts collected from evidence sources, (ii) the relations that constitute the extended ontology, and (iii) the observed quality of evidence sources. Based on these data structures we propose trend detection experiments to measure not only significant changes in the domain, but also in the conceptualization suggested by user feedback.

1 INTRODUCTION

Ontologies are the backbone of the Semantic Web. Due to the highly dynamic characteristics of many domains, it is necessary to keep ontologies up-to-date to ensure their usefulness. A common definition of ontology evolution is the "timely adaptation of an ontology to the arising changes and the consistent management of these changes" (Haase and Stojanovic, 2005). In this position paper we focus on the timely adaptation of an ontology learning framework to changes arising in a heterogeneous set of evidence sources in dynamic domains, and touch the consistent management of changes only implicitly. In contrast to other research projects investigating ontology dynamics, we are not only interested in keeping lightweight ontologies up-to-date, but especially in the management and the fine-grained (regarding the evidence sources and time periods) analysis of sources of change, in the detection of trends and patterns, and in making the reasons for change traceable.

We propose an ontology learning system that traces confidence dynamics on the level of (i) evidence sources, (ii) the results of the ontology learning algorithms, and (iii) the quality of input sources. For these tasks we use specialized matrix-based data structures to capture dynamic confidence aspects. Ontology dynamics literature distinguishes the management of changes performed by the user (Noy et al., 2006; Vrandecic et al., 2005) and systems that focus on learning and updating ontologies dynamically (Alani et al., 2006; Novacek et al., 2007). The proposed framework belongs to the second category, but tightly integrates user input into the learning cycle to validate the results of the learning algorithms and optimize the performance of the algorithms over time.

Section 2 provides an overview of related work. Section 3 then briefly introduces the ontology learning and confidence dynamics framework. The envisioned trend detection experiments and the formal description of the matrix-based data structures follow in Section 4. A discussion of contributions and future work concludes the paper in Section 5.

2 RELATED WORK

This paper focuses on dynamic aspects of an ontology learning system, and also touches the integration of user feedback into the learning process. Ontology learning refers to the (semi-)automatic generation of

Wohlgenannt G., Weichselbraun A., Scharl A. and Sabou M..
Confidence Management for Learning Ontologies from Dynamic Web Sources.
DOI: 10.5220/0004111101720177
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 172-177
ISBN: 978-989-8565-30-3

ontologies, aiming to reduce the effort involved in the expensive task of manual ontology construction. Ontology learning typically involves corpus linguistics to extract semantically similar terms to form clusters of meaning (Wohlgenannt et al., 2009), and approaches such as the lexico-syntactic patterns or rules inspired by Hearst (Hearst, 1992).

Similarly to ontology learning, most ontology dynamics approaches rely on a single source of evidence to derive changes, usually text-based sources – as domain text is an abundant resource. Tools such as EVOLVA (Zablith et al., 2010), SPRAT (Maynard et al., 2009) and FLOR (d'Aquin et al., 2008) belong to this category. In contrast to other approaches which focus on just one aspect, EVOLVA aims to cover both to the adaptation of ontologies and the management of changes (Zablith et al., 2009).

Except for initial efforts in the RELExO framework (Maynard and Aswani, 2010), which is a tool for semi-automatic ontology refinement based on Formal Concept Analysis, there are no ontology dynamics tools to our knowledge that utilize data from multiple and heterogeneous sources. Ontology dynamics triggered concurrently by heterogeneous sources remain an open research question. Our proposed system flexibly integrates a large number of evidence sources and methods (based on unstructured, structured and social data) into the learning algorithms.

Dataset dynamics is also a novel research trend in the linked data community. In contrast to our work, which focuses on domain ontologies and only integrates and interprets data relevant to the domain, the linked data community monitors the evolution of the linked data cloud as a whole (Popitsch and Haslhofer, 2010; Umbrich et al., 2010).

We propose to use an evidence confidence matrix (see Section 4.1), for the integration of evidence from multiple sources, as well as for ontology evolution and trend detection experiments. Cimiano et al. (Cimiano et al., 2009) already suggested the use of confidences in Text2Onto. They calculate a confidence for each learned object and store it in a Probabilistic Ontology Model (POM) to allow more sophisticated ways of user interaction and visualization. Cimiano's work has focused on the change management aspect in ontology evolution, and not on trend detection or advanced pattern analysis, as we aim to.

Flouris et al. (Flouris et al., 2006) use a different approach by applying the work in belief revision theory to the field of ontology evolution. They stress that automatic and computer-based evolution instead of manual change management is necessary and desirable in many contexts, a statement which resonates well with the approach proposed in this paper.

For the ongoing evaluation of the learned ontologies we use crowdsourcing and Games with a Purpose (GWAP) (Ahn and Dabbish, 2008) to evaluate and optimize the influence of respective evidence sources and extraction methods within the learning algorithms. GWAPs typically solve computational problems that are easy for humans but hard to tackle for machines, and motivate users with the incentive schemes of games. Existing GWAPs in the field of ontology learning like OntoPronto (Siorpaes and Hepp, 2008) and Guess What ?! (Markotschi and Voelker, 2010) do not offer a similar integration and feedback loop between the games and the learning algorithm.

3 THE ONTOLOGY LEARNING FRAMEWORK

This section describes the ontology learning framework which is the foundation for the trend detection experiments proposed in the paper. We will focus on the components directly related to ontology dynamics and confidence management, but also touch the other aspects for the sake of comprehensibility and completeness.

Figure 1 gives a graphical overview of the system and its workflow. The ontology learning process starts from a typically small seed ontology (a few concepts and relations). The seed ontology is static, thereby making a comparison of the learned parts and their dynamics more meaningful. The first step in the process is the collection of evidence data for the given seed concepts. We collect evidence data, i.e., terms related to the seed concepts, from a variety of sources (e.g., domain text) with a variety of methods (e.g., co-occurrence statistics or Hearst pattern). Section 4.1 gives more information about evidence sources. To track the dynamics of the domain the documents and other underlying data originate from the period of time in question. We will compare weekly and monthly intervals when building new ontologies.

All evidence is collected in a graph data structure called the *semantic network*. The semantic network includes the seed concepts and all extracted terms connected to the seed concepts with typed links. The type of links reflects the evidence source and the observed strength of relation. The evidence confidence matrix (ECM) (see Section 4.1) constitutes an additional representation of the semantic network – the ECM also includes the dimension of time for use in subsequent trend detection experiments.

The semantic network is then transformed into a spreading activation network. Spreading activation is a search technique inspired by the human brain and

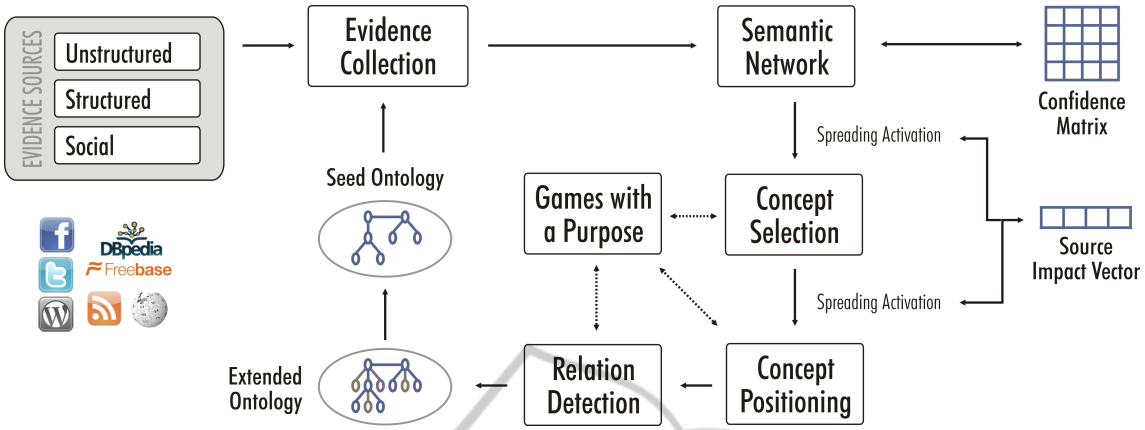

Figure 1: Ontology Extension Architecture System Diagram.

its cognitive models. The source impact vector (SIV, see Section 4.2), which reflects user-generated confidence into quality of evidence sources, contributes to the link weights in the spreading activation network, making high quality sources have more impact on the resulting ontology. The SIV evolves over time according to user feedback harvested from GWAPs. In the next step spreading activation detects the (e.g., 20) most relevant concept candidates to be included in the extended ontology. A round of user feedback via GWAPs (see section 3) accepts or rejects these new concept candidates.

Another run of spreading activation then positions the new concepts in the ontology, i.e., it selects the most related concept from the seed ontology to connect the new concept to. Every new concept gets connected to exactly one seed concept. User feedback confirms the position or suggests a better one (see below). The spreading activation algorithm outputs the relation strength between every seed concept and every new concept. We capture this learning algorithm-based confidence information in a relation strength matrix (see Section 4.3) for trend detection experiments and to support user selection of relations if needed (see below).

Finally, relation type detection with techniques to identify taxonomic relations as described in (Liu et al., 2005) and methods to label non-taxonomic relations (Weichselbraun et al., 2010b) conclude the learning process.

The ontology learning system works in an iterative fashion, where the newly generated extended ontology functions as seed ontology in the next learning cycle. The ontology building process stops when the ontology reaches the desired size and granularity level.

The underlying ontology learning components have been thoroughly evaluated and published in (Liu et al., 2005), (Weichselbraun et al., 2010a) and (Weichselbraun et al., 2010b). The framework presented in this paper extends this work by introducing dynamic data structures for confidence management, the source impact vector (SIV) and through mechanisms for its adaptation according to user feedback, as we detail next.

Low-overhead Forms of User Feedback

It is evident from the system description (Figure 1) that rapid user feedback captured for example through Games with a Purpose (GWAPs) is a key component. We will extend our game portfolio with new Facebook-based games similar to the already successfully deployed *Sentiment Quiz* (Rafelsberger and Scharl, 2009). The Sentiment Quiz addressed major challenges of GWAPs such as how to attract and retain players, how to ensure the generation of high quality data, and how to effectively aggregate results.

In the first game (or first task in an integrated game) players confirm if a new concept candidate is relevant to the domain. If the concept is not relevant (a certain level of agreement among users is needed), the user evaluation terminates and the concept is pruned. Another game indicates if the connection between the seed concept and the new concept is correct. If it is not correct the game lets players suggest the appropriate seed concept to connect to. For this task, the GWAP ranks the seed concepts to select from by the relation strength from the relation strength matrix.

To use the scarce resource of user feedback optimally, we do not re-evaluate concepts or relations in subsequent intervals, where the number of skipped intervals depends on inter-player agreement on the respective item.

4 DYNAMIC CONFIDENCE MANAGEMENT

This section introduces the planned confidence management and trend detection experiments as well as three data structures that are crucial for tracking temporal changes and trends: the evidence confidence matrix (Section 4.1), the source impact vector (Section 4.2) and the relation strength matrix (Section 4.3).

4.1 Evidence Confidence Matrix

As already mentioned, our confidence management and ontology dynamics approach is data-driven and relies on evidence extracted from heterogeneous sources, which is then collected into a graph data structure and also the evidence confidence matrix (ECM). We distinguish three basic types of evidence sources: unstructured, social and structured. We have repeatedly and successfully used the webLyzard suite of Web mining tools (www.weblyzard.com) for generating high-quality domain corpora. The result of the mirroring and domain detection process is a number of corpora for the respective period: e.g., a US news media corpus, a British news media corpus, a Fortune 1000 corpus, etc. We will select the level of granularity (e.g., all news media from a country vs. single news media) according to the number of documents available. The algorithms apply term extraction methods such as co-occurrence statistics and Hearst patterns to the corpora (Liu et al., 2005).

To collect evidence from social media we use the TagInfoService interface of the easy Web Retrieval Toolkit (www.semanticlab.net/index.php/eWRT) to get related terms for the label of a seed concept from sources such as Twitter, Flickr and Del.icio.us (Weichselbraun et al., 2010a). Querying the DBpedia SPARQL endpoint with the seed concept labels and specific properties such as *dcterms:subject*, or the use of Scarlet (Sabou et al., 2008), constitute the structured data evidence sources.

The total number of evidence sources relates to the number of extraction methods multiplied by the number of corpora (for text sources). Examples of evidence sources (and *evidences*) for the seed concept "greenhouse effect" in a climate change ontology are:

- Sentence-level Co-occurrence in Australian News Media
  (*"greenhouse effect"* $\xrightarrow{es_i}$ *"carbon dioxide"*),
- Related tags from Twitter
  (*"greenhouse effect"* $\xrightarrow{es_i}$ *"petrol"*),
- DBpedia-query `dcterms:subject`
  (*"greenhouse effect"* $\xrightarrow{es_i}$ *"kyoto protocol"*), ...

The *semantic network* connects collected terms with the seed ontology via directed weighted links. The *ECM* is an additional form to store evidence data and includes a temporal dimension. The system generates ECMs to represent all possible relations between a seed concept C_s and a candidate concept C_c. The ECM contains the observed connection strength between the two concepts for all evidence sources. The second dimension is the temporal one. Equation 1 shows a ECM for seed concept C_s and candidate concept (term) C_c with the dimensions evidence source *es* and time *t* and the corresponding confidence values c_{es_i,t_j}.

$$ECM_{C_s,C_c} = \begin{bmatrix} c_{es_1,t_1} & c_{es_1,t_2} & \cdots & c_{es_1,t_m} \\ c_{es_2,t_1} & c_{es_2,t_2} & \cdots & c_{es_2,t_m} \\ & \vdots & & \\ c_{es_n,t_1} & c_{es_n,t_2} & \cdots & c_{es_n,t_m} \end{bmatrix} \quad (1)$$

The ECM supports a number of trend detection experiments such as: (i) observe patterns in the evidence data between two concepts on a fine-grained per evidence source basis, (ii) compare the occurrence patterns for a new concept/term among multiple seed concepts, to see the evidence strength between those and how it evolves, (iii) aggregate evidence sources (e.g., all news media sources) and trace such patterns on a higher level, (iv) compare (aggregated) sources to see which types of sources promote a new concept at what point in time, (v) study the characteristics of evidence sources themselves.

4.2 Source Impact Vector

The source impact vector (SIV) contributes to the weights in the spreading activation network because its values reflect user generated confidence (quality) and suggested impact of an evidence source in the learning process. Evidence sources which tend to provide low quality terminology should have low impact, and vice versa. The initial settings stem from heuristics and metrics such as Google PageRank.

User feedback yields an optimized SIV. Neural network learning techniques such as backpropagation allow adjusting the weights of the spreading activation network to the users' perception of the domain and set the corresponding values to optimize the source impact vector. To prevent overfitting to a specific ontology, we plan to keep the SIV within predefined intervals.

Equation 2 presents a SIV for a point in time t_i. It contains the impact values I for evidence sources es_j.

$$SIV_{t_i} = \begin{bmatrix} I_{es_1} & I_{es_2} & \cdots & I_{es_n} \end{bmatrix} \quad (2)$$

The dynamic adaptation of the source impact vectors according to user feedback is novel. By introducing a temporal dimension and storing source impact values over time, we obtain a source impact *matrix*. By exploring this temporal aspect, we can study the trends in source impact: (i) patterns in the quality of single evidence sources in a specific domain, (ii) patterns in aggregated views on source impact (e.g., all social media sources), (iii) patterns across domains to explore the suitability of specific sources for a certain domain.

The *concept selection* and *concept positioning* phases have different characteristics, and, therefore, an evidence source might be better suited to one than the other. Therefore we plan to experiment with using two separate SIVs corresponding to support values for these two tasks.

4.3 Relation Strength Matrix

Spreading activation in the process of *concept positioning* yields the relation strength, i.e., the learning algorithm-based confidence, between all seed and new concepts in the extended ontology. Additionally to the application of relation strength values as described in Section 3, these data also provide the base for interesting trend detection experiments when studied from a temporal viewpoint. The relation strength matrix (RSM) as given in Equation 3 shows the relation strength values rs for any relation r_{ij} between seed concept i with new concept j.

$$RSM = \begin{bmatrix} rs_{r_{11}t_1} & rs_{r_{11}t_2} & \cdots & rs_{r_{11}t_n} \\ rs_{r_{12}t_1} & rs_{r_{12}t_2} & \cdots & rs_{r_{12}t_n} \\ \vdots & & & \vdots \\ rs_{r_{21}t_1} & rs_{r_{21}t_2} & \cdots & rs_{r_{21}t_n} \\ \vdots & \vdots & & \vdots \\ rs_{r_{mn}t_1} & rs_{r_{mn}t_2} & \cdots & rs_{r_{mn}t_n} \end{bmatrix} \quad (3)$$

The RSM can give interesting insights into the *dynamics* of the domain. A permanent change in learning algorithm-based confidence (relation strength) between concepts suggests the meaning of a concept (especially in taxonomic relations), or its relation to other concepts has changed in the outside world. For example, the meaning of the concept "energy source" might have changed when the predominant relation between "energy source" and "fossil fuel" gets weaker over time, and the relation to "alternative energy" intensifies.

It will then be interesting to examine if user feedback is in line with shifts in terminology as computed by the algorithms, or if users do not perceive these domain changes. Conversely, if user feedback differs from the past, but the underlying evidence is unchanged, then a *shift in user perception and conceptualization*, but not a change in the domain itself, is indicated.

We will also take an overall view on the characteristics of relations, e.g., is it common in the domain to have dominant and unambiguous relations between seed and new concepts, or are there mostly connections of similar strength from a seed concept to many new concepts. Experiments will point out if (parts of) a domain are getting more fuzzy and connected or if reverse effects occur.

5 DISCUSSION

This position paper presents a framework for learning lightweight ontologies and studying dynamic confidence values assigned to its various elements. In regular - e.g., weekly or monthly - intervals, the framework generates ontologies starting from a small seed ontology by integrating heterogeneous evidence extracted in the respective period. Spreading activation algorithms detect new concepts and position them into the ontology. User feedback with games with a purpose validates new concepts and relations, and provides data to manage and optimize the confidence in specific evidence sources. The main contributions of the approach are (i) special data structures that facilitate powerful and fine-grained dynamic confidence management of ontological elements, (ii) trend detection in those data structures, referring to input data (evidence) and the resulting verified ontologies, (iii) the integration of rapid user feedback cycles, (iv) experiments to measure ontology dynamics characteristics such as change in a domain (because the "world" has changed) or change in conceptualization. Future work will focus on the execution of the proposed confidence management and trend detection experiments.

ACKNOWLEDGEMENTS

The work presented in this paper was developed within DIVINE (www.weblyzard.com/divine), a project funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.ffg.at/fit-it).

REFERENCES

Ahn, L. and Dabbish, L. (2008). Designing Games with a Purpose. *Communications of the ACM*, 51(8):58–67.

Alani, H., Harris, S., and O'Neil, B. (2006). Winnowing ontologies based on application use. In Sure, Y. and Domingue, J., editors, *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 185–199. Springer.

Cimiano, P., Maedche, A., Staab, S., and Voelker, J. (2009). Ontology learning. In Staab, S. and Rudi Studer, D., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 245–267. Springer Berlin Heidelberg.

d'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. (2008). Towards a New Generation of Semantic Web Applications. *IEEE Intelligent Systems*, 23(3):20 – 28.

Flouris, G., Plexousakis, D., and Antoniou, G. (2006). Evolving ontology evolution. In Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., and Stuller, J., editors, *SOFSEM*, volume 3831 of *Lecture Notes in Computer Science*, pages 14–29. Springer.

Haase, P. and Stojanovic, L. (2005). Consistent evolution of owl ontologies. In *Proceedings of the Second European Semantic Web Conference, Heraklion, Greece*, pages 182–197.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545, Nantes, France.

Liu, W., Weichselbraun, A., Scharl, A., and Chang, E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management*, 0(1):50–58.

Markotschi, T. and Voelker, J. (2010). Guess What?! Human Intelligence for Mining Linked Data. In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at the International Conference on Knowledge Engineering and Knowledge Management (EKAW)*.

Maynard, D. and Aswani, N. (2010). Bottom-up Evolution of Networked Ontologies from Metadata (NeOn Deliverable D1.5.4).

Maynard, D., Funk, A., and Peters, W. (2009). SPRAT: A Tool for Automatic Semantic Pattern-based Ontology Population. In *International Conference for Digital Libraries and the Semantic Web (ICSD-2009)*, Trento, Italy.

Novacek, V., Laera, L., and Handschuh, S. (2007). Semi-automatic integration of learned ontologies into a collaborative framework. In *Proceedings of the International Workshop on Ontology Dynamics (IWOD 2007) at ESWC 2007, Innsbruck, Austria*.

Noy, N. F., Chugh, A., Liu, W., and Musen, M. A. (2006). A framework for ontology evolution in collaborative environments. In Cruz, I. F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 544–558. Springer.

Popitsch, N. and Haslhofer, B. (2010). Dsnotify: handling broken links in the web of data. In Rappa, M., Jones, P., Freire, J., and Chakrabarti, S., editors, *WWW*, pages 761–770. ACM.

Rafelsberger, W. and Scharl, A. (2009). Games with a purpose for social networking platforms. In *HT '09: Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 193–198, New York, NY, USA. Association for Computing Machinery.

Sabou, M., d'Aquin, M., and Motta, E. (2008). Exploring the Semantic Web as Background Knowledge for Ontology Matching. *Journal on Data Semantics*, XI.

Siorpaes, K. and Hepp, M. (2008). OntoGame: Weaving the semantic web by online games. In Bechhofer, S., Hauswirth, M., Hoffmann, J., and Koubarakis, M., editors, *5th European Semantic Web Conference (ESWC)*, volume 5021, pages 751–766. Springer.

Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., and Decker, S. (2010). Towards dataset dynamics: Change frequency of linked open data sources. In Bizer, C., Heath, T., Berners-Lee, T., and Hausenblas, M., editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Vrandecic, D., Pinto, H. S., Tempich, C., and Sure, Y. (2005). The diligent knowledge processes. *J. Knowledge Management*, 9(5):85–96.

Weichselbraun, A., Wohlgenannt, G., and Scharl, A. (2010a). Augmenting lightweight domain ontologies with social evidence sources. In Tjoa, A. M. and Wagner, R. R., editors, *9th International Workshop on Web Semantics, 21st International Conference on Database and Expert Systems Applications (DEXA 2010)*, pages 193–197, Bilbao, Spain. IEEE Computer Society Press.

Weichselbraun, A., Wohlgenannt, G., and Scharl, A. (2010b). Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data & Knowledge Engineering*, 69(8):763–778.

Wohlgenannt, G., Weichselbraun, A., and Scharl, A. (2009). Integrating Structural Data into Methods for Labeling Relations in Domain Ontologies. In *20th International Workshop on Database and Expert Systems Application (DEXA-2009); 8th International Workshop on Web Semantics*, pages 94–98, Austria.

Zablith, F., d'Aquin, M., Sabou, M., and Motta, E. (2010). Using Ontological Contexts to Assess the Relevance of Statements in Ontology Evolution. In Cimiano, P. and Pinto, H. S., editors, *Proc. of EKAW - Knowledge Engineering and Management by the Masses*, volume 6317 of *Lecture Notes in Computer Science*, pages 226–240. Springer.

Zablith, F., Sabou, M., d'Aquin, M., and Motta, E. (2009). Ontology evolution with evolva. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., and Simperl, E. P. B., editors, *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 908–912. Springer.