Towards a Multi-level Approach for the Maintenance of Semantic Annotations

Silvio Domingos Cardoso[1,2], Chantal Reynaud-Delaître[2], Marcos Da Silveira[1], Ying-Chi Lin[3], Anika Groß[3], Erhard Rahm[3] and Cédric Pruski[1]

[1]*LIST, Luxembourg Institute of Science and Technology, 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg*
[2]*LRI, University of Paris-Sud XI, Orsay, France*
[3]*Institute of Computer Science, Universität Leipzig, P.O. Box 100920, 04009 Leipzig, Germany*

Keywords: Semantic Annotation, Annotation Maintenance, Ontology Evolution, Life Sciences.

Abstract: Semantic annotations are often used to enrich documents as clinical trials and electronic health records. However, the usability of these annotations tends to decrease over time due to the evolution of the domain ontologies. The maintenance of these annotations is critical for tools that exploit them (e.g., search engines and decision support systems) in order to assure an acceptable level of performance. Despite the recent advances in ontology evolution systems, the maintenance of semantic annotations remains an open problem. In this paper, we introduce, based on previous experiments, the main components of a multi-level approach towards the automatic maintenance of semantic annotations. We further provide examples for strengthening our proposal.

1 INTRODUCTION

The use of Knowledge Organization Systems (KOS) (Hodge, 2000), such as classification schemes, controlled terminologies, thesauri or ontologies in the medical field to annotate medical data is gaining interest over the last years (Gimenez et al., 2012; Funk et al., 2014; Yimam et al., 2016). Usually, KOS elements are used to annotate documents such as clinical reports or medical images in order to make their semantics explicit for humans and software applications. The KOS entities (concepts, properties, relationships, etc.) are associated with documents producing semantic annotations (Da Silveira et al., 2015). This process is commonly made by humans or automatic annotators and brings many benefits for end users such as, enhancing the retrieval of relevant information for decision support or improving semantic interoperability between systems (Uren et al., 2006).

However, the dynamic nature of medical knowledge forces to continuously revise KOS content. Thus, semantic annotations based on previous versions of the KOS can be impacted and loose their validity. Therefore, mechanisms to adapt these impacted annotations to the new version of KOS are required. In our previous work (Groß et al., 2012; Cardoso et al., 2016), we have shown a strong correlation between the modification of KOS elements and the modification of semantic annotations. We also manage to categorize the various evolution that can affect KOS and associate these changes with modifications of elements defining annotations.

In the literature, three families of approaches dealing with annotation maintenance can be found. The first one addresses the problem of automatic detection of inconsistent annotations (Eilbeck et al., 2009; Qin and Atluri, 2009; Köpke and Eder, 2011; Zavalina et al., 2015). However, mechanisms to support the correction of impacted annotations are not proposed. The second family of approach put the focus on the automatic detection and manual correction of invalid annotations (Maynard et al., 2007; Auer and Herre, 2007; Burger et al., 2010; Abgaz, 2013). However, these approaches only consider basic ontology changes, e.g., deletion and addition of concepts in ontology while more complex changes are important to consider and requires human intervention to perform the maintenance which is hardly applicable in the medical domain by virtue of the huge amount of annotations to adapt. Last, the most advance works implement an automatic correction of the annotation (Luong and Dieng-Kuntz, 2006; Tissaoui et al., 2011;

Cardoso S., Reynaud-DelaÃőtre C., Da Silveira M., Lin Y., GroÃ§ A., Rahm E. and Pruski C.
Towards a Multi-level Approach for the Maintenance of Semantic Annotations.
DOI: 10.5220/0006230104010406
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 401-406
ISBN: 978-989-758-213-4

Park et al., 2011; Frost and Moore, 2014). This is mostly done based on reasoning techniques which rely of the logic formalism of the KOS. However, as medical KOS are expressed using lightweight description logic, these technique must be adapted.

The literature review highlights that there is no annotation maintenance/adaptation framework able to cope with the specificity of the medical domain e.g., size of the KOS, amount of annotations. In this paper we discuss the foundation of a (semi-)automatic approach to manage semantic annotations when their underlying KOS evolve over time without re-annotating the documents. We further justify our ideas based on previous experiments and examples.

We structure the remainder of this paper as follows: In Section 2 we define the annotation model used throughout this paper. Section 3 introduces our ideas towards the (semi-)automatic maintenance of semantic annotations. In Section 4 we conclude the paper by outlining future work.

2 ANNOTATION MODEL

The development of novel annotation maintenance approaches requires an appropriate annotation model covering evolution and quality aspects for annotations. We refer to our previously described annotation model (Groß et al., 2012; Cardoso et al., 2016) and give a brief overview of the main aspects.

A single annotation is defined as $a = (i, c, \{q\})$ where an instance item $i \in I_u$ is annotated with an ontology concept $c \in ON_v$, and a set of quality indicators $\{q\} \in Q$. An instance might be an electronic health record (EHR) or a question item from a case report form (CRFs) as used within clinical trials. In general, a concept can be used to annotate many items and an item might be annotated with several concepts. Different quality indicators can be used to retain quality, reliability and provenance information for each annotation, e.g. by attaching numerical confidence values, categorical ratings or evidence codes (Groß et al., 2009). Note that the quality of automatically generated annotations can vary significantly depending on the used methods, tools and their configurations (Funk et al., 2014).

Both, instance data and ontologies, underlie continuous changes. Hence, we denote I_u as an instance in the version u and ON_v is an ontology in the version v. In this proposition paper, we focus on maintaining annotations due to evolution of ontology. We include further elements in the annotation model to better trace KOS changes and to correctly update the annotations. For instance, we retain the position of an annotation within an instance item ($offset$) since items can cover several concepts. The $offset$ can be useful, e.g. to link concepts from different versions with the same part of an item. We further consider the semantic relationship between a KOS concept and an item or the annotated part of an item. For instance, one item can be annotated as equivalent to a concept, more/less specific, partial match, etc. The semantic type of an annotation is useful to update outdated annotations. For instance, instead of removing an impacted annotation after concept deletion, one could preserve the annotated item by linking it to the superclass of the removed concept and changing its semantic type to “less specific”. As additional provenance information, our annotation model includes an element to indicate which concept attribute (e.g., title, synonym, preferred terms, etc.) has mainly been used to produce an annotation. This can be valuable during the maintenance process, for example, to decide whether a basic attribute change is relevant and might entail an annotation modification.

3 FOUNDATION FOR SEMANTIC ANNOTATION MAINTENANCE

As discussed in Section 1, our long term objective is to design a (semi-)automatic approach for maintaining semantic annotations valid over time if the underlying KOS is evolving without a complete reannotation of the document and by guaranteeing a high quality in the annotation after maintenance. We have analyzed the evolution of several KOS of the medical domain and we identified the behavior of annotations under different scenarios. We rely on these findings to derive different aspects to take into consideration for the maintaining semantic annotations. It can be seen as a multi-level approach that can be split according to inputs, process and outputs. It allows us to optimize the annotation maintenance task by considering at each step more information of different nature to maintain annotation that remain invalid after the previous step.

3.1 Maintenance Process

The different maintenance processes we have identified consist in: i) Automatically detecting inconsistent annotations caused by the evolution of the underlying KOS; ii) Using information gained from the evolution of the KOS only to adapt impacted annotations; iii) Using information of external KOS to maintain annotations that could not be maintained by considering local resource; iv) Using change patterns to finalize

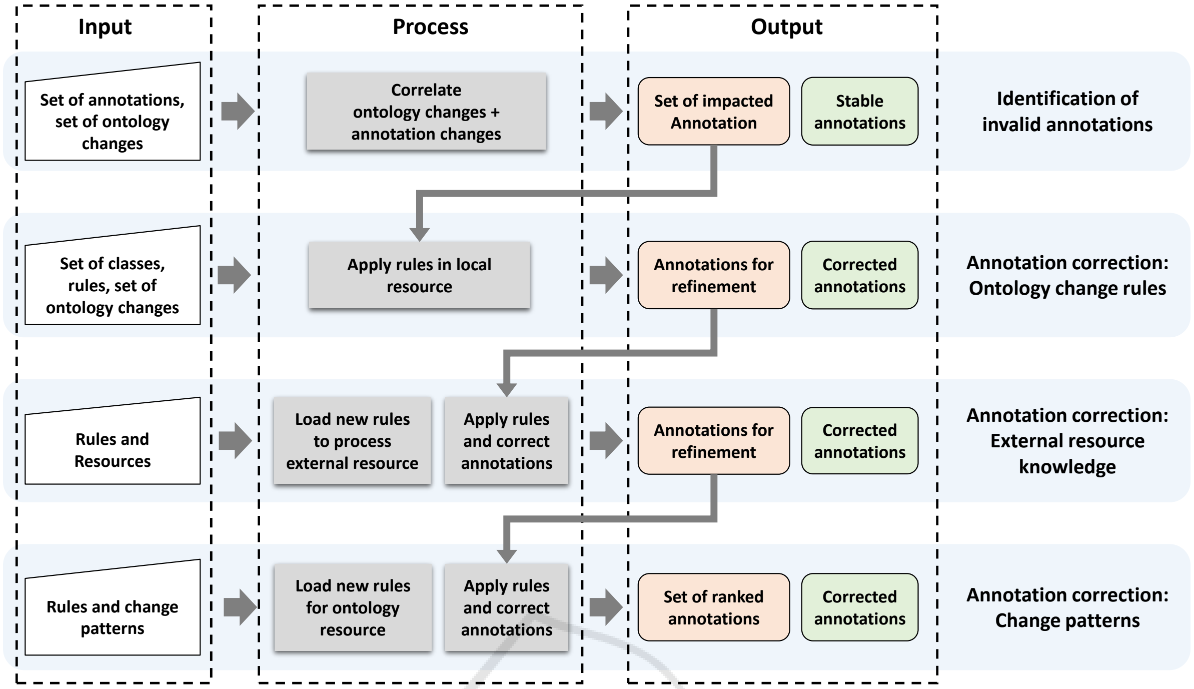
Input
Process
Output
Set of annotations, set of ontology changes
Correlate ontology changes + annotation changes
Set of impacted Annotation
Stable annotations
Identification of invalid annotations
Set of classes, rules, set of ontology changes
Apply rules in local resource
Annotations for refinement
Corrected annotations
Annotation correction: Ontology change rules
Rules and Resources
Load new rules to process external resource
Apply rules and correct annotations
Annotations for refinement
Corrected annotations
Annotation correction: External resource knowledge
Rules and change patterns
Load new rules for ontology resource
Apply rules and correct annotations
Set of ranked annotations
Corrected annotations
Annotation correction: Change patterns

Figure 1: A multi-level framework for supporting annotation maintenance.

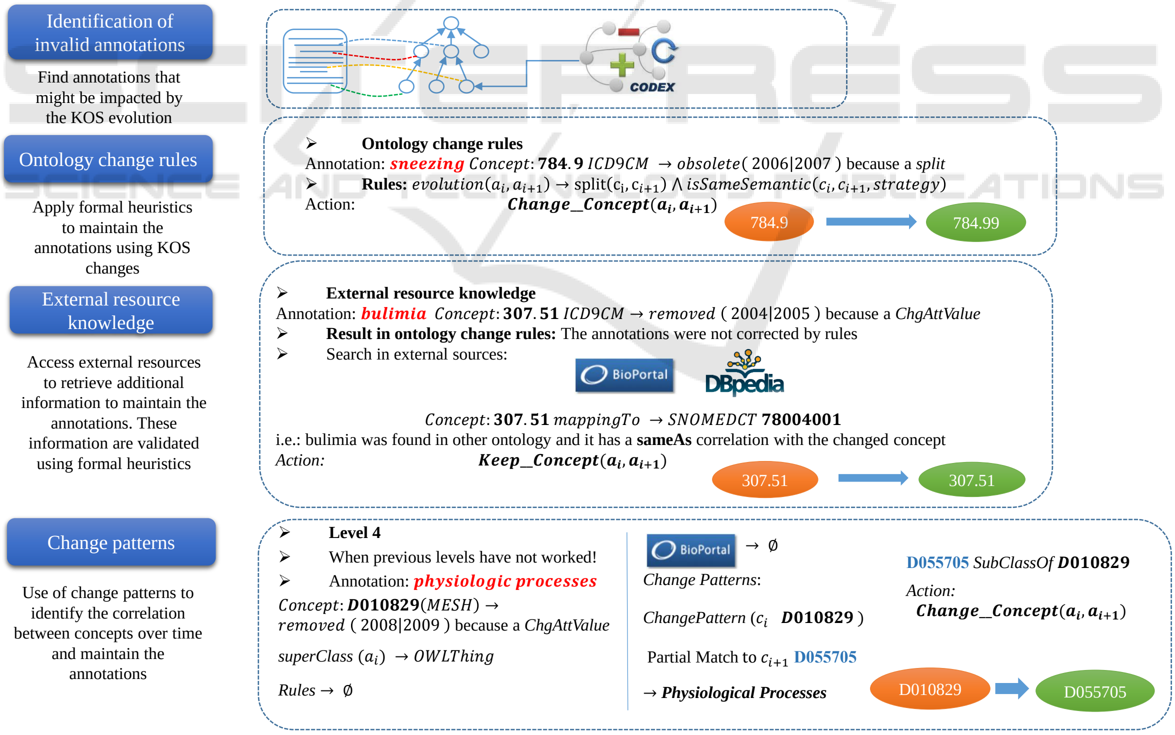
Identification of invalid annotations
Find annotations that might be impacted by the KOS evolution
CODEX
Ontology change rules
Apply formal heuristics to maintain the annotations using KOS changes
Ontology change rules
Annotation: sneezing Concept: 784.9 ICD9CM → obsolete(2006|2007) because a split
Rules: evolution(a_i, a_i+1) → split(c_i, c_i+1) ∧ isSameSemantic(c_i, c_i+1, strategy)
Action: Change_Concept(a_i, a_i+1)
784.9
784.99
External resource knowledge
Access external resources to retrieve additional information to maintain the annotations. These information are validated using formal heuristics
External resource knowledge
Annotation: bulimia Concept: 307.51 ICD9CM → removed (2004|2005) because a ChgAttValue
Result in ontology change rules: The annotations were not corrected by rules
Search in external sources:
BioPortal
DBpedia
Concept: 307.51 mappingTo → SNOMEDCT 78004001
i.e.: bulimia was found in other ontology and it has a sameAs correlation with the changed concept
Action: Keep_Concept(a_i, a_i+1)
307.51
307.51
Change patterns
Use of change patterns to identify the correlation between concepts over time and maintain the annotations
Level 4
When previous levels have not worked!
Annotation: physiologic processes
Concept: D010829(MESH) → removed (2008|2009) because a ChgAttValue
superClass (a_i) → OWLThing
Rules → ∅
BioPortal → ∅
Change Patterns:
ChangePattern (c_i D010829)
Partial Match to c_i+1 D055705
→ Physiological Processes
D055705 SubClassOf D010829
Action:
Change_Concept(a_i, a_i+1)
D010829
D055705

Figure 2: Examples illustrating the behaviour of the framework at each level.

the maintenance and optimize the quality of the set of adapted annotations.

- **Identification of invalid annotations**: It consists in identifying invalid annotations by analyzing the evolution of the associated KOS. To this end, it takes as input a set of annotations and two successive versions of the used KOS namely K_n and K_{n+1}. The identification of concepts that have changed between K_n and K_{n+1} can be obtained using an ontology *Diff* tool (Hartung et al., 2013; Noy et al., 2002) as well as additional information specifying the type of changes that have affected these concepts. As it is the case for ontology mapping adaptation (Groß et al., 2013), such information plays a key role in the maintenance task because it will determine the type of correction to apply to the annotations in the next levels. For instance, the deletion of a concept attribute can lead to the deletion of annotations but the deletion of the same attribute value in the context of a split of concept can lead to the migration of the annotation to the evolved version of the concept (i.e. the result of the split). It is therefore crucial to consider not only basic ontological changes (i.e. addition/deletion of concept) as it is the case in existing approaches for annotation maintenance but complex changes (i.e. split/merge of concepts) to optimize both the maintenance process and the quality of the adapted annotations.

- **Annotation correction using ontology change rules**: It consists in using information derived from the set of annotations itself as well as the data of the *Diff* between the two KOS versions K_n and K_{n+1} coming from the previous level to adapt the identified invalid annotations. At this level, the correction of annotations can be specified in rules that combine the context of evolution of the KOS and the status of the annotations. Under these conditions, the rules must specify the maintenance action to perform. For instance, we observed cases where an annotation was impacted because in its new version the label of concept associated with this particular annotation adopted the plural form. Therefore the corresponding rule can look like:

  “If the label of concept is set to its plural form then do not change the annotation”

  The type of ontological change contained in the *Diff* allows to propose more elaborated correction rules acting directly on the element of the annotation model like the $offset$. For instance, if the attribute used to annotate the text contained in an EHR is modified (e.g., a new word was added at the end of the label), then we check if we can find this modification in the text of the document to annotate (e.g., if the new word is also adjacent to the old text) by checking the information located at the $offset$ position. The corresponding rule is:

  “If the label of concept increase and the data located beside the offset of the annotation is equal to the added word then increase the offset”

  Another example is depicted in Figure 2. The annotation *sneezing* associated with the concept having as code 784.09 in ICD-9-CM version 2006 is no more valid in 2007. It is the direct consequence of the split of concept 784.9 between 2006 and 2007.

  In the example of Figure 2 the depicted rule checks if a concept was split. Basically, it specifies if the concept 784.99 from the new version of ICD-9-CM was engendered by a split of concept and whether it has a label (or an attribute value) which is equal to the same text of the annotation. It also verifies that the new version of the concept 784.9 (if it still exists) has no label of concept that fully match the text of the annotation. As a result, the action to maintain this annotation is to change the source concept of the annotation to 784.99.

- **Annotation correction using external resource knowledge**: It consists in using information inferred from external knowledge sources to maintain the annotations that could not be corrected using local resources of the previous level. Actually, in many cases the drift of ontological concepts can be characterized only by considering the semantic relationships provided by other ontologies (Pruski et al., 2016). Often labels of concept are completely different, from the syntactic point of view, before and after evolution. Therefore, considering local resources only does not allow to characterize their evolution and, in turn, cannot be reused for annotation maintenance purpose. The example depicted in Figure 2 about the evolution of the label of concept 307.51 of ICD-9-CM “Bulimia” in 2004 to “Bulimia nervosa” in 2005 shows another use case that requires external knowledge source. Applying existing approaches on annotation associated with this concept would simply lead to the deletion of the annotations. But the consideration of external resource (here mappings between ICD-9-CM and SNOMED CT provided by Bioportal) tells that these two terms are synonyms therefore the annotation can be kept. Nevertheless, the nature of the external knowledge resources can vary. Whether RDF datasets like BIO2RDF (Belleau et al., 2008) or expressive OWL ontologies contained in Bioportal (Noy et al., 2009) are considered, the inferred information can be of different quality and can affect the quality of the maintenance process.

- **Annotation correction using change patterns**: At this stage, information provided by the *Diff* and the use of external resources are not sufficient to maintain invalid annotations. The analysis of the morphosyntactic form of concept labels can reveal information to take decision about the maintenance of annotation. This technique has already been explored in the context of ontology mapping adaptation (Dos Reis et al., 2015) but remains less relevant in terms of quality in the resulting maintenance decisions. Change Patterns are modifications observed in attribute values of a concept using linguistic-based features to identify the correlation between concepts over time. For instance, a *Partial Copy* between concepts is computed if and only if there exists a partial overlap between words from an attribute present in the KOS version K_n and an attribute in the new KOS version K_{n+1} (i.e., the attribute a_0 becomes a_1).

For instance, the annotation "Physiologic processes", shown at the bottom of Figure 2 produced using MeSH in period 2008/2009 was removed. This is due to a change in the attribute value in the definition of the concept D010829 leading to "Physiological Phenomena". Assuming the following conditions: i) we do not have information inside the ontology to handle with this change, ii) the super class from concept D010829 is *Thing* iii) external resources do not provide the necessary information to make decision, the application of four change patterns (*total copy, total transfer, partial copy, partial transfer*) considering only the attributes in the same sub-ontology e.g., the sub classes from concept D010829 allow to change the concept associated to this annotation from D010829 to D055705.

3.2 Output

Our approach was designed to process the annotations according to different levels of granularity, but the outputs only contain three kinds of data.

The first one refers to the nature of the annotations. It makes the distinction between annotations impacted by the evolution of the underlying KOS and non impacted annotations. We described in details these annotations in (Cardoso et al., 2016).

At the levels dealing with the correction of the annotations, the outputs are: i) the corrected annotations and ii) the set of annotations that need further investigation. Once corrected, the annotations are also enriched with evolution information making future modifications easier and enhancing their quality.

If invalid annotations remain, the definition of another levels exploiting different kind of information for maintenance purpose need to be implemented. The complexity of the evolution affecting KOS, the nature of the annotation, the specificities of the kind of object to annotate need to be taken into account in the definition of the additional levels. The rules that are used at each level also need to be defined by considering the quality of the adapted annotations.

4 CONCLUSION

We have presented a multi-level approach towards the (semi-)automatic maintenance of the annotations turned invalid after the evolution of their associated KOS. Our proposal is based on literature review as well as experimentations and consists in the progressive integration of complex information of different nature and various sources for correcting invalid semantic annotations without re-annotating documents. As future work, we will put the stress in the definition and validation of such a framework. Since our annotation framework will be used to (semi-)automatically correct outdated annotations, the used methods will need careful evaluation according to the quality of the produced results. For future work, we also plan to evaluate the different maintenance approaches using several annotation datasets from the biomedical domain such as annotated CRFs or EHR.

ACKNOWLEDGEMENTS

The National Research Fund (FNR) of Luxembourg and Deutsche Forschungsgemeinschaft (DFG) support this work under the ELISA research project.

REFERENCES

Abgaz, Y. M. (2013). *Change impact analysis for evolving ontology-based content management*. PhD thesis, Dublin City University.

Auer, S. and Herre, H. (2007). *A Versioning and Evolution Framework for RDF Knowledge Bases*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716.

Burger, T., Morozova, O., Zaihrayeu, I., Andrews, P., and Pane, J. (2010). Report on methods and algorithms for linking user-generated semantic annotations to semantic web and supporting their evolution in time.

Cardoso, S. D., Pruski, C., Silveira, M. D., Lin, Y., Groß, A., Rahm, E., and Reynaud-Delaître, C. (2016). Leveraging the impact of ontology evolution on semantic annotations. In *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, pages 68–82.

Da Silveira, M., Dos Reis, J. C., and Pruski, C. (2015). Management of dynamic biomedical terminologies:

Current status and future challenges. *Yearbook of Medical informatics*, 10(1):125–133.

Dos Reis, J. C., Dinh, D., Da Silveira, M., Pruski, C., and Reynaud-Delaître, C. (2015). Recognizing lexical and semantic change patterns in evolving life science ontologies to inform mapping adaptation. *Artificial intelligence in medicine*, 63(3):153–170.

Eilbeck, K., Moore, B., Holt, C., and Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, 10(1):67.

Frost, H. R. and Moore, J. H. (2014). Optimization of gene set annotations via entropy minimization over variable clusters (emvc). *Bioinformatics (Oxford, England)*, 30(12):1698–1706.

Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K. B., Hunter, L. E., and Verspoor, K. (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):1–29.

Gimenez, F., Xu, J., Liu, Y., Liu, T. T., Beaulieu, C. F., Rubin, D. L., and Napel, S. (2012). Automatic annotation of radiological observations in liver CT images. In *AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, Illinois, USA, November 3-7, 2012*.

Groß, A., Dos Reis, J. C., Hartung, M., Pruski, C., and Rahm, E. (2013). Semi-automatic adaptation of mappings between life science ontologies. In *International Conference on Data Integration in the Life Sciences*, pages 90–104. Springer.

Groß, A., Hartung, M., Kirsten, T., and Rahm, E. (2009). Estimating the quality of ontology-based annotations by considering evolutionary changes. In *International Workshop on Data Integration in the Life Sciences*, pages 71–87. Springer.

Groß, A., Hartung, M., Prufer, K., Kelso, J., and Rahm, E. (2012). Impact of ontology evolution on functional analyses. *Bioinformatics*, 28(20):2671–2677.

Hartung, M., Groß, A., and Rahm, E. (2013). Conto–diff: generation of complex evolution mappings for life science ontologies. *Journal of biomedical informatics*, 46(1):15–32.

Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. Reports - Descriptive.

Köpke, J. and Eder, J. (2011). Semantic invalidation of annotations due to ontology evolution. In Meersman, R., Dillon, T., Herrero, P., Kumar, A., Reichert, M., Qing, L., Ooi, B.-C., Damiani, E., Schmidt, D., White, J., Hauswirth, M., Hitzler, P., and Mohania, M., editors, *On the Move to Meaningful Internet Systems: OTM 2011*, volume 7045 of *Lecture Notes in Computer Science*, pages 763–780. Springer Berlin Heidelberg.

Luong, P.-H. and Dieng-Kuntz, R. (2006). A rule-based approach for semantic annotation evolution in the coswem system. In *Canadian Semantic Web*, volume 2 of *Semantic Web and Beyond*, pages 103–120. Springer US.

Maynard, D., Peters, W., and Sabou, M. (2007). Change management for metadata evolution.

Noy, N. F., Musen, M. A., et al. (2002). Promptdiff: A fixed-point algorithm for comparing ontology versions. *AAAI/IAAI*, 2002:744–750.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440.

Park, Y. R., Kim, J., Lee, H. W., Yoon, Y. J., and Kim, J. H. (2011). Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products. *BMC Bioinformatics*, 12(1):1–7.

Pruski, C., Dos Reis, J. C., and Da Silveira, M. (2016). Capturing the relationship between evolving biomedical concepts via background knowledge. In *the 9th Semantic Web Applications and Tools for Life Sciences International Conference*.

Qin, L. and Atluri, V. (2009). Evaluating the validity of data instances against ontology evolution over the semantic web. *Information and Software Technology*, 51(1):83 – 97.

Tissaoui, A., Aussenac-Gilles, N., Hernandez, N., and Laublet, P. (2011). Evonto - joint evolution of ontologies and semantic annotations. In Dietz, J., editor, *International Conference on Knowledge Engineering and Ontology Development (KEOD), Paris, 26/10/2011-29/10/2011*, pages 226–231.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14 – 28.

Yimam, S. M., Biemann, C., Majnaric, L., Šabanović, Š., and Holzinger, A. (2016). An adaptive annotation approach for biomedical entity and relation recognition. *Brain Informatics*, 3(3):157–168.

Zavalina, O. L., Kizhakkethil, P., Alemneh, D. G., Phillips, M. E., and Tarver, H. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information &amp; Knowledge Management*, 14(01):1550005.