International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 3 (2011) pp. 290 -297

A Generic Evaluation Framework for Knowledge-Based Infrastructures: Design and Applications

Juan José Bosch Vicente[1, 2], Fanny Klett (IEEE Fellow)[3]

[1] Fraunhofer Institute of Digital Media Technology, Ilmenau, Germany
[2] Universitat Pompeu Fabra, Barcelona, Spain
juanjo.bosch@gmail.com

[3] German Workforce ADL Partnership Laboratory, Ilmenau, Germany
fanny.klett.de@adlnet.gov

***Abstract*: In this article, the authors introduce a generic framework for the evaluation of knowledge-based infrastructures, and present its applicability based on technologies developed within the Core Technology Cluster (CTC) of one of the most research-intensive programs of the German government toward new technologies for acquisition, processing, presentation and delivery of information on the Internet, the THESEUS program: (http://theseus-programm.de). The main components of the framework are presented, and the evaluation process is described, including also the identified steps that have to be completed in advance, such as the selection of appropriate evaluation criteria, metrics, and references. The flexibility of the presented framework enables its application for a wide range of technologies. This article reflects the application of the proposed framework for evaluation of software components that deal with key aspects toward the realization of the semantic web, namely: ontology reasoning, ontology mapping, machine learning algorithms used for learning semantic annotation from unstructured data, and document structure recognition.**

***Keywords*:** Knowledge-Based infrastructures, Semantic Web, Evaluation framework, Statistical machine learning evaluation, Ontology matching evaluation, Ontology reasoning evaluation

I. Introduction

The main motivation for the design of the evaluation framework proposed in this article is the increasing need for a generic methodology applicable to the evaluation of various technologies, particularly the knowledge-based technologies developed within the CTC of the THESEUS program. This program represents an important contribution to the creation of a new internet-based knowledge infrastructure that allows for a fast and effective knowledge processing [1].

Two broad objectives of evaluation are identified in ISO 14598 and considered in the proposed framework: (1) to identify problems so that they can be rectified, and (2) to compare the quality of a product with alternative products or against requirements [2]. A more refined basis for the evaluation of quality is presented in the quality model introduced in ISO 9126-1 [3]. This model distinguishes between three approaches to quality: internal quality, external quality, and quality in use. The internal and external quality approaches refer to software products, while the quality in use approach refers to the effect of the product that is being used. Furthermore, ISO 9126 introduces six characteristics of the software quality attributes: functionality, reliability, usability, efficiency, maintainability, and efficiency [3]. Each of these characteristics is subdivided into sub-characteristics, which can be measured by internal or external metrics. Fig. 1 illustrates the relationship between quality attributes and their measures. The main focus of the proposed evaluation framework is not on the technology (software product in terms of ISO 9126-1) itself, but on the effects of its use in determined contexts. Thus, this article concentrates on measuring the external quality, and the quality in use.

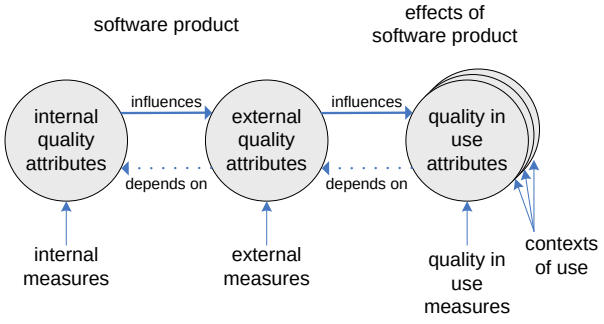

Figure 1. Quality in the lifecycle (adapted according to ISO 9126-1 [3])

In the following sections, the evaluation framework and its elements are presented initially in a generic way, followed by the implementation of this framework in particular application scenarios, which involve the design of a specific testing methodology: (1) The scenario presented in Section 3 *Application of the generic framework to ontology reasoning evaluation* focuses on ontology reasoning technologies. (2) The scenario addressed in Section 4 *Application of the generic*

*framework to ontology matching evaluatio*n deals with ontology mapping. (3) The scenario reflected in Section 5 *Application of the generic framework to statistical machine learning evaluation* relates to the use of statistical machine learning methods for the extraction of semantic relations, the recognition of named entities, and additionally, the recognition of the structure of text documents.

II. Evaluation Framework

As explained in the previous section, in order to perform the evaluation, it is necessary to measure the appropriate quality attributes of the developed technologies in accordance with their context of use. The proposed evaluation framework, firstly introduced in [4] and subsequently detailed during further research work, significantly extends the framework proposed in ISO 9241-11 [5], by considering both, the context of use of the software, and simultaneously a set of references required for the evaluation process. Fig. 2 provides an overview of the relationship between the main elements of the framework: (1) a set of references, and (2) the evaluation process itself.

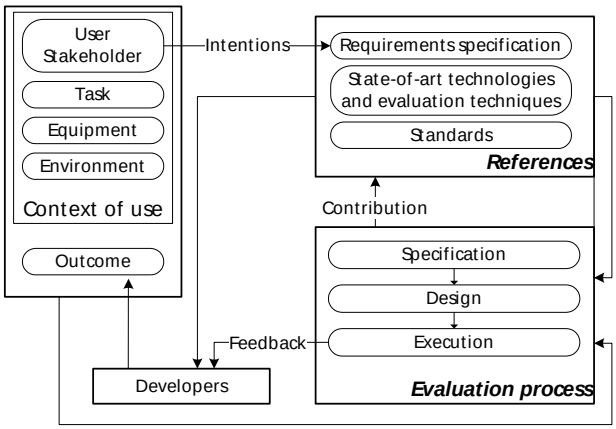

Figure 2. Graphical representation of the evaluation framework

The evaluation process lays the following consecutive steps out: (1) specification, (2) design, and (3) execution. The evaluation process requires as inputs both, the set of references, and the product, e.g. the software components, in their intended context of use. Against this background, the output of the evaluation process represents the feedback to the developers to be considered in the next iteration of the development cycle, which corresponds to a formative evaluation.

The proposed framework can furthermore serve a summative evaluation that is commonly completed at the end of a development process. It also allows the developers to compare (or benchmark) the quality of the developments against alternative state-of-the-art technologies during the development cycle. These benchmarks, which are typically being performed with finished products in summative evaluations processes, provide an early and useful feedback to the developers on the strengths and weaknesses of their algorithms in relation to other approaches.

The following subsections put each of the elements of the proposed framework and their structure in the centre of consideration.

A. References

This subsection emphasizes the importance of the set of references that are identified to serve as an input for the evaluation process. This set concerns the requirements specified by the users and/or stakeholders, the established standards, such as ISO standards for Quality, User-Centered Design (ISO 9126 [3], ISO 13407 [6], etc.), and the state-of-the-art technologies and evaluation techniques as follows:

(1) The specification of the needs of the users and/or stakeholders is a crucial task in the development process. Its output addresses the requirements specification, which represents an essential reference for both, developers and evaluators.

(2) Standards are useful for the evaluation process, since they provide accredited guidelines and procedures, which can support the process of designing the evaluation methodology toward various areas of technology.

(3) State-of-the-art technologies should be considered, in order to allow for the qualitative comparison between the developed components and their alternatives. Moreover, state-of-the-art evaluation techniques offer a basic reference for the selection of both, appropriate metrics and evaluation methodology.

The basic set of references this subsection focuses on, need to be specified for each of the intended application scenarios, and extended or adapted where appropriate. The following subsection provides the detailed description of the evaluation process itself.

B. Evaluation process

This subsection deals with the consecutive steps of the evaluation process as noted in the beginning of this Section. It also addresses the relationship between these three steps, and the required references.

(1) The “Specification” step refers to the definition of the purpose of the evaluation. Against this background, the requirements specification represents a basic reference, since the main intentions and objectives of the software tests derive from these documents. Once the purpose of the tests is defined, the “Specification” step approaches the selection of appropriate metrics, rating levels and criteria to be used during the evaluation process. The requirements specification should provide quantitative rating levels for the selected quality attributes. Additionally, it is recommended to consider the state-of-the-art technologies in the corresponding application scenario, especially when no particular rating levels have been specified by the users and/or stakeholders. Appropriate rating levels should be set to benchmark the quality of the developments with alternative technologies. Also related standards can support the selection of appropriate metrics and criteria, which need to be selected for the evaluation of each of the components. King [7] presents some practical criteria for the selection and definition of metrics, such as reach the highest value for perfect quality; reach the lowest value for worst possible quality; be monotonic; be clear and intuitive; correlate well with human judgment; be reliable and exhibit as little variance as possible; be cheap to set up and apply; and finally, be automatic.

(2) The “Design” step of the evaluation process concerns the evaluation methodology and deals, therefore, with the

set-up of an evaluation plan according to the above-mentioned "Specification" step. Along with the selection or creation of the appropriate tools, which typically allow performing the evaluation in an automatic, objective and repeatable way, also databases or corpora selection are in the focus of this step.

(3) The "Execution" step involves the measurement of the previously selected characteristics, the comparison by use of the selected criteria, and finally, the assessment of the results. This step provides feedback to the developers to serve the subsequent iteration of the design process.

The next sections address the implementation of the presented generic framework in selected specific application scenarios.

III. Application of the generic framework to an ontology reasoning evaluation

This section demonstrates the application of the generic evaluation framework toward a specific testing methodology for an ontology reasoning application scenario that results from the THESEUS program [1]. The following subsections address the scenario and the testing methodology itself, and provide details about the developed benchmarking tool.

A. Ontology reasoning scenario

Ontology reasoning is a fundamental part of the semantic technologies. The main challenge is to derive implicit information from explicit information in terms of logical consequence. Ontology reasoning can also be used for validation and deduction purposes. The adequate selection of a reasoner depends on the characteristics of the reasoning task to be performed. There are applications, for example in the medical area, where highly accurate results are important as the answers represent possible diseases the patient may suffer from. In applications where the time performance is more critical than the correctness of the reasoning results, approximate reasoning becomes a key requirement.

The performance of ontology reasoning systems is largely dependent on the size of the ontologies, the queries, and the expressivity of the language. The most commonly used evaluation metrics are: load time, query response time, and the soundness and completeness in terms of precision and recall [8], calculated by comparing the answers of the tested reasoner against the results provided by a reference reasoner. The quantitative evaluation of the reasoners dealing with approximate reasoning benefits from the analysis of the variation of the completeness and soundness of the results with respect to the gain in speed. Guo et. al. [11] present a supportive list of recommendations and requirements for the creation of knowledge based systems benchmarks.

Recent benchmarks performed in the field have shown that most reasoners are efficient in only some of the above-mentioned aspects. It is thus necessary to perform an evaluation, which investigates the behavior of the reasoning components by involving small as well as large ontologies in terms of the Resource Description Framework (RDF) triples [9], [10].

B. Specific testing methodology

In order to properly apply the generic framework for the evaluation of ontology reasoning software components, it is necessary to analyze their characteristics [11]. According to the use case requirements [1], the most important characteristic is the query response time, which largely depends on the size of the ontologies, and the queries used for reasoning. Also the completeness and soundness of the results are important in the evaluation of approximate reasoners.

By referring to the general framework explained in the previous section, the evaluation metrics selected in the "Specification" step involve load time, query response time, and precision and recall. The "Design" step that addresses the creation of the evaluation methodology, refers now in the particular application to the selection of the adequate ontologies, queries, and the development of a reasoners benchmarking tool.

The general "Execution" step concerns the measurement of the selected metrics, and the assessment of the results, which depends on the intended use of the technologies. The concrete evaluation methodology created in the "Execution" step is illustrated in Fig. 3. A set of ontologies and a set of queries are applied to both, the tested reasoner and the reference reasoner. Then the query response times of the tested and reference reasoners are measured and compared. The soundness and completeness of the tested reasoner is investigated by comparing the answers against the answers from the reference reasoners, which are sound and complete. The precision (P) and recall (R) are calculated as follows: P=correct/(correct+incorrect), R=correct/(correct+miss), where "miss" corresponds to the results not found by the tested reasoner, "correct" corresponds to the results, which were correctly found, and "incorrect" corresponds to the results, which were incorrectly computed as part of the answers.

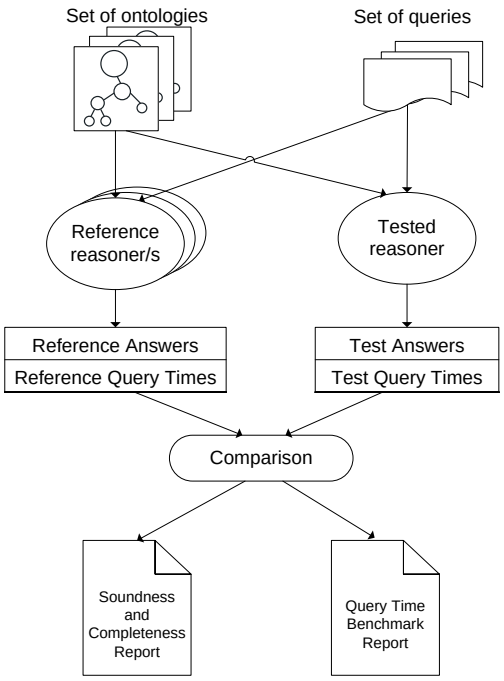

Figure 3. Evaluation methodology for the reasoning components

The described methodology forms the basis for the creation of the reasoner benchmarking tool that automatically performs the measures, and compares them with the references.

C. *Reasoner benchmarking tool*

The developed benchmarking tool targets at serving the "Execution" step of the evaluation process. The benchmarking tool measures the query response time of the system under test and compares it with the reference reasoners, such as Pellet [12] and KAON2 [13]. The answers of the queries provided by the tested reasoner are also compared against the answers provided by the reference reasoners according to the evaluation methodology introduced in the previous subsection.

Fig. 4 points toward the features that the benchmarking tool offers for the selection of the appropriate ontologies and query files, as well as the number of consecutive queries to be executed, in order to minimize the effect of caching issues [11].

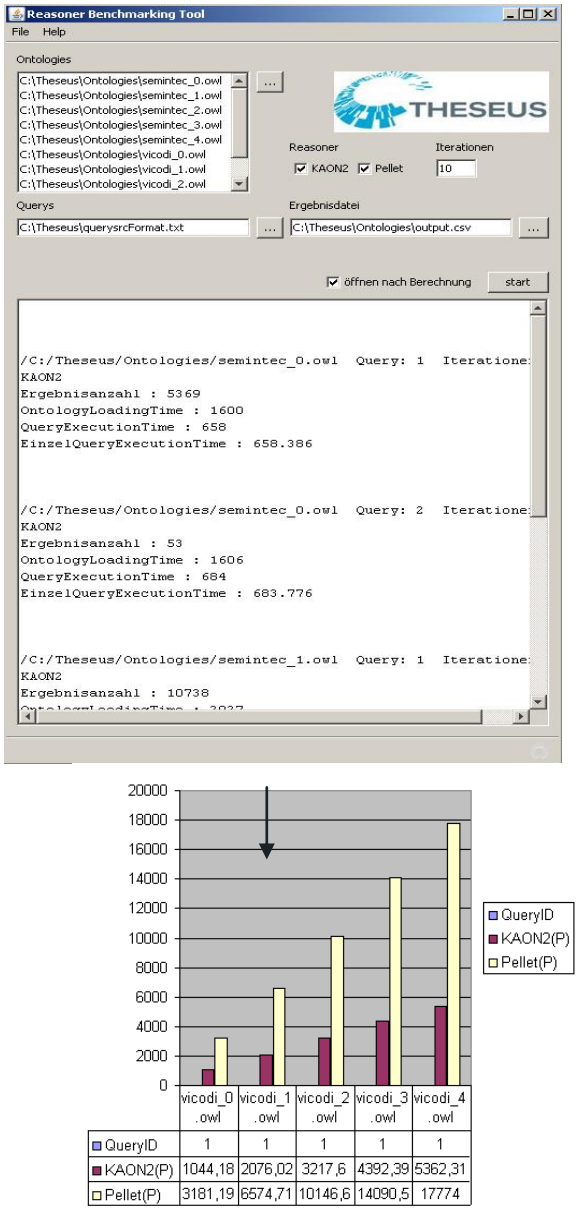

Figure 4. Screenshot of the developed reasoning benchmarking tool and the integrated visualization of the results

The benchmarking tool provides the opportunity to select the reasoners for the benchmarking process, as well as the output file, where the measures of the query response times will be written. An integrated visualization option allows for the display of the results to facilitate the comparison of the query response time performance of the system under test with the alternative approaches.

The visualized results shown in Fig. 4 illustrate the fact that the benchmarking tool can additionally be used to benchmark the query response time of a group of reasoners without the specification of a system under test.

IV. Application of the generic framework to an ontology matching evaluation

This section demonstrates the application of the generic evaluation framework toward a specific testing methodology for an ontology matching application scenario that results from the THESEUS program [1]. The following subsections address the scenario and the testing methodology itself, and provide details about the resources used in the evaluation process.

A. *Ontology matching scenario*

The concept of matching ontologies means that for each concept in terms of relation, and property in a given ontology, a corresponding concept has to be located in the second ontology, with the same or closest meaning. This is a very important matter toward the realization of the semantic web, since it deals with homogenization and interoperability aspects. Thus, ontology matching can benefit many tasks, such as query answering, data translation, navigation, and also tasks that use knowledge and data that are distributed in several ontologies.

The performance of the matching components depends on the size of the ontologies. Hence, an important requirement for the evaluation is the inclusion of series of ontologies with increasing size, in order to be able to measure the ability of the matching algorithms to cope with them.

The comparison of the performance of different matching tools can be difficult since human experts did not agree on how to merge ontologies by now, and we do not yet have a good enough metric for comparing ontologies [14]. Further, one of the difficulties in this area of interest is that there is no consensus on the merits of the various approaches, or on their classification. In fact, many overviews of ontology mapping vary in their approaches and are incompatible [15]. Qualitative evaluation may thus be subjective and produce different results depending on the evaluator.

In order to overcome these difficulties, several initiatives have been undertaken, which deal with the evaluation of ontology matching components. In line with this development, the next subsection addresses a feasible specific testing methodology for ontology mapping.

B. *Specific testing methodology*

The quantitative evaluation of the mapping software components is mostly performed in terms of precision, recall and composite measures such as the F-measure (harmonic mean of precision and recall), while the time response of the

algorithms is not the most relevant aspect according to the requirements specification.

Therefore, the most important metrics to be applied in the "Specification" step for ontology mapping evaluation are precision and recall. In this context, precision can be defined as the share of real correspondences between the two ontologies among all found, while recall can be defined as the share of real correspondences between the two ontologies that are found. The proposed metrics are calculated by the comparison of the output of the mapping algorithms with a manually created reference mapping. The precision (P) and recall (R) are calculated in this context as follows: P= tp/(tp+fp), R=tp/(tp+fn), where "fn" corresponds to the matches between the ontologies not found by the algorithm, "tp" corresponds to the matches that were correctly found, and "fp" corresponds to the matches found by the algorithm but not present in the reference set of mappings. The F1-measure is calculated as follows: F1= 2·P·R/(P+R)

Another measure introduced in [16] is the "Overall" measure, where P is the precision, and R is the recall:

$$Overall = R \cdot \left(2 - \frac{1}{P} \right) \quad (1)$$

This "Overall" measure estimates the effort that it would cost a user to modify the proposed match result to the intended result. It is only valid for values of precision above 0.5 (meaning that at least half of the proposals are correct). Otherwise, the overall accuracy is negative, meaning that it would take the user more effort to remove the false positives and add the missing matches than to complete the matching manually. Obviously, the best result is obtained when both, precision and recall are equal to 1.0.

A measure of the "distance" between ontologies can also be supportive [17]. Generalized precision and recall are useful to overcome some limitations of the classical precision and recall. With the classical metrics, "an alignment may be very close to the expected result and another quite remote from it and both return the same precision and recall", as detailed in [18]. Thus, the generalized metrics measure the distance between the alignments, instead of the strict equality.

The particular evaluation methodology for the ontology mapping component created to fit the general "Design" step explained in Section 2 deals with the comparison of the produced alignments with a reference alignment. In this step, the systematic benchmark series should be designed to identify the weaknesses and strengths of the matching algorithms. For instance, the tests progressively discard information, in order to evaluate how the matching algorithms treat situations when information is lacking.

The following subsection addresses the selection of the tools, and resources needed for the evaluation, such as the ontologies to be matched, and the reference mappings. Additionally, the comparison with alternative approaches performed in the "Execution" step is also described.

C. Tools and resources for an ontology mapping evaluation

Some initiatives arose to handle ontology alignment evaluation, and tried to establish a consensus for the evaluation of available methods for schema matching and ontology integration. One example refers to the Ontology Alignment Evaluation Initiative Campaign (OAEI). The main objective of the OAEI is to "improve the quality of ontology matching algorithms by continuous comparison with other methods" [19]. It also aims at providing high quality benchmarks that can be used for comparing systems. Another main goal of the initiative is to enhance the evaluation methods and metrics used.

The OAEI covers several tracks, which focus on different domains, and especially on different characteristics of the ontology matching task, such as scalability, ability to deal with decreasing information in the ontologies, etc. The OAEI also provides useful resources for the evaluation of ontology matching technologies, which allow for comparing the results obtained with the tested algorithms against the results obtained by the participants in the contest. The resources available include the ontologies to be matched, and the reference mapping used to compute the previously introduced metrics. Additionally, there are tools, which can be used to perform the evaluation in the "Execution" step. Since the OAEI 2010 edition it is also possible to perform the evaluation of the algorithms as web services, in collaboration with the SEALS project [20], which is developing a reference infrastructure for the evaluation of semantic technologies.

V. Application of the generic framework to a statistical machine learning evaluation

This Section refers to the application of the generic framework to a scenario covering the usage of statistical machine learning algorithms in two main settings, as presented in the following subsections.

A. Statistical machine learning scenario

Statistical machine learning techniques have been used in a large variety of applications from failure prediction in automotive assembly plants [21] to the semantic web. The following explanations demonstrate the applicability of the generic framework for two settings with a broad coverage of applications in the semantic web, namely (1) the semantic annotation of unstructured data, which deals with a main bottleneck, the manual annotation; and (2) the recognition of the structure of textual documents, which is a helpful step prior to the textual analysis.

(1) Semantic annotation in the context of the evaluation deals with the generation of RDF triples from textual data. Two main subtasks are: Named Entity Recognition (NER), and Semantic Relation Extraction (SRE). NRE involves the identification of pieces of text with entities, such as person, location, dates, etc., while SRE addresses the identification of predefined relations between pairs of entities in a text.

State-of-the-art algorithms are able to achieve near-human performance for named-entity recognition, scoring F-measures around 95% as stated in [22] in the Message Understanding Conference (MUC-6), where the focus was set on NER for persons, locations, and organizations [23]. However, the achieved performance in both, NRE and SRE, depends on the language, the domain, and the entities or relations of interest. In other evaluation forums, such as the Conference on Computational Natural Language Learning (CoNLL-2003) [24], where language-independent NER was

in the centre of consideration, or BioCreAtIvE II targeted at the biomedical domain, the best F-measure achieved by the participants was below 90%.

(2) The structure recognition of textual documents, such as scientific papers, file cards, letters, etc., refers to a textual document as an input of the software component to be evaluated, while the output represents the extracted layout structure, such as headings, footnotes, title, abstract, etc. Moreover, the evaluated structure recognition component deals also with the visualization and refinement option for the structure recognition results.

B. Specific testing methodology

The requirements specification serves as an important reference to bring into the focus of evaluation the appropriate entities, relations, domain and language, on the one hand, and to facilitate the selection of the appropriate corpora, tools, and rating levels, on the other hand.

The particular application of the generic "Specification" step aims at the selection of the traditional metrics: precision, recall and F-measure for the NER and SRE evaluation. According to [25], [26], the Receiver Operating Characteristics (ROC) curve, and more specifically, the area under this curve may be used as a single-number measure. Also the specific AUC score has been theoretically and empirically proved to provide better results than accuracy in the evaluation of learning algorithms [25], [26]. Consequently, the authors recommend its application it in the evaluation process.

The evaluation methodology created in the "Design" step also involves the comparison of the automatically computed results against a golden standard. The corpora used in the evaluation process were selected from initiatives which deal with the quantitative evaluation of SRE and NER, such as the Automatic Content Extraction program (ACE) or the Message Understanding Conference (MUC-6). In the biomedical domain, the GeneRIF (Gene Reference Into Function) dataset was used. It consists of 5720 GeneRIF sentences retrieved from 453 randomly selected Entrez Gene database entries [27]. The task targeted at the identification of relations between diseases and genes from a set of concise phrases.

In the case of the structure recognition, a corpus of scientific papers was created, and manually annotated, with the structure of each of the parts. The considered labels applied were: "Abstract", "Authors", "FigureCaption", "Headline", "References", "Text", "Title" and at last, "Ignore" in case none of the previous label was applicable. An evaluation tool was developed to perform the "Execution" step, by means of a ten-fold-cross validation. This tool computes precision, recall and F-measure for each field of interest, since the results typically depend on their similarity with other fields. For instance, it is easier to recognize the field 'page-number' than fields like 'abstract' or 'biography' [28].

In order to evaluate the capability of the visualization and refinement option for the structure recognition results, the authors propose a quantitative evaluation of the visual interface, based on the measurement of the improvement rate of the results achieved when users are involved, compared to the fully automatic classification results. The time required by the interactive process can be computed, and effectiveness and efficiency can be then measured. Following the definitions from [5], effectiveness is the accuracy and completeness with which users achieve specified goals, and efficiency concerns resources spent in relation to the accuracy and completeness with which users achieve goals. Consequently, the F-measure can be seen as the effectiveness of the system, and the efficiency can be calculated by dividing the effectiveness by the time needed to process the document and get the results.

A measure of the improvement of the effectiveness of the structure recognition system when the visualization component is being used for the correction of the results, in relation to the additional time required, is computed by calculating the slope of the curve in Fig. 5.

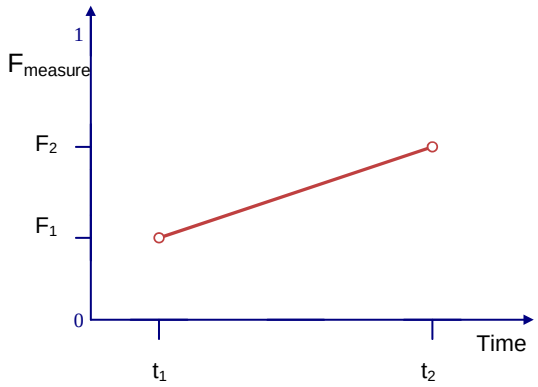

Figure 5. Improvement of the F-measure by using the visualization and refinement option

In Fig. 5, t1 represents the time needed to process the document without a user involvement, t2 represents the time needed to process the document with a user involvement, F1 represents the F-measure achieved without a user involvement, and F2 represents the F-measure achieved with a user involvement. The slope is calculated as follows: $\Delta F/\Delta t = (F2 - F1)/(t2 - t1)$

To assess the results, the authors recommend considering the requirements specification, since it may not be possible every time to involve users in the structure recognition process, or only for a limited amount of time.

VI. Conclusions and future work

This article introduced a generic framework for the evaluation of knowledge-based technologies. The authors demonstrate its applicability by implementing a specific testing methodology, and executing respective evaluation of software components dealing with ontology reasoning, ontology matching, NER, SRE, and document structure recognition.

This framework covers the principles of the iterative system design process, which offer benefits for both, the evaluators and the developers. The first iteration of the evaluation within [1] has been completed, and feedback from the evaluation process has been provided to the developers to facilitate the improvement of their components in the following iteration of the development cycle. The same procedure applies for the evaluation framework itself. The developers' feedback is iteratively applied to enhance the evaluation methodology.

Generally, the objective of this article and the underlying work, is to improve - based on an advanced research and development in the area of knowledge engineering, the state-of-the-art in quality assurance of software components, which target at the acquisition, processing and better use of

knowledge resulting from multiple sources. The appropriate evaluation of the technologies involved is a crucial step, in order to achieve this goal. Therefore, our further work will focus on evolving and applying the generic framework in challenging application scenarios targeting at ontology learning from textual data, situation-sensitive dialogue processing components, semantic information visualization techniques, and further arising technologies. This is expected to open up new opportunities for the effective development of knowledge-based infrastructures and their successful implementation in pioneering applications.

Acknowledgment

This research has been partially supported by the THESEUS Programme funded by the German Federal Ministry of Economics and Technology (BMWi). The first author would also like to thank La Caixa Fellowship program.

References

[1] URL: http://theseus-programm.de/home/

[2] Information technology -- Software product evaluation -- Part 1: General overview, ISO/IEC 14598-1, 1999.

[3] Software engineering — Product quality — Part 1: Quality model, ISO/IEC 9126-1, 2001.

[4] P. Dunker, J.J. Bosch, J. Liebetrau, "Towards evaluating the core technology cluster of the german research project THESEUS", ImageCLEF Workshop on Multimedia Information Retrieval Evaluation, Aarhus, Denmark, 2008.

[5] Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability, ISO 9241-11, 1998.

[6] Human-centered design processes for interactive systems, ISO 13407, 1999.

[7] M. King, "Living up to standards". *In Proceedings of the EACL Workshop on Evaluation Initiatives in Natural Language Processing*, pp. 65-72, Budapest, Hungary, 2003.

[8] C.J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworth & Co. (Publishers) Ltd., 1979.

[9] A. Agostini, C. Bettini, D. Riboni, "A performance evaluation of ontology-based context reasoning," in *Proceedings of the 5th Annual IEEE Int. Conf. Pervasive Computing and Communications, 4th International Workshop on Context Modeling and Reasoning*, pp. 3-8, 2007.

[10] X.H. Wang, D.Q. Zhang, T. Gu,.; H.K. Pung; , "Ontology based context modeling and reasoning using OWL" *In Proceedings of the 2nd IEEE Annual Conf. Pervasive Computing and Communications – Workshops* , White Plains, NY, USA, pp. 18- 22, 2004.

[11] Y. Guo, Z. Pan and J. Heflin, "A requirements driven framework for benchmarking semantic web knowledge base systems". *In IEEE Trans. Knowl. Data Eng.: Special Issue: Knowledge and Data Engineering in the Semantic Web Era*, Vol. 19, No. 2, pp. 297-309, 2007.

[12] E. Sirin, B. Parsia, B. Grau, A. Kalyanpur, Y. Katz: "Pellet: a practical OWL DL reasoner", *Journal of Web Semantics*, Vol 5, No. 2, pp 51-53, 200.

[13] B. Motik, R Studer, "KAON2 – a scalable reasoning tool for the semantic web". In *Proceedings of the 2nd European Semantic Web Conference (ESWC'05)*, Heraklion, Greece, 2005.

[14] J. Sampson, "Measuring the Quality of Ontology Mappings: A multifaceted approach." *Doctoral Symposium at INTEROP-ESA 2005, First International Conference on Interoperability of Information Systems*, Geneva, Switzerland, 2005.

[15] F., van Harmelen, "Ontology mapping: a way out of the medical tower of Babel?" *10th Conference on Artificial Intelligence in Medicine (AIME 05)*, Aberdeen, Scotland: IOS Press, 2005.

[16] S. Melnik, H. Garcia-Molina, E. Rahm: "Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching". In *Proceedings of the 18th International Conference on Data Engineering (ICDE 2002)*, San Jose, CA, 2002.

[17] N. Noy, M. Musen, "Evaluating Ontology-Mapping Tools: Requirements and Experience". *In Proceedings of OntoWeb-SIG3 Workshop at the 13th International Conference on Knowledge Engineering and Knowledge Management*, Siguenza, Spain, pp. 1-14, 2002.

[18] M. Ehrig, J. Euzenat, "Relaxed Precision and Recall for Ontology Matching". In Ben Ashpole, Jérôme Euzenat, Marc Ehrig, and Heiner Stuckenschmidt, editors, *In Proceedings of the K-Cap 2005 workshop on Integrating ontologies* , pages 25–32, 2005

[19] URL: http://oaei.ontologymatching.org/

[20] URL: http://www.seals-project.eu/

[21] A. Pandian, A.. Ali, "A Review of Recent Trends in Machine Diagnosis and Prognosis Algorithms". *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*. Vol.2, 2010 , pp.320-328

[22] G. D, Zhou, and J. Su, "Named entity recognition using a hmm-based chunk tagger". In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL '02)*, 2002, pp. 473–480.

[23] R. Grishman, B. Sundheim, "Design of the MUC-6 evaluation. In: *MUC6 '95: Proceedings of the 6th conference on Message understanding*, 1995.

[24] F. Erik, T. Sang, De F. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". *In Daelemans, W., Osborne, M., eds.: Proceedings of CoNLL*, 2003.

[25] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition* Vol. 30, No. 7, 1997, pp. 1145-1159.

[26] J. Huang, C. Ling, "Using AUC and Accuracy in Evaluating Learning Algorithms", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 3, March, 2005, pp. 299-310.

[27] D. Maglott, J. Ostell, K. Pruitt, T. Tatusova, "Entrez Gene: gene centered information at NCBI". Nucleic Acids Res 2005.

[28] S. Klink, T. Kieninger, "Rule-based Document Structure Understanding with a Fuzzy Combination of Layout and Textual Features". *IJDAR - International Journal on Document Analysis and Recognition*, Vol. 4, No. 1, Springer, 2001, pp. 18-26.

Author Biographies

Juan José Bosch Vicente earned his Telecommunications Engineering degree in the Universitat Politècnica de Valencia, Spain in 2004. During his university studies he also had yearly stays at the National Institute of Applied Sciences Lyon, France and the University of Sheffield, England. He has experience in both industry and research environments.

Mr. Bosch worked firstly for Hewlett Packard, and then Electronic Data Systems, where he was involved in software development projects. In 2007 he joined the Fraunhofer Institute for Digital Media Technology, in Ilmenau, Germany where he firstly worked in the evaluation of knowledge based technologies, and machine learning techniques in order to advance the semantic web. The research presented in this paper reflects the work carried out while his stay in this institution. Furthermore he was involved in the creation of a real time 3D monitoring system, dealing with the control of safety risks in human-robot interaction.

Mr. Bosch is currently at the Universitat Pompeu Fabra, Barcelona, Spain, interested in the application of machine learning algorithms to the analysis and description of audio and music. He has published papers at several international conferences and workshops.

Fanny Klett holds a Ph.D. in Electronic Media Technology from Ilmenau University of Technology. She established the Business Area Data Representation and Interfaces at the Fraunhofer Institute Digital Media Technology. In 2009, she assumed the Directorship of the German Workforce Advanced Distributed Learning Partnership Laboratory, which is run in cooperation with the US Government.

Dr. Klett's research and development interests are directed to new technologies in the area of information and data management and their evaluation, competency and job performance management, collaborative systems as well as sensor and data fusion for monitoring man-machine cooperation.

Dr. Klett was a visiting scientist at the Institute for Computer-Supported New Media, Graz University of Technology, and is being invited as visiting lecturer at many European universities. She actively works in standardization bodies such as the IEEE Learning Technology Standards Committee, the IEEE Standards Association, the US Advanced Distributed Learning Initiative, the Learning Education and Training Systems Interoperability Association, and the ISO SC36 for learning, education and training.

Dr. Klett chaired and served on more than 20 conference planning and program committees (UNESCO, IEEE, APSCE, etc.) She is associated editor of the IEEE Education Society and ASEE Electrical and Computer Engineering Division joint publication "The Interface" and serves on the reviewer board of the IEEE Transactions on Education, and the IEEE Educational Technology and Society Journal.

Dr. Klett is IEEE Fellow and serves the IEEE Educational Activities Board and the IEEE Computer Society Member and Geographic Activities Board by representing Region 8 (Europe, Middle East, and Africa). She is also Member of the Sponsor Executive Committee and Secretary of the IEEE Learning Technology Standards Committee chartered by the IEEE Computer Society Standards Activity Board, and Member of the Council and the Academic Board of the European Association for Education in Electrical and Information Engineering.

Dr. Klett has published more than 80 technical and invited papers and book chapters and organized numerous Special Sessions and Workshops on various digital information research topics at leading international conferences.