# Interactive Data Integration with MappingAssistant

Jan Noessner[1], Faraz Fallahi[2]
Eva Maria Kiss[2], and Heiner Stuckenschmidt[1]

[1] KR & KM Research Group
University of Mannheim, B6 26, 68159 Mannheim, Germany
{jan,heiner}@informatik.uni-mannheim.de
[2] ontoprise GmbH,
An der Raumfabrik 29, 76227 Karlsruhe, Germany
{fallahi,kiss}@ontoprise.de

**Abstract.** Due to the heterogeneity of distributed systems data integration is a main success factor in real-life business. Applying semantic web technologies for matching data is one successful approach for data integration. Ontoprise uses ontologies as target schema for integrating different sources like databases, text-files and ontologies. However, the created target ontology and the corresponding mapping-rules might be error-prone. Hence, we developed the conflict resolution framework MappingAssistant which detects wrong rules or facts on the instance level in an interactive way. In this demo we present the MappingAssistant framework and an evaluation which emphasizes that users are used to investigate data on the instance level.

## 1 Introduction and Process Description

In data integration much work has been invested in producing automated alignments with ontology matching systems [2]. However, alignments produced by automated ontology matching algorithms are still error-prone and, therefore, need to be supervised by a human domain expert. In real-world scenarios users are usually confronted with ill-labeled concepts. Hence, the domain expert is used to check the data on the instance level. However, existing applications like AgreementMaker [1] present alignments mostly on the schema level.

The MappingAssistant simplifies the alignment evaluation process by investigating data on the instance level. In the example shown in Figure 1 the user selected *FamilyCar* in the target schema. In the next step, the user identifies those instances which have been classified incorrectly. Due to the amount of instances a user can be faced to diagnose we utilize different clustering techniques in order to reach data simplification. Attribute-driven combinations of weighted hierarchical and partial clustering algorithms, as mostly described in [3], are therefore utilized. In our example the *MX5_Mieta* is a two seated car and, thus, not a *FamilyCar*. Based on this information, a diagnostic algorithm asking

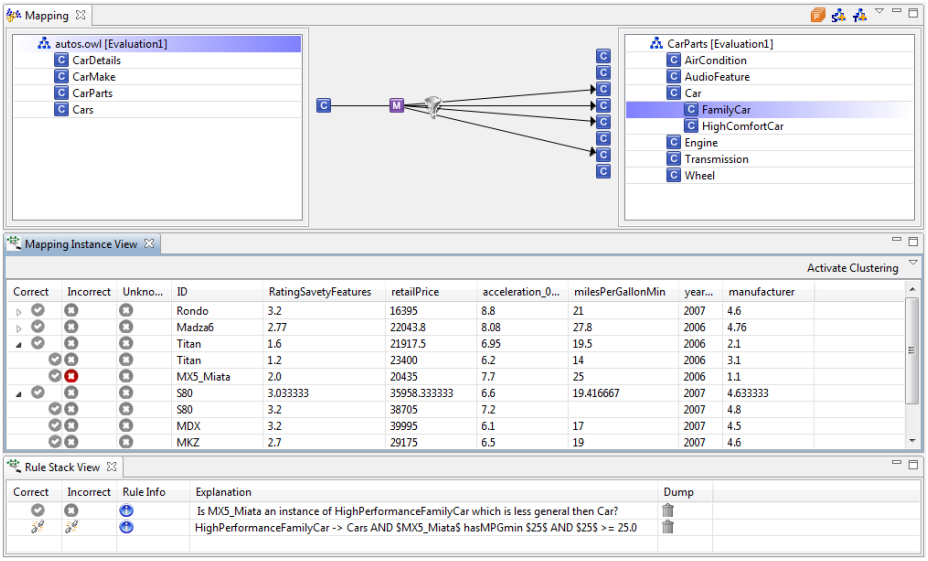


**Fig. 1.** MappingAssistant Perspective and Views

human-understandable questions (see Section 2) leads to the identification of the wrong rule or fact. If the wrong rule or fact has been identified, this information is also visualized in the Rule Stack View. The approach is implemented as an extension of the OntoStudio Ontology Engineering Workbench [4].

We demonstrate the use of MappingAssistant on a dataset constructed from a publicly available ontology from the car domain[3]. Due to license constraints we are not allowed to publish the MappingAssistant online, but we provide a video at `http://www.youtube.com/watch?v=72abBBTfl_E`[4].

## 2 Proof-Tree Algorithm for Diagnosis

We developed a proof-tree [5] based approach which minimizes the amount of questions the systems asks the user in order to determine wrong rules or facts. When the user depict an instance as incorrect, we generate a prooftree for the corresponding concept-assertion like *FamilyCar(MX5_Mieta)* in our example. Since the user evaluation is correct by assumption, the prolog-based proof-tree must contain at least one wrong node. In order to determine this wrong node our approach traverses the proof-tree in a way that the amount of user questions are minimized for a correct, as well as for an incorrect answer of the user. These questions are presented to the user in natural language based sentences, as shown in Figure 1.

[3] `http://gaia.isti.cnr.it/~straccia/download/teaching/SI/2006/Autos.owl`
[4] To best view the video adjust your resolution to 1080p.

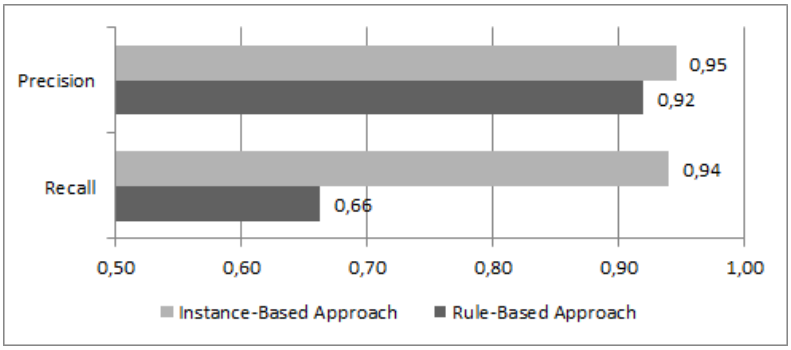


**Fig. 2.** Precision and Recall of Probands: Aggregates the precision and recall values for both, our instance-based and the state-of-the-art rule-based approach.

## 3 Evaluation

For evaluation we compared our instanced-based approach against a state-of-the-art rule-based one. In each task the probands were asked to find as many wrong mapping rules as possible within 10 minutes. Therefore, we constructed two test datasets containing 10 wrong mapping rules each. In our experiment we swapped the chronological order and the datasets used for the approaches. At the end, all probands had to answer a short questionnaire.

In the instance-based approach, probands where asked to identify wrong instances using the graphical user interface illustrated in Figure 1. Expert background knowledge concerning the instances was simulated by providing a sheet with representative instances and explanations. For the rule-based approach, probands had to identify the mapping-rules by checking the correctness of the mapping rule itself. In particular, mistakes had to be found either in wrong concept assignments or wrong filter attributes. For both tests, explanations on the functionality were provided beforehand.

We executed the study with 22 probands. The results of the probands' performance are displayed in Figure 2 and 3. Most of the wrong mapping-rules have been identified correctly in both, the instance- and rule-based approach, reaching precision values of 0.95 and 0.92, respectively. However, in case of the instance-based approach, almost all existing wrong mapping rules could be identified (recall 0.94), while in the rule-based approach participants missed one third of the wrong mapping rules (recall 0.66). In fact, almost all of the participants correctly identified more wrong mapping-rules in the instance-based approach than in the concept-based approach (91 % in Figure 3). Vice versa, none of the participants identified more wrong mapping-rules.

Overall, the study emphasizes that identifying wrong mapping-rules on the instance-level is more convenient. This is consistent with the results from the questionnaire, which can't be presented here due to space constraints.

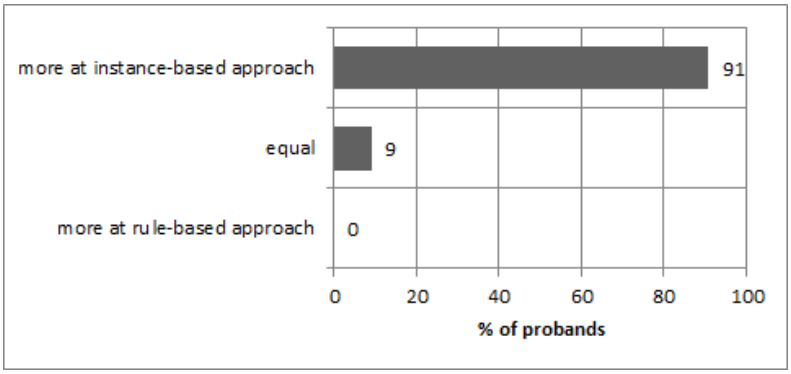


**Fig. 3.** Comparison of Number of Identified Wrong Mapping-Rules: 91% of the probands identified more wrong rules using the instance-based approach.

## 4 Conclusion and Future Work

We developed an interactive conflict resolution framework for identifying wrong rules or facts on the instance level. It includes data simplification techniques for avoiding the presentation of thousands of instances to the user and a diagnostic proof-tree algorithm which identifies the wrong rule or fact with the help of user questions asked in natural language. Furthermore, we verified our approach studying a user evaluation which confirmed the simplicity of our approach compared to state-of-the-art approaches.

In the second phase of our project we are concerned with correcting wrong mapping-rules. Our focus will lay on specializing mapping-rules utilizing inductive logic programming. MappingAssistant will be extended by a component which recommends improved specialized mapping-rules on the basis of background knowledge and individuals marked as correct or incorrect by the user.

## References

1. I. Cruz, F. Antonelli, and C. Stroe. AgreementMaker: efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, 2(2):1586–1589, 2009.
2. J. Euzenat, A. Ferrara, C. Meilicke, A. Nikolov, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn. Results of the ontology alignment evaluation initiative 2010. In *Proc. of ISWC workshop on Ontology Matching, Athens*, pages 73–95. Citeseer, 2010.
3. J. Hair, W. Black, B. Babin, R. Anderson, and R. Tatham. *Multivariate data analysis*, volume 7. Prentice hall Upper Saddle River, NJ, 2009.
4. A. Maier, H. Schnurr, and Y. Sure. Ontology-based information integration in the automotive industry. *The SemanticWeb-ISWC 2003*, pages 897–912, 2003.
5. A. Walker. Prolog/Exl, an inference engine which explains both yes and no answers. In *Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 1*, pages 526–528. Morgan Kaufmann Publishers Inc., 1983.