Chapter 1
Evaluating Semantic Web Service Technologies: Criteria, Approaches and Challenges

Ulrich Küster
Institute of Computer Science, Germany

Birgitta König-Ries
Institute of Computer Science, Germany

Matthias Klusch
German Research Centre for Artificial Intelligence, Germany

ABSTRACT
In recent years, a huge amount of research effort and funding has been devoted to the area of semantic web services (SWS). This has resulted in the proposal of numerous competing approaches to facilitate the automation of discovery, composition and mediation for web services using semantic annotations. However, despite of a wealth of theoretical work, too little effort has been spent towards the comparative experimental evaluation of the competing approaches so far. Progress in scientific development and industrial adoption is thereby hindered. An established evaluation methodology and standard benchmarks that allow the comparative evaluation of different frameworks are thus needed for the further advancement of the field. To this end, a criteria model for SWS evaluation is presented and the existing approaches towards SWS evaluation are comprehensively analyzed. Their shortcomings are discussed in order to identify the fundamental issues of SWS evaluation. Based on this discussion, a research agenda towards agreed upon evaluation methodologies is proposed.

INTRODUCTION
To foster reuse, state of the art software engineering has been driven over decades by the trend towards more and more component based software development. In recent years another trend towards more and more distributed and more loosely coupled systems could be observed. Service oriented architectures (SOAs) are the latest product of this long-reaching development. Web services in particular have become increasingly popular and are currently the most prominent implementation of a SOA. The grand vision of the web service paradigm is to have...
a rich library of ten thousands web services available online that provide access to information, functionality or resources of any kind and that can be easily integrated into existing applications or composed in a workflow-like fashion to form new applications.

Even though this promising technology has already proven to be an effective way of creating widely distributed and loosely coupled systems, the integration of the services is still labor intensive and thus expensive work. Thus – following the vision of the semantic web (Berners-Lee et al., 2001) – the idea of semantic web services (SWS in the following) was introduced (McIlraith et al., 2001), applying the principles of the semantic web to the web service paradigm.

SWS related research has attracted a huge amount of effort and funding recently. Within the sixth EU framework program alone, for instance, at least 20 projects with a combined funding of more than 70 million Euros dealt directly with semantic services. This gives a good impression of the importance being put on this field of research. The huge amount of effort (and money) spent into SWS research has resulted in numerous proposals of ontology based semantic descriptions for component services (Klusch, 2008b). Based on such descriptions, a plethora of increasingly sophisticated techniques and algorithms for the automated or semi-automated dynamic discovery, composition, binding, and invocation of services have been proposed (Klusch, 2008a).

However, despite of this wealth of theoretical work, recent surveys have shown that surprisingly little effort has been spent towards the comparative evaluation of the competing approaches (Küster et al., 2007b, Klusch and Zhing, 2008). Until recently there were no comparative evaluations and it was impossible to find two systems which had been evaluated on the same use cases. Evaluations were mostly concentrated either on artificially synthesized datasets under questionable assumptions or based on one or two use cases for which it was not clear, whether they were reverse engineered from the solution. In other words: "There are many claims for such technologies in academic workshops and conferences. However, there is no scientific method of comparing the actual functionalities claimed. [...] Progress in scientific development and in industrial adoption is thereby hindered" (Lausen et al., 2007).

There are striking parallels to this situation in the history of related areas:

"[in the experiments] …there have been two missing elements. First [...] there has been no concerted effort by groups to work with the same data, use the same evaluation techniques, and generally compare results across systems. The importance of this is not to show any system to be superior, but to allow comparison across a very wide variety of techniques, much wider than only one research group would tackle. [...] The second missing element, which has become critical [...] is the lack of a realistically-sized test collection. Evaluation using the small collections currently available may not reflect performance of systems in large [...] and certainly does not demonstrate any proven abilities of these systems to operate in real-world [...] environments. This is a major barrier to the transfer of these laboratory systems into the commercial world."

This quote by Donna Harman (Harman, 1992) addressed the situation in text retrieval research prior to the establishment of the series of TREC conferences in 1992 but seems to perfectly describe the current situation in SWS research. Harman continued:

"The overall goal of the Text REtrieval Conference (TREC) was to address these two missing elements. It is hoped that by providing a very large test collection and encouraging interaction with other groups in a friendly evaluation forum, a new thrust in information retrieval will occur."