REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-02-

0039

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		15 May 97 - 30 Jun 01

4. TITLE AND SUBTITLE
Scalable Knowledge Composition

5. FUNDING NUMBERS
F49620-97-1-0339

6. AUTHOR(S)
Gio Wiederhold

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Stanford University
651 Serra Street, Room 260
Stanford, CA 94305-4125

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
801 N. Randolph Street Room 732
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-97-1-0339

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The objective Scalable Knowledge Computing project was to enable reliable composition of information scalably from multiple autonomous sources. We enable interoperation among information sources by defining application-sensitive rules (articulation rules) that define precisely the correspondence among the terms used to describe the distinct resources, databases, knowledge-bases or information on the web. Anyone who needs information from multiple websites, since it is not available in one single site, is aware of the amount of effort required to perform the simplest of composition tasks. Our aim is to provide a system that makes reliable interoperation among information sources a reality.

20020215 103

14. SUBJECT TERMS	15. NUMBER OF PAGES
	3
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Final Report F49620-97-1-0339

Scalable Knowledge Composition

Gio Wiederhold

December 2001

The objective Scalable Knowledge Computing project was to enable reliable composition of information scalably from multiple autonomous sources.

We enable interoperation among information sources by defining application-sensitive rules (articulation rules) that define precisely the correspondence among the terms used to describe the distinct resources, databases, knowledge-bases or information on the web. Anyone who needs information from multiple websites, since it is not available in one single site, is aware of the amount of effort required to perform the simplest of composition tasks. Our aim is to provide a system that makes reliable interoperation among information sources a reality.

1. We addressed the original HPKB challenge problems, as set out by DARPA in 1997. While we, as a small independent project could not compete in scale and speed, we demonstrated that our answers were factually better, because we could access and combine source information. For instance, to obtain answers about OPEC and security council membership we accessed www.OPEC.com and www.UN.org in addition to the CIA factbook and generated correct answers, whereas the projects that relied only in the CIA factbook provided answers that were wrong relative to the real-world status, since the factbook did not provide the needed temporal information to recognize the lack of overlap among these two conditions for several countries. It is obvious that going to the sources is always more reliable than relying on a secondary compilation, and SKC enables that strategy [JSV98].
2. Our system is based on an interoperation system proposed by Karp [Kar96]. We extended it to not only work using databases, but also using knowledge bases and other information sources. In Karp's system, each database comes with a schema which is saved in a Knowledge Base of Databases(KoD). Correspondingly, we assume that associated with each information source is an ontology. However, we do not require all ontologies to be saved in a central repository like the KoD [MWK00, MWD01].
3. In order to match terms based on their meanings we processed two dictionaries, Webster's (public) and Oxford English (licensed), to enable matching based on a semantic network created from the links implicit in the words listed and their definitions, a nexus. These networks exceed by an order of magnitude those that have been manually created, as Wordnet. Using the Nexus repository we can, for instance, match `buyer' from a car-sales site with `owner' from a car registration site, even though there is no hint in the spelling of these words that they refer to the same set of people. We have applied this technique to information available

about NATO-countries governmental structures. The terms here vary greatly, as prime-minister vs. president, parliament vs. congress, and the like. We achieved an automatic match of 70% of the terms that had been linked manually. This capability will be crucial in many business and military situations, for instance when ordering materiel, supplies, and services from multiple autonomous suppliers and internal warehouses [Jan00].

4. We enhanced the articulation generator that matches terms in ontologies to include other heuristics based on word similarity and ontology graph structure. A word-relator, using a corpus of documents related to the topics of discourse, generates a similarity measure based on the context in which words appear. Words appearing in similar contexts get a higher score. A structural similarity generator compares two ontology graphs and tries to match terms that appear in similar "neighborhoods" in two ontologies. A weighted average of the scores generated by the several articulation generation heuristic routines gives us a score on the basis of which terms in ontologies are matched. Experiments done on two catalogues obtained from different sources in the construction industry show that we achieved a match of 70-80% with very few false positives.
5. No automatic method can reliably generate precise and minimal articulations. The articulations generated automatically need to be verified by a expert familiar to the two domains and the application for which the articulation is being generated. We have built a simple GUI prototype that displays the two ontologies, their articulation and enables the expert to ratify the articulation. The expert's response is logged and used in future articulation generation.
6. Our articulations are small intersections of the base terminologies and ontologies and hence easy to maintain, even as our knowledge improves, base capabilities change, and applications become more demanding. We expect that these ontologies will be combined in many important applications. To serve that requirement we have developed an algebra over ontologies, which allows reliable and arbitrary combinations of base and derived ontologies, providing scalability without massiveness. The algebra is the formal basis for enabling query optimizations. We have identified the properties of the algebraic operators. Query optimization algorithms depend heavily upon these properties and enables us to scalably compose information without compromising reliability [MW01].

References:

[JSV98] Jan Jannink, Pichai Srinivasan, Danladi Verheijen, and Gio Wiederhold: "Encapsulation and Composition of Ontologies"; Proc. AAAI Summer Conference, Madison WI, AAAI, July 1998.

[Kar96] Peter Karp: "A Strategy for Database Interoperation"; Journal of Computational Biology, Vol. 2, No. 4, pp 573-583, 1996.

[Jan00] Jan Jannink: "A Word Nexus for Systematic Interoperation of Semantically Heterogenous Data Sources", Ph.D. Thesis, Department of Computer Science, Stanford University, Stanford, CA,
2000.

[MWK00] Prasenjit Mitra, Martin Kersten and Gio Wiederhold: "A Graph-Oriented Model for Articulation of Ontology Interdependencies", In Proc. EDBT, Konstanz, Germany, Springer-Verlag, pp 86-100, 2000.

[MWD01] Prasenjit Mitra, Gio Wiederhold, and Stefan Decker: A Scalable Framework for Interoperation of Information Sources. In Proc. 1st International Semantic Web Working Symposium (SWWS `01), Stanford University, Stanford, CA, July 29-Aug 1, 2001.

[MW01] Prasenjit Mitra, Gio Wiederhold: An Algebra for Semantic Interoperability of Information Sources. In Proc. 2nd. IEEE Symp. on BioInformatics and Bioengineering, BIBE 2001, Bethesda, MD, Nov. 2001.