

Collaborative Research:  
**Developing a Learner-Centered Metathesaurus for Science,  
Mathematics, Engineering and Technology Education**

Final Report, December 2004<sup>1</sup>  
NSF Award DUE-0121743

## 1. Summary

SMETE.ORG is a comprehensive collection of high quality STEM (science, technology, engineering & mathematics)<sup>2</sup> digital learning resources and value-added services that contribute to the overall NSDL program at NSF. We allow authors and other submitters to *add* learning resources ‘anywhere, ‘anytime’ through our Web-based cataloging system. We review each learning object, consistent with library cataloging standards, to check for critical metadata fields and operation on one of the intended platforms. We use reviews/comments as quality feedback on highly used selected items. We have approximately 8,000 learning objects and 3,000 registered users. We support the National Science, Technology, Engineering and Mathematics Education Digital Library (NSDL) program by engaging in collaborative research, deployment and services related to enhanced harvesting, federated search, interoperability, scalability, usability, and personalization. We have developed extensive software environments and toolkits. With funding from another grant, we have developed a service that permits any collection or service to search our extensive repository of learning objects at [www.smete.org](http://www.smete.org). The service, based on the emerging SOAP specification utilizes standards-based methods to query the repository and receive item-level metadata of learning objects matching the query. We have also created a personalization service called *My Portfolio* that allows users to create a user profile and enables them to save found learning resources to their personal workspace. We also provide a recommender service using hybrid collaborative filtering, usage logs, and profile information. A strength of the SMETE.ORG approach has been to transfer cutting-edge research into the features and services it provides. A summary of the targeted research in metathesaurus development, implementation and testing is provided in this report.

## 2. Metathesaurus Development

Metathesaurus technology is used to formalize relations among structures of knowledge (expressed as linked thesaurus concepts) from many viewpoints and enable cross- and multi-disciplinary browsing and searching. A metathesaurus maintains a set of key concepts of the domain, semantic relationships between concepts such as synonyms, hyponyms (a word that is more specific than a given word), and also mappings from various source thesauri to the key concepts. Relatively few of these have been constructed; those that have been built were generally hand-coded at great expense. The process has typically been to figuratively overlap existing structures and “pin” them together at common points. Through this targeted research grant, SMETE.ORG developed a machine-learning framework from which to build metathesauri; preliminary results have been incorporated in NEEDS and SMETE.ORG.

---

<sup>1</sup> This is a collaborative grant with Dr. William H. Wood at the University of Maryland, Baltimore County. Dr. Wood’s final report will be submitted separately.

<sup>2</sup> Recently NSF switched the “T” and “M” in the ordering of the disciplines so that SMETE stood for SMET Education.

### 3. Summary of Foundation in Interoperability and Federated Search

This targeted research in metathesaurus development was greatly enhanced by the opportunity to deploy and test the results of the research within the interoperability framework of the SMETE.ORG site with partial funding from another NSF gran: DUE-0127580. Interoperability enables collections to extend their reach by increasing the opportunity for discovery of resources and providing additional resources related to its local collection. SMETE.ORG developed and implements federated search with multiple collections using two approaches: (1) harvesting and (2) agent-based federation [1,3-4]. Both approaches recognize the contributions and intellectual property<sup>3</sup> of partner collections and have developed in close collaboration with the partner. These low barrier-to-entry mechanisms enable the SMETE.ORG approach to meet or exceed the metadata requirements of NSDL Policy Committee and Core Integration team implementation (by using richer metadata element sets) and provide agreement-based access to metadata.

The harvesting approach relies upon the Open Archive Initiative-Protocol for Metadata Harvesting and custom gateways to build a centralized metadata repository from partners, as well as provide access to locally cataloged collections. In some cases (e.g., LON-CAPA and the Michigan Teacher Network), IEEE Learning Object Metadata-based local application profiles is transformed to normalized application profiles using tools developed by SMETE.ORG as it is harvested.

The second approach federates collections by deploying agents to perform searches of a partner's metadata, report back to the portal and present aggregated search results. SMETE.ORG developed a federated search server-client specification to enable others to remotely query SMETE.ORG , and uses this as a basis for how it interoperates with other collections. A partner collection implements a web-services framework to describe their metadata format, query and response mechanisms supported and other key features supported by their client via a Web Service Definition Language (WSDL) file. SMETE.ORG is then able to query the partner collection and aggregate its results with simultaneous queries of other collections (returned via a SOAP transport layer)<sup>4</sup>. In return the partner collection is able to query the SMETE.ORG catalog in a similar fashion. Several other digital libraries have taken advantage of our API for interoperability. The service is closely modeled on the Lucene query syntax, the Open Archives Initiative Protocol for Metadata Harvesting structure for the response, and the Google Web API for the request envelope, thus reducing the amount of new code collections and services must write to utilize the service if they desire to develop their own client applications. We also provide ready-to-implement clients in both Perl script and Java packages for our partners for fast prototyping of federated search service with us. SMETE.ORG and MERLOT have implemented this specification to enable cross-federated search between the NEEDS and MERLOT collections.<sup>5</sup>

---

<sup>3</sup> SMETE.ORG policy states that it will not retransmit harvested metadata, without permission, in accordance with agreements with partners.

<sup>4</sup> Paepcke, A., R. Brandriff, G. Janee, G. Larson, B. Ludaeschre, S. Melnik and S. Raghavan, Search Middleware and the Simple Digital Library Interoperability Protocol, D-Lib Magazine, 6(3), March 2000.

<sup>5</sup> For more details see [www.SMETE.ORG/smete/public/soap/index.jhtml](http://www.SMETE.ORG/smete/public/soap/index.jhtml) and [fedsearch.merlot.org/main/search.jsp](http://fedsearch.merlot.org/main/search.jsp).

Both harvesting and agent-based federation are challenging because each collection provider uses a custom metadata format, along with particular program and storage resources, to store metadata for materials in their collection. SMETE.ORG 's interoperability efforts have benefited greatly from the development of a metadata cross-walk tool for use with arbitrary data sources that converts metadata stored in arbitrary formats and in different access containers to representations of IEEE Learning Object Metadata or Dublin Core XML bindings.

SMETE.ORG has also implemented tools to enable advanced interoperability of NSDL collections. SMETE worked with LON-CAPA ([www.lon-capa.org](http://www.lon-capa.org)) to integrate SMETE.ORG 's personalization service with their learning management system, and Utah State University to send a user's collection of digital learning resources in *My Workspace* to the Instructional Architect ([ia.usu.edu](http://ia.usu.edu)).

The following collections are now searchable from the [www.SMETE.ORG](http://www.SMETE.ORG) portal using our various forms of interoperability:

1. Center for Highly Interactive Computing in Education: 3 learning objects
2. Computer Science Teaching Center: 70 learning objects
3. DLESE: 1,904 learning objects
4. Eisenhower National Clearinghouse: 31 learning objects
5. LearningOnline Network with CAPA: 3,168 learning objects
6. Michigan Teacher Network: 85 learning objects
7. Math Forum: 20 learning objects
8. Mathematics Association of America: 10 learning objects
9. National Library of Virtual Manipulatives for Interactive Math: 154 learning objects
10. NEEDS (National Engineering Education Delivery System): 1,788 learning objects
11. Stanford Paper Bike Project: 377 learning objects

SMETE.ORG participated with the Core Integration team and the Digital Library for Earth System Education (DLESE) in the deployment and testing of the DLESE OAI tool. We have made the National Engineering Education Delivery System (NEEDS) Premier Award Winners from 1997-2003 available at <http://www.SMETE.ORG/oai/provider?verb=Identify>.

#### **4. Metathesaurus Application to Interoperability and Federated Search**

A major challenge to federated search across multiple collections concerns the different vocabularies used by collection providers. Lacking a standardized controlled vocabulary set, each collection provider often employs their own vocabularies to describe the content of the resource, the metadata fields, the various pedagogical uses of the resource and the general categories used to classify the resource. Different organizational structures and goals can even lead to different vocabularies within the same discipline. This makes searching over resources from multiple providers difficult and inefficient in that a user is incapable of being familiar with all controlled vocabularies employed by different collections and knowing their differences in order to conduct a successful query. Developing a meta-thesaurus helps to bridge the differences in thesauri from various collections. It also enables us to develop a recommender mechanism to automatically expand queries submitted by the user by including similar vocabularies in various

collection thesauri that the user is not familiar with, hence increasing the effectiveness of the federated search.

Our approach in metathesaurus generation is to identify a candidate list of keyphrases through computational text analysis and then use latent semantic analysis to identify useful synonyms. The synonyms are used to extend the search terms used in a user query with the goal of increasing the recall rate across multiple disciplines.

Keyphrases have been used extensively in IR systems to facilitate information exchange, organize information and assist information retrieval. Automation of keyphrase generation is essential for the timely creation of keyphrases for large repositories in new domains where previous thesauri do not exist or for metacollections in which keyphrases that are meaningful across disparate collections are needed. We developed an automated keyphrase extraction algorithm using a non-dominated sorting multi-objective genetic algorithm. The “clumping” property of keyphrases is used to judge the appropriateness of a phrase and is quantified by a condensation clustering measure proposed by Bookstein. The objective is to find the smallest phrase set that has the best precision, as measured by average condensation clustering. We implemented a multi-objective genetic algorithm to find the optimal set of keyphrases that displayed the lowest dispersion (highest condensation) level and minimized the number of phrases. Using our framework, additional preferences could be added to the optimization model if needed. Trade-off information between different objective functions can be gained from the final generation. More details on this approach can be found in [2].

Incorporating the results from the keyphrase extraction and latent semantic analysis researches, we’ve developed and implemented a tool for suggesting related terms when users submit queries on SMETE. This tool can assist users in recognizing their true information needs and identifying the most appropriate query terms (phrases) when the users are not familiar with the terminologies in the target domain. The phrase set we use in our system draws from three sources: keyphrases extracted with the algorithm mentioned in the previous paragraph, subject headings assigned to the resources, and past queries submitted by the users. We associate the keyphrases and subject headings with the resources they originated from. User queries are associated to the resources users downloaded/viewed subsequently after they submit the queries. Latent semantic analysis is then carried out on these associations to find the synonyms and antonyms among the phrases. Currently we’ve provided the service using this tool through our search result pages. If related terms are found based on our synonym database, they are shown on the search result pages. Users are provided with the chance of searching with those terms again by clicking on the link associated with the terms if they think the terms are more appropriate than their original queries. An example of the query recommender tool in action can be seen in Fig. 1 on the following page or viewed at: [http://www.needs.org/needs/public/search/search\\_link.jhtml?keyword=design](http://www.needs.org/needs/public/search/search_link.jhtml?keyword=design).

## **5. Conclusions and Recommendations**

In order to test out the automatic generation of keyphrases for use in a metathesaurus, a human evaluation procedure was carried out to assess the quality of extracted phrases, using a controlled set of abstracts from a collection of design conference papers. Our results report that over 90% of the phrases were acceptable keyphrases for engineering design. In addition, 80% of the author-

assigned keyphrases used in the documents showed up in the generated lists as well. The results indicate that our proposed MOGA algorithm can extract a reasonably good keyphrase set just by processing a collection of documents in a particular domain without any prior training or domain-specific knowledge. We believe such an approach could greatly reduce the effort of developing domain-specific thesauri and updating established thesauri more efficiently. We envision that the generated phrases as a whole can be an efficient indexing tool as well as a tool for reformulating user queries through query expansion. A subset of the generated phrases that are pertinent to a certain topic can also be used to assist authors or editors assigning keywords to academic papers on that topic.

By examining the distribution of occurrences of keyphrases in the final population, we found that phrases tend either to be present in almost all solutions or exist in very few solutions. We hypothesize that phrases that have fewer occurrences are the ones that can be used to tune the optimals along the Pareto curve. A detailed sensitivity analysis will be needed to validate this hypothesis.

We are currently conducting a more extensive evaluation for a metathesaurus across engineering disciplines in the area of engineering education in the NEEDS collection within SMETE.ORG. The evaluators, in this case, are engineering faculty and students in higher education.

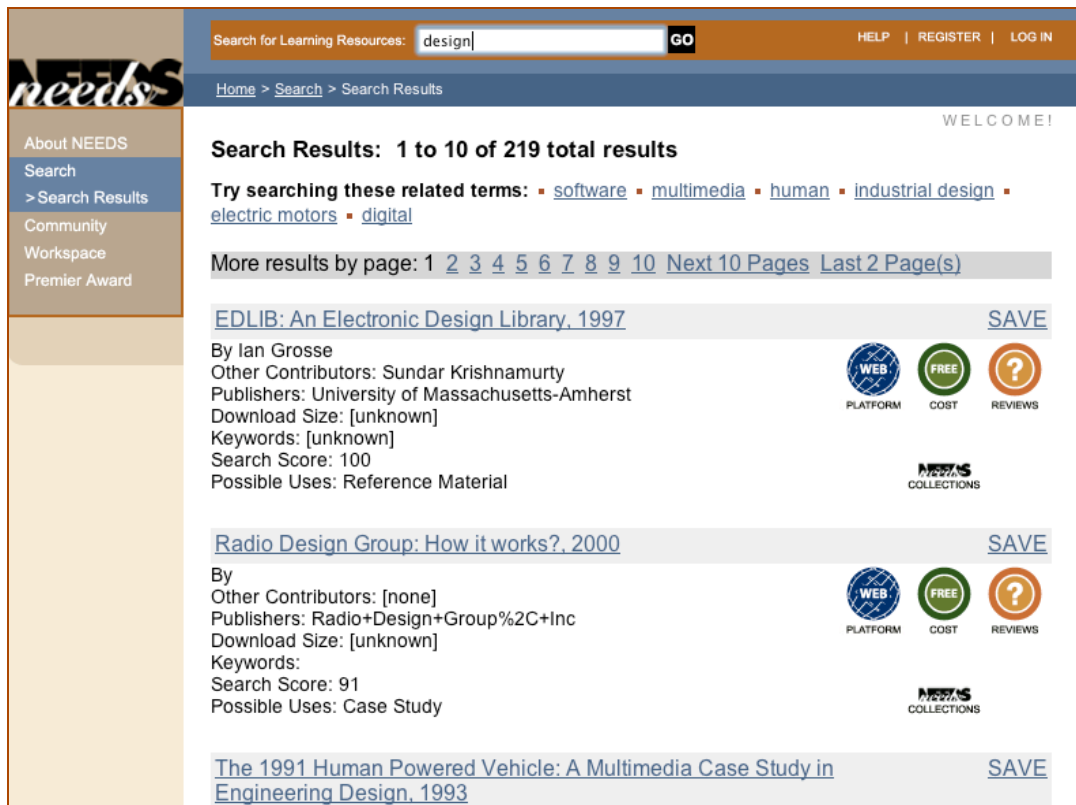


Fig. 1: Results using the query "design" in NEEDS.ORG. The system recommends a recursive search over more specific related terms of "software", "multimedia", "human", "industrial design", "electric motors" or "digital".

## 6. Publications and Products from NSF Grant DUE-0127580

With partial funding from this grant, the following publications, outreach presentations and web products were completed.

### Conference Proceedings

1. Dong, A., E. Fixler and A.M. Agogino, "A Case Study of Policy Decisions for Federated Search Across Digital Libraries," *Proceedings of ICDL 2004* (International Conference on Digital Libraries).
2. Wu, J. and A. M. Agogino, Automating Keyphrase Building with Multi-Objective Genetic Algorithms, " *Proceedings of the Hawaii International Conference on System Science*, HICSS, 2004.

### Technical Reports

3. Dong, A., E. Fixler, A.M. Agogino, M.J. Koning-Bastiaan, and S. Shamseldin, "A Web Services Approach to Federated Search Across Digital Libraries". Working Paper # 03-0103-3, 2003. Berkeley Expert Systems Technology Laboratory, 6102 Etcheverry Hall, UC Berkeley, Berkeley, CA 94720-1740.
4. Wu, J., E. Fixler, and A.M. Agogino, "Translating Between Native Java Objects and LOM: A Tools to Assist Digital Libraries Exchange Information." Working Paper # 03-0203-3, 2003. Berkeley Expert Systems Technology Laboratory, 6102 Etcheverry Hall, UC Berkeley, Berkeley, CA 94720-1740.

### Outreach Presentations

5. Agogino, A.M. (for L. Zia, 2001). "Successful Partnering" at the "Forging Library Partnerships in the Networked Age," Clark Kerr Campus, UC Berkeley, Nov. 2, 2001.
6. Agogino, A.M., "Bringing the Educational Experience of NEEDS and SMETE.ORG to NSDL", AAAS Annual Meeting: Science as a Way of Life, (13-18, Feb. 2003, Denver, CO), CD ROM.
7. Agogino, A.M., "Ubiquitous Wireless Infrastructure to Support Mobile Learning," HP/CITRIS 2004 Workshop on Planetary-Scale Applications, Wed., May 26, UC Berkeley.
8. Dong, A. (2001). Enhancing Interoperability for the National SMETE Digital Library Program, UC Berkeley Digital Library Seminar, October 15, 2001.
9. Dong, A. and Agogino, A. (2002). Who's Out There? Building Community through Recommendations, MERLOT International Conference 2002, Atlanta, GA.
10. Dong, A., Koning-Bastiaan, M. & Muramatsu, B. (2002). Implementing Federated Search Across Collections, MERLOT International Conference 2002, Atlanta, GA.
11. Dong, A. (2002). Planning to Make the Best Use of the National Science Digital Library, Project Kaleidoscope Workshop on The World Wide Web: Strengthening the Undergraduate Learning Environment, United States Air Force Academy, Colorado Springs, Colorado, February 8 - 10, 2002.

## **Websites and Web Reports**

12. SMETE digital library at [www.smete.org](http://www.smete.org). The SMETE digital library is the gateway to a comprehensive collection of science, mathematics, engineering and technology (SMET) education content and services to learners, educators, and academic policy-makers.

## **PhD Dissertations**

13. Jia-Long Wu, “Building Subject Headings for a Unified Language System for Engineering Design,” doctoral dissertation, UC Berkeley, Expected submission: Spring 2005.

## **Acknowledgements and Disclaimer**

This report is based upon work partially supported by the National Science Foundation under Grant No. DUE-0127580. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.