2005

A Component-Based Approach for Scientific Services for Education and Research (Scientific SEARCH)

Rajani S. Sadasivam
University of Massachusetts Medical School

Murat M. Tanik
University of Alabama - Birmingham

Linda Casebeer
University of Alabama - Birmingham

See next page for additional authors

A Component-Based Approach for Scientific Services for Education and Research (Scientific SEARCH)

Authors
Rajani S. Sadasivam, Murat M. Tanik, Linda Casebeer, David Allison, Jill Gemmill, Jason Lynn, Barrett Bryant, Yi-Fang Wu, Michael Bieber, and Leon Jololian

A Component-Based Approach for Scientific Services for Education and Research (Scientific SEARCH)

Rajani Sadasivam[1], Murat M. Tanik[1], Linda Casebeer[1], David Allison[1], Jill Gemmill[1], Jason Lynn[1], Barrett Bryant[1], Yi-Fang Wu[2], Michael Bieber[2], Leon Jololian[3]
[1]*University of Alabama at Birmingham;*
[2]*New Jersey Institute of Technology;*
[3]*New Jersey City University*
rajani@uab.edu, mtanik@uab.edu

Abstract

Today's challenge for retrieving digital information by users such as "students," "educators," or "researchers" is coping, more than ever before, with the excessive data and information available. The problem is further compounded because of the way scientific knowledge is structured, in terms of expert interviews, articles, conference coverage, journal scans etc. Great progress has been made in digital library research. The NSF/NSDL through their initiatives has assembled a great set of tools and techniques that hold significant potential. Many projects are now underway applying these tools and techniques to meet the information needs of different user communities. The primary focus of Scientific SEARCH project is enhancing access to high-quality learning materials and resources, modules, and other digital objects targeted towards scientific consumer and scientific producer. The project will use a multi-phased approach to achieve the objective. The paper describes the first-phase work submitted to NSF 04-542 solicitation.

Index Terms—***Component Based, Digital Library Research, Web services.***

1. Introduction

What information consumes is rather obvious; it consumes the attention of its recipients.

Herb Simon[1]

The great promise of the Internet is that it has the potential to provide us with a wealth of reliable resources and information to improve our productivity. Today's challenge for retrieving digital information by scientists ("students," "educators," or "researchers") is coping, more than ever before, with the excessive data and information available. As the number and content of information sources increases, the demand for further efficiencies in access also increases. The problem is further compounded because of the way scientific knowledge is structured, for example in terms of expert interviews, articles, conference coverage, and journal scans. Hence, it is difficult for users to find answers to specific questions that will be directly applicable to their needs and assess their quality. In recent studies, health care specialists have cited the largest barriers to seeking scientific information on the Internet as too much information to scan and information that is not specific enough to address their particular questions [2, 3]. Everyone seems to be overwhelmed with too much information that consumes too much time and energy to retrieve – efficient criteria and services for finding user-specific certified high-quality information are needed.

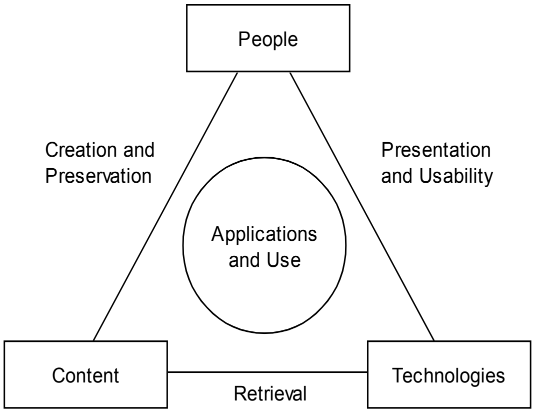

Figure 1. Conceptual Model of digital library research [4]

2. Focus of Scientific Search digital library research.

Table 1: Motivational Scenario

Scientific Consumer

An attorney at the Federal Trade Commission (FTC) calls Dr. X, a university professor, to help FTC with a case they are pursuing. The case involves a company that is marketing a dietary supplement made from leaves of the lemon tree and claiming that it is safe and effective for producing weight loss. FTC asks Dr. X to prepare a report giving her expert opinion as to whether there is *competent and reliable scientific evidence* that the product is safe and effective for producing weight loss.

To do so, Dr. X will **(1)** attempt to gather all information that directly bears on this point; and **(2)** synthesize that information into an opinion and a report. Item **(1)** is our focus here.

If Dr. X is being *very thorough*, in addition to relying on her accumulated background knowledge, Dr. X might:

1) Conduct a search of multiple reference databases including, but not necessarily limited to:

A) PubMed
B) Science Citation Index
C) Agricola
D) Psycinfo
E) Dissertation Abstracts International

2) Conduct a search of the abstracts of conferences that carry such material (e.g., the annual *Experimental Biology* meetings) which are posted on the web.

3) Conduct a search of the United States Patents database (WWW.USPTO.gov) for related patents and patent applications.

4) Conduct a search of the USDA, NIH, and NSF's online searchable databases of funded grants to identify any ongoing work related to this topic.

5) Conduct a search of the FDA's website and adverse event databases.

6) Conduct a search of the European Union's patent database.

7) Search *Lexus-Nexus* databases of legal cases for any relevant information.

8) Search the National Toxicology Program's database of reports (http://ntp-server.niehs.nih.gov/index.cfm?objectid=7DA86165-BDB5-82F8-F7E4FB36737253D5)

8) Combine all of the information retrieved into a unified 'pre-report' eliminating duplicate references and abstracts that appeared in more than one database.

9) Download the full text PDF version of all documents that are freely available to Dr. X.

10) Download the full text PDF version of all documents for which the fee is less than some dollar amount (e.g., $30).

11) Search *Books in Print* for any books on this topic, add the references to the unified pre-report, and then order all books that are available from the University library and purchase those that are not but are available for less than ***D*** dollars from www.Amazon.com.

12) Obtain the email addresses for the corresponding authors of all documents identified and send them an email asking for a PDF of their document that Dr. X has not yet acquired it and also asking if they are aware of any new information on this topic that they can share.

13) Combine all information obtained into a unified and organized 'pre-report' that simple lists/contains what was identified and from which source.

14) Store all information on a single directory on Dr. X's hard drive.

Scientific Provider

Dr. Y studies body weight and energy regulation using a variety of approaches including secondary analysis of archival data. Working from an evolutionary biology perspective, he hypothesizes that recent secular increases in obesity rates are, in part, the result of assortative mating for adiposity (fatness).

Assortative mating is a pattern of nonrandom mating, most often *positive assortment*, in which the probability that 2 individuals mate is positively related to their degree of phenotypic similarity. Assortative mating increases genetic variance in a population even though it does not affect allele frequencies (it does affect genotype frequencies). It can be shown that accepting 3 propositions implies that assortative mating is contributing to increased obesity prevalence: 1) human adiposity variations have a genetic component, 2) the adiposity threshold for defining obesity was historically above the population median, and 3) humans assortatively mate for adiposity.

Dr. Y believes that ample evidence already exists to support propositions 1 and 2, but wishes to evaluate whether evidence is well established in the literature for proposition 3 and, if not, he wishes to conduct such a study using archival data.

To do so, Dr. Y takes the following steps:

1) He thoroughly searches for and examines the extant literature on this topic. He begins by executing a literature search and retrieval 'expedition' much like that described for Dr. X in our other hypothetical scenario.

2) Dr. Y then searches for additional articles that have spouses in large cohort studies where indicators of adiposity (e.g., body mass index, BMI; Kg/m^2) are likely to be available. He wants to find all the existing datasets that information on BMI for spouses so that he can evaluate the spousal correlation for BMI as an indicator of assortative mating for adiposity.

Each promising article is downloaded and then checked for a statement about whether the data are publicly available. If so, steps are taken to obtain the data. If no such statement is made, an email address is obtained from the corresponding author either from the paper itself or by going to the institutional website directory, and a letter is emailed to the author in which Dr. Y introduces himself and his project and asks if the data can be made available.

3) Dr. X visits websites of well-known data repositories including, but not limited to:

A) The ICPSR (http://www.icpsr.umich.edu/)
B) The Henry A. Murray Research Archive (http://www.murray.harvard.edu/mra/index.jsp)
C) The US HHS archive (http://www.hhs-stat.net/scripts/result.cfm?lk=5)
D) The UK Data Archive (http://www.data-archive.ac.uk/)

4) Dr. Y conducts a Google search to identify additional data archives he may not be familiar with and subsequently searches them.

5) Dr. Y searches each data archive for all data sets containing spouses or couples prior to marriage and that contain height and weight (the information needed to calculate BMI) on the participants. He ascertains which datasets are freely available and downloads all of those datasets and their accompanying documentation to an organized file structure on his hard drive.

6) For any useful dataset available at a cost, he pays for and downloads any costing less than some amount he prespecifies.

7) Finally, for each dataset, he files an Institutional Review Board (IRB) exemption request to allow him to analyze archival data on humans that contains no identifying information. In this situation, when the purpose of the study, PI, etc are the same for each study, completing the IRB exemption form is exceptionally simple and can be done in a rote manner, simply replacing the name of the dataset on each form.

Great progress has been made in digital library research [1,5]. The NSF/NSDL through their initiatives has assembled a great set of tools and techniques that hold significant potential. Many projects are now underway applying these tools and techniques to meet the information needs of different user communities. The primary focus of Scientific SEARCH project is automating/aiding the steps described in Table 1 (Motivational Scenario Table). The project will use a multi-phased approach to achieve the objective. The paper describes the first-phase work submitted to NSF 04-542 solicitation.

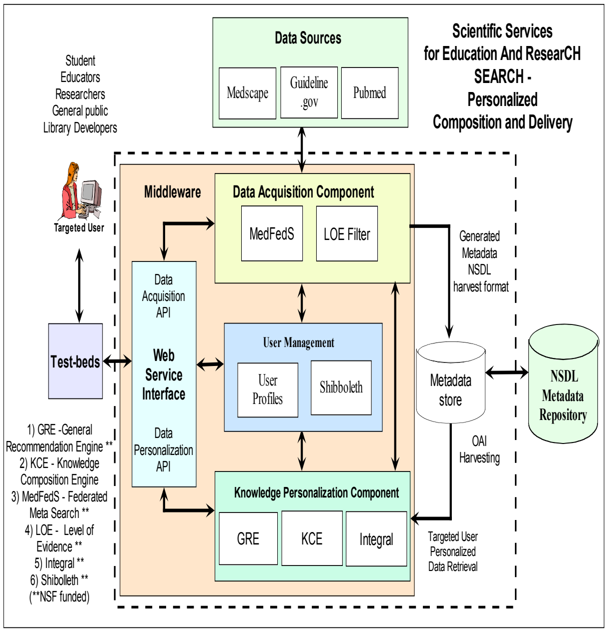
Student
Educators
Researchers
General public
Library Developers
Data Sources
Medscape
Guideline .gov
Pubmed
Scientific Services for Education And ResearCH SEARCH - Personalized Composition and Delivery
Targeted User
Middleware
Data Acquisition Component
MedFedS
LOE Filter
Generated Metadata NSDL harvest format
Data Acquisition API
Web Service Interface
Data Personalization API
Test-beds
User Management
User Profiles
Shibboleth
Metadata store
NSDL Metadata Repository
OAI Harvesting
Knowledge Personalization Component
GRE
KCE
Integral
Targeted User Personalized Data Retrieval
1) GRE -General Recommendation Engine **
2) KCE - Knowledge Composition Engine
3) MedFedS - Federated Meta Search **
4) LOE - Level of Evidence **
5) Integral **
6) Shibolleth **
(**NSF funded)

Figure 2: Scientific Search Component Approach

3. Impact of first phase of Scientific Search project

In the Scientific SEARCH project, obesity is used as an archetype for interdisciplinary research where needs, documents, and information span many disciplines, utilize many approaches, and are growing at an extraordinary rate [6]. NSF/NSDL has noted that even after considerable achievements in digital library research, there are still challenges in interdisciplinary information access. In the conceptual model (Figure 1) depicted in "Knowledge Lost in Information" [1] and detailed in the "DELOS-NSF Working Group" report [4] they strongly indicate that "interdisciplinary research areas along the three edges and the center of the triangle" are focused areas that traditional research teams currently neglect. In the first phase ([7, 8]) "Scientific SEARCH" project expects to significantly advance the state-of-the art in the critical areas of "retrieval" and "presentations and usability" through applying existing techniques and concepts from different domains. By leveraging prior NSF supported projects at New Jersey Institute of Technology (NJIT) (http://web.njit.edu/~wu/grants.php, [9-19]) and UAB [20-24], the Scientific SEARCH project will directly address interdisciplinary needs with a focus on users interested in the domain of obesity-related research and education. We expect to make the following significant contributions to state-of-the-art:

1. Deliver "actionable information" through utilizing level of evidence-based practice guidelines (**certified and graded quality**) (Table 1, [25]), and
2. Enhance knowledge personalization through integrating Knowledge Composition Engine (KCE) with NSF-sponsored General Recommendation Engine (GRE) and

IntegraL integration infrastructure (**need-focused presentation**) [9-19].

4. Proposed work

The project will design and implement lightweight and customizable Scientific SEARCH middleware services (depicted in Figure 2) that will greatly enhance the ability to find higher quality information. The efforts will be three-fold. First, the project will leverage upon existing NSDL approaches (GRE, IntegraL, MedFeds http://web.njit.edu/~wu/grants.php [9-19]) in our development efforts. In addition, the project will develop KCE to increase efficiency of access to information. Second, the project will develop federated administration of authentication and authorization systems using NSDL recommended Shibboleth [26] to integrate other library communities into the system. Third, the project will develop test-bed services to demonstrate the usage of library services and encourage participation within and outside the collaborating institutions. The generic criteria and services that we develop for identifying and suggesting resources within the obesity domain will be efficiently and effectively applicable to other domains.

5. Scientific Search approach.

5.1. Web Services integration.

The project will use Web Services approach for integrating the different components of the Scientific SEARCH. IBM defines a Web service as a collection of functions that are packaged as a single entity and published to the network for use by other programs [27, 28]. One early example is Microsoft Passport, a convenient authentication service hosted by Microsoft. Web services are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. In the Web Service model, Extensible

Markup Language (XML) based standards Simple Object Access Protocol (SOAP) is used for communication, Web Services Description Language (WSDL) is used for Service definition, and Universal Description, Discovery and Integration (UDDI) is used for Service discovery [29, 30]. Each of the Scientific SEARCH components will be encapsulated with Web Service wrappers (WSDL) and invoked in sequence based on users needs. SOAP will be used as the messaging protocol.

5.2. Federated Authentication and Authorization

The Scientific SEARCH project will use NSDL recommended Shibboleth [26] architecture, based on open standards for Federated Identity and Web Services, for federated administration of authentication and authorization of the Scientific SEARCH. Federations are formed via a trust relationship among a set of resource providers and consumers. The resource provider trusts each federation member to authoritatively identify members of its own community and to provide attribute information necessary at a remote access decision point. The Scientific SEARCH system will use existing authentication systems to establish user identity; for example, at UAB the single sign on "blazerID" system is already well known and used by all members of the university community for access to financial, human resource, and course-based information. Attributes available from existing authoritative directories will be combined with Scientific SEARCH's specific role information; for example, those identified for the role "content editor" will be provided with different access permissions than content users [20-23]. Using this approach, each person will gain access to the system using familiar log-in credentials and will navigate and access the system according to role/attribute information made available through Shibboleth. This approach assembles a complex system from existing components that can be distributed while sharing a common authentication and authorization system environment. As a result, the user can experience customized and integrated content presented from multiple sources without requiring a single monolithic system at the backend.

5.3. Level of Evidence based practice guidelines for actionable information

Over the past two decades, many attempts have been made to synthesize health related scientific literature into practice guidelines for the purpose of developing benchmarks of care. As these practice guidelines were being developed, it became clear to the panels authoring them that there were various levels of medical evidence and that these various levels created dilemmas for those attempting to draw conclusions from the literature. Various systems has been developed for grading evidence, but in all cases the attempt has been to address the continuum of evidence from meta-analyses, based on a broad spectrum of anecdotal and well-controlled clinical evidence. The two most common sources of such information are: 1) the Agency for Healthcare Quality and Research's National Guideline Clearinghouse (AHQRNGC, www.guideline.gov); and 2) the Cochrane Collaboration (www.cochrane.org). The Cochrane Collaboration attempts to control the level of evidence by reviewing only the literature considered to be the highest level of evidence, i.e. data produced by randomized controlled trials that meet strict eligibility criteria. A number of systematic Cochrane Reviews have been conducted related to obesity. Some examples include [31-33]. AHQRNGC, however, includes thousands of guidelines and relies on a system of grading the evidence. This grading of evidence allows students, researchers, healthcare practitioners and their patients, the lay public, to determine the strength of the evidence used to make specific recommendations.

Table 2: Conclusion Grades

Grade I:	The evidence consists of results from studies of strong design for answering the question addressed. The results are both clinically important and consistent with minor exceptions at most. The results are free of any significant doubts about generalizability, bias, and flaws in research design. Studies with negative results have sufficiently large samples to have adequate statistical power.
Grade II:	The evidence consists of results from studies of strong design for answering the question addressed, but there is some uncertainty attached to the conclusion because of inconsistencies among the results from the studies or because of minor doubts about generalizability, bias, research design flaws, or adequacy of sample size. Alternatively, the evidence consists solely of results from weaker designs for the question addressed, but the results have been confirmed in separate studies and are consistent with minor exceptions at most.
Grade III:	The evidence consists of results from studies of strong design for answering the question addressed, but there is substantial uncertainty attached to the

	conclusion because of inconsistencies among the results of different studies or because of serious doubts about generalizability, bias, research design flaws, or adequacy of sample size. Alternatively, the evidence consists solely of results from a limited number of studies of weak design for answering the question addressed.
Grade Not Assignable:	There is no evidence available that directly supports or refutes the conclusion.

These schemes for grading evidence can be applied to any content related to healthcare. A number of practice guidelines on the management of obesity are catalogued by the AHQRNGC. These guidelines assign grades or levels of evidence to the information reviewed to produce guidelines. An example from the Institute for Clinical Systems Improvement (ICSI)'s Prevention and management of obesity (mature adolescents and adults) is shown in Table 2 [25].

5.4. Web Based portal environment.

A web-based portal will be developed to provide user access to the different features of the Scientific SEARCH. The project will customize the Content Management Software (CMS), Drupal, for its needs [34]. Based on our experiences, the advantage of using Drupal is multifold:

- Drupal provides a well established and tested software base. Many companies and research organizations use Drupal for their Web portal needs.
- Drupal provides a structured Web framework on which the project can develop the necessary interfaces for the researchers working on the project. Drupal's modular approach provides for easier integration. Drupal modules can be written that can invoke the Web Service components to provide access to the features of the Scientific SEARCH. Once developed, these modules can easily be integrated with the CMS following well documented instructions [34].
- The rich toolset that Drupal provides, such as revision control, user management, and other administration features can be reused for our software system. The projects expect that this approach would greatly reduce the development time of the Scientific SEARCH.
- Drupal also provides a content management feature, which can be used to document high level descriptions of the project thus proving an informative portal for our research collaborators. Since Drupal uses the MySql database it is easy for us to extend the Drupal database to build the necessary tables required by the Storage component.
- The collaborative features of Drupal provide us an opportunity to conduct online collaboration between the participants sited at diverse locations.

Other CMS provides similar features but Drupal compares favorably to them according to independan analysis. [35] .

5.5. Components of Scientific Search

The different components of the Scientific SEARCH (Figure 2) will be exposed through a lightweight Web Service API using standardized protocols SOAP and WSDL.

5.5.1. Data Acquisition Component.

MedFedS Federated Search: Scientific SEARCH project proposes to develop a federated search, MedFedS, which combines search results from Medscape.com, PubMed, and the National Guidelines Clearinghouse. MedFedS will be based on NJIT's Highlight metasearch engine [36]. When a user inputs a query, MedFedS dispatches the query to the above three medical portals and databases through corresponding wrappers. Each wrapper translates the query into an acceptable request to its search engine, submits the request, retrieves documents, and parses them into an internal structured format. MedFedS merges and ranks these documents by the averages of their orders in the individual sources. Merging consists of duplicate removal and document ranking. To rank document snippets, a relevancy indicator is computed. The relevancy indicator for a document snippet is defined as the average of the rank orders of the document snippet, in all sources. The federated search will help users obtain more relevant documents from more sources.

Level of Evidence (LOE) filter: Unlike regular search engines that organize returned documents solely based on the similarity between a query and all search returned hits, MedFedS will combine document similarity measures and level of evidence (LOE) criteria for the ranking of relevant documents. Therefore, we anticipate that when returned hits are augmented with LOE, users can easily skim through the list of returned documents and identify useful ones. We will investigate how adding LOE reduces the recall effort needed to locate sufficient relevant documents. We shall follow the approach that proved successful for our Highlight metasearch engine, but make a research contribution by customizing it for the medical domain and adding LOE ranking criteria. In order to use LOE as ranking criteria, we will ask medical researchers to help us identify sample documents for each evidence level to obtain keywords. To produce a keyword list for each evidence level, medical experts familiar with medical study procedures will help us mark sample documents with keywords associated with the LOE they are assigned to. We will use identified evidence keywords from an evidence level to form a pseudo document, and previously unprocessed documents will be classified at the level with which they share the highest similarity value. The LOE information will be recorded as part of metadata for each document, once it becomes available. During a search session, after returned documents are collected from each digital collection by the MedFedS, they will be combined to form a unified list. For all documents with the LOE recorded in the metadata database, the information will be retrieved and displayed as part of a returned hit. For previously unprocessed documents which do not have corresponding LOE, the LOE filter described above will be applied to determine this information, based on the evidence keywords they contain. Users can choose to receive returned hits by similarity only, by LOE only, or by the combination of the two. Our current design of the last ranking method, which combines similarity and LOE, will sort returned documents the following way: for each relevance interval, (91%-100%, 81-90%, etc), documents will be sorted based on LOE. For example, a returned document of level 3a evidence that is 97% relevant to the query will be ranked lower than a document of level 1a evidence and 93% relevant. A catalog will be deployed (at UAB) to store the generated LOE Meta information and it will be recorded as part of metadata for each new document once they are processed. Subsequently, the catalog metadata information will be used in the Data acquisition process for new searches to increase speed and reliabilty.

5.5.2. Knowledge Personalization Component.

In order to customize KCE and GRE for Scientific SEARCH to generate the targeted knowledge (in composed format as well as generating recommendation), all users will be encouraged to register and set up a profile about their specialty (e.g. family practice, internal medicine, Ob/Gyn, etc), and their research interests (e.g. diabetes, obesity, etc) through the NSDL user management module before starting a search. Users will also be provided forms to evaluate the information provided, which will be stored with their profile. These grades will be used to assess the operation of the Knowledge Personalization Component and improve its search performance.

IntegraL: Scientific SEARCH project will integrate with other NSDL resources through the IntegraL project [http://is.njit.edu/integral]. In addition to the GRE recommendations, the IntegraL engine automatically will add link anchors to recognized key phrases and other identifiable elements in every page displayed. When the user clicks on a link anchor, IntegraL will generate a pop-up list of links to related resources and services for that particular element both within Scientific SEARCH resources and within other NSDL libraries and resources, customized to the users' current task. Similarly, IntegraL automatically will add links within other participating NSDL resources to relevant Scientific SEARCH resources. IntegraL has the effect of virtually integrating Scientific SEARCH resources and services with those of the rest of the NSDL. The IntegraL project will provide a Masters student who will write the integration wrapper necessary to connect Scientific SEARCH to the NSDL IntegraL infrastructure, under the guidance of the NJIT team.

Customized Recommendation: Scientific SEARCH project will provide recommendations of additional relevant documents based on collaborative filtering (CF), content-based filtering (CB), and knowledge-based filtering (KB) through our NSDL General Recommendation Engine (GRE) which we shall specially customize for Scientific SEARCH. All user search queries will be first processed by the federated search engine to obtain an initial list of search hits. At the same time, the GRE engine will start generating and combining the three kinds of recommended documents. The purpose of CB recommendations is the same as the "similar pages" link in Google – they help users find more relevant documents to a specific document without the need to change the initial query. CB recommendations are obtained from the document-document similarity matrix created during document indexing.

CF generates recommendations based on the browsing patterns of users who have similar interests. This is especially important to medical professionals because medicine is a well defined domain and each specialty has advanced to different complex knowledge. Searching large medical databases with implicit help from similar practitioners will greatly reduce the effort needed. To produce accurate CF recommendations, user profiles and click streams from user browsing activities are needed. Once users are grouped based on their interests and browsing patterns, recommendations are made based on the users' clickstream patterns and ratings (if available). Not-before-seen items for a particular user can be rated and ranked based on the ratings and clickstreams obtained from users in the same group.

KB aims at providing recommendations based on users' knowledge domain areas. It works by mapping users' needs to document metadata and other salient features extracted from documents. Users' needs can be represented as simple keywords about their goals and tasks or documents they found useful, and these can be stored in their profiles. For items to be recommended, information about their features/characteristics is stored, as well as the relationships between these items and features. Both users and documents are modeled using the same knowledge representation, so that they can be matched.

It must be emphasized that the major contribution and difference between results of original GRE and those of Scientific SEARCH customizable recommendation is that recommended documents, whether CF, CB or KB, will all be augmented with level of evidence (LOE) information for more efficient and effective browsing.

Knowledge Composition Engine (KCE): In addition to the GRE/Integral recommendations, The Scientific SEARCH project will provide composed knowledge modules through the development of the KCE. Similar to the GRE, all user search queries will first be processed by the federated search engine to obtain an initial list of search hits. KCE development will focus on knowledge composition. Here, we note that the challenge of integrating information for knowledge composition has many commonalities with integrating software components. In fact, when we consider that the source of information is becoming increasingly software components deployed on networks, be it a local area network or the Internet, information integration is a corollary to integrating components. Information comes in different formats (e.g., text, Adobe-pdf, MS-Word), shapes (e.g. written, graphic, images, videos), and levels (e.g., abstract, detailed, expert). Efforts to integrate related but incongruous information from disparate sources on the Web has eluded researchers and information seekers despite the many advances through NSDL and other initiatives [1]. Similarly, efforts to integrate independently developed software components have not been particularly successful. Relying on established standards, it is possible to mediate mismatches between components through the introduction of adapters, even when different standards are in use. The de-facto component standards in existence today address the wiring level interaction. Yet there is much to gain from standards targeted at higher levels, such as the operational and semantic levels of components. Our approach decomposes the external behavior of components into three independent views based on data, function, and control. Mismatches between components in each of these three views can then be separately reconciled through a mapping process that relies on archetypes of reusable designs. This process forms the basis for implementing smart mediating adapters between components. At the wiring level, components can be made to communicate with each other through the use of bridges and proxies that allow one component to request the invocation of a service from another component and then have that invocation translated into a format that the target component can accept. In this way, the problem defined at the wiring level has been addressed through mediation rather than through standardization. The solution is workable because in the presence of multiple standards, it is possible to convert from one set of standards into another, through the use of adapters [37, 38]. Component integration starts by removing mismatches at the wiring level. But this is only the first step. The focus of our research is to lay the mechanisms for defining the type of mismatches that must be overcome at the next level - the operational level. Mismatches between components at the operational level are often traced to three basic elements: data, function, and control [39]. This is due to the fact that in the design of components, developers use different models for the data imported/exported by the component, the services assumed to exist within the execution environment, and the control mechanism used in invoking services between components. The underlying models used by two components, which are expected to interact but which have been developed under two different models, will invariably have mismatches. By using the principle of separation of concerns, we propose to separate the issues related to each of the three basic elements and then deal with each element independently. It is expected that Meta data standards [40] can be mapped for describing the characteristic of each type of element (data, function, and control) within the component design model. Having these meta-data for each component, it will become possible to mediate the mismatches between components in a way similar to what is possible at the wiring level.

5.5.3. Integration with NSDL core infrastructure. We will incorporate several NSDL core integration features. In Section 5.4.3, we have described a detailed approach for integration Shibboleth as the main authentication and authorization mechanism in Scientific SEARCH. MedFeds uses NSDL search in its federated search mechanism. Moreover, through the use of IntegraL and GRE in the Scientific SEARCH project, people using the NSDL search service will see Scientific SEARCH links and recommendation embedded within the search results. Additionally, we will work to incorporate the KCE in the NSDL search results. We will deploy OAI server to aid in the Harvesting process of the Scientific SEARCH catalog. The Scientific SEARCH results will also utilize NSDL harvested Metadata when applicable.

6. Significance of first phase of Scientific Search project.

As mentioned in Section 2, the primary focus of the Scientific SEARCH project is to aid/automate the steps described in Table 1. The project will use a multi-phased approach to achieve the objective. In the first phase, the focus is on the following:

1. Delivering actionable information" through utilizing level of evidence-based practice guidelines (**certified and graded quality**) (Table 1, [25]), and
2. Enhancing knowledge personalization through integrating Knowledge Composition Engine (KCE) with NSF-sponsored General Recommendation Engine (GRE) and IntegraL integration infrastructure (**need-focused presentation**). [9-19]:

The first phase affects the search steps described in the Table 1 scenario. Although, the project does not include all the databases mentioned in the Scenario it will provide a model for integration of federated databases. The project significantly enhances as well as reduces time to find relevant information by providing personalized actionable information. For example, Dr. X in the Table 1 scenario looking for actionable Grade 1 (well-verified) information targeted to the case pursued can customize the Scientific SEARCH to his specific needs. Hence, Dr. X will be able to spend more time reviewing the materials for content and spend less time reviewing the materials for its validity. In another scenario, Dr. Y who is looking to experimentally further verify a hypothesis can customize for lower grade information (Grade II and under) for his needs. In both cases, Dr. X and Dr. Y will be provided relevant information - personalized to his domain, interests, and needs - reducing search time and improving productivity. Deep evaluation has been designed to assess the impact of the project in various scenarios.

7. Conclusion and future work.

In this paper, we describe the first phase work of the Scientific SEARCH project. The paper describes a three fold effort: First, the project will leverage upon existing NSDL approaches (GRE, IntegraL, MedFeds http://web.njit.edu/~wu/grants.php [9-19]) in our evelopment efforts. In addition, the project will develop KCE to increase efficiency of access to information. Second, the project will develop federated administration of authentication and authorization systems using NSDL recommended Shibboleth [26] to integrate other library communities into the system. Third, the project will develop test-bed services (Web based) to demonstrate the usage of library services and encourage participation within and outside the collaborating institutions. In the future phases of the project, the project will develop infrastructure to address additional steps besides search in the Table 1 scenario. We will incorporate into Scientific SEARCH using the Web Service API [41] [42] provided two functionalities:

1. Ability to search Google and filter its results by concepts (extending MedFedS) and using level of evidence grade criteria (extending LOE filter) and
2. Ability to search and purchase Amazon search books of Amazon.com through the Scientific SEARCH interface.

Furthermore, we expect to integrate more databases into the Scientific Search after demonstrating the success of the first phase of the project.

8. References

[1] "Knowledge Lost in Information," Report of the NSF Workshop on Research Directions for Digital Libraries, NSF Award no. IIS-0331314, Chatham, MA, June 15-17, 2003.

[2] N. L. Bennett, L. L. Casebeer, et al., "Physicians' Internet information-seeking behaviors," *J Contin Educ Health Prof*, vol. 24, 1, pp. 31-8, Winter, 2004.

[3] L. Casebeer, N. Bennett, et al., "Physician Internet medical information seeking and on-line continuing education use patterns," *J Contin Educ Health Prof*, vol. 22, 1, pp. 33-42, Winter, 2002.

[4] C. Ching-chih and K. Kiernan, "Report of the DELOS-NSF working group on digital imagery for significant cultural and historical materials,"DELOS-NSF working group on Significant Cultural and Historical Materials, 2002.

[5] L. Zia, "Growing a National Learning Environments and Resources Network for Science, Mathematics, Engineering, and Technology Education," *D-Lib Magazine*, vol. 7, 3, March, 2001.

[6] B. Muhlhausler, "The "big picture" in obesity research," *Science*, vol. 300, 5622, pp. 1091-2, May 16, 2003.

[7] A. Ertas, T. Maxwell, et al., "Transformation of Higher Education: The Transdisciplinary Approach in Engineering," *IEEE Transactions on Education*, vol. 46, 2, pp. 289-295, May, 2003.

[8] M. M. Tanik and A. Ertas, "Transdisciplinary Engineering Education and Research Model," *Journal of Integrated Design and Process Science*, vol. 4, 4, pp. 1-11, 2000.

[9] Y. B. Wu and C. X., "Extracting Features from Web Search Returned Hits for Hierarchical Classification," presented at Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE'03), Las Vegas, 2003.

[10] Y. B. Wu, Q. Li, et al., "KIP: A Keyphrase Identification Program with Learning Functions," presented at Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, 2004.

[11] Y. B. Wu, S. L., et al., "Finding More Useful Information Faster from Web Search Results," presented at Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03), New Orleans, 2003.

[12] M. a. S. O. K. Bieber, "On Generalizing the Concept of Hypertext," *Management Information Systems Quarterly*, vol. 16, 1, pp. 77-93, 1992.

[13] M. Bieber, "Hypertext and Web Engineering," *ACM Hypertext'98 Proceedings*, pp. 277-278, 1998.

[14] J. Bieber and M. Bieber, "Relationship Analysis: A Technique to Improve the Systems Analysis Process," *submitted to the Journal of the AIS*, 2005.

[15] I. Im and A. Hars, "Finding Information Just for You - Knowledge Reuse Using Collaborative Filtering Systems," presented at Proceedings of the ICIS Conference, New Orleans, 2001.

[16] J. Yoo, J. Catanio, et al., "Relationship Analysis in Requirements Engineering," *Requirements Engineering*, vol. 9, pp. 238-247, 2004.

[17] J. Yoo and M. Bieber, "Finding Linking Opportunities through Relationship-based Analysis," presented at Hypertext 2000 Proceedings, San Antonio, 2000.

[18] Z. Li and I. Im, "Recommender Systems: A Framework and Research Issues," presented at Americas Conference on Information Systems (AMCIS), Dallas, TX, 2002.

[19] Q. Li, Y. B. Wu, et al., "Incorporating Document Keyphrases in Search Results," presented at Proceedings of the Americas Conference on Information Systems (AMCIS), New York, 2004.

[20] J. Gemmill, S. Chatterjee, et al., "ViDe.Net Middleware for Scalable Video Services for Research and Higher Education," presented at ACM Southeastern Conference, Savannah, GA, 2003.

[21] J. Gemmill and J. Lynn, *Directory Services Middleware for Multimedia Conferencing, Using ITU-T Recommendation H.350 and IETF Informational RFC 3944*: LuLu Press, 2005.

[22] J. Gemmill, A. Srinivasan, et al., "Middleware for Scalable Real-time Multimedia Communications Cyberinfrastructure," *Journal of Internet Technology*, vol. 5, 4, pp. 405-420, 2004.

[23] R. Puljala, R. Sadasivam, et al., "Middleware: Single Sign On Authentication and Authorization for Groups," presented at ACM Southeastern Conference, Savannah, GA, 2003.

[24] R. S. Sadasivam, M. M. Tanik, et al., "Cyberinfrastructure Development - A Component Based Approach with Software Agents," presented at IDPT 2003, Austin, Texas, December 3-5, 2003.

[25] I. f. C. S. I. (ICSI), "Prevention and management of obesity (mature adolescents and adults),"Institute for Clinical Systems Improvement (ICSI), 2004.

[26] M. Erdos and S. Cantor, "Shibboleth Architecture Draft V05."

[27] K. Channabasavaiah, K. Holley, et al., "Migrating to a Service-Oriented Architecture, Part 1," in *developerWorks > Web services*, ftp://www6.software.ibm.com/software/developer/library/ws-migratesoa.pdf, 2003.

[28] "Web Services," http://www.ibm.com/us/, 2003.

[29] D. Geer, "Taking Steps to Secure Web Services," *IEEE Computer*, vol. 36, 10, pp. 14 - 16, October, 2003.

[30] "SOAP and WSDL," http://www.w3.org/TR/, 2003.

[31] S. Pirozzo, C. Summerbell, et al., "Advice on low-fat diets for obesity," *Cochrane Database Syst Rev*, 2, pp. CD003640, 2002.

[32] H. Moore, C. Summerbell, et al., "Dietary advice for treatment of type 2 diabetes mellitus in adults," *Cochrane Database Syst Rev*, 2, pp. CD004097, 2004.

[33] K. Shaw, C. Del Mar, et al., "Exercise for obesity (Protocol for a Cochrane Review)," *The Cochrane Library*, 3, 2002.

[34] Drupal.org, "Drupal Content Management System," vol. 2005, http://www.drupal.org, 2005.

[35] D. Michelinakis, "Open Source Content Management Systems: An Argumentative Approach,"The University of Warwick: Warwick Manufacturing Group, A report submitted for the award of MSc Electronic Business Management, August, 2004.

[36] R. S. Bot, Y.-F. Brook, et al., "A Hybrid Classifier Approach for Web Retrieved Documents Classification," presented at Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, 2004.

[37] L. K. Jololian, F. J. Kurfess, et al., "Data, Function, and Control as Elements of Component Integration," presented at IDPT 2003, Austin, Texas, 2003.

[38] L. K. Jololian, "A Framework for a Meta-Semantic Language for Smart Component Adapter," *Journal of Systems Integration*, vol. 10, 3, pp. 269-297, 2001.

[39] R. Deline, "A Catalog for Resolving Packaging Mismatch," presented at Fifth Symposium on Software Reusability, May 21-23, 1999.

[40] S. E. Dennis, S. Candler, et al., "An Indexing Standard for Sharing Health Education Multimedia Resources: The Health Education Assets Library (HEAL) Metadata Schema," presented at 37th Hawaii International Conference on System Sciences (HICSS-37), Big Island, Hawaii, January 5-8., 2004.

[41] "Amazon Web Services," http://www.amazon.com/gp/browse.html/002-1932355-1742457?node=3435361, 2004.

[42] "Google Web Services," vol. April 2005, http://www.google.com/apis/, 2005.