SBIR PHASE II
FINAL REPORT

PROJECT TITLE:

A Model Management System for Numerical Simulations of Subsurface Processes

Our research team has successfully completed the objectives of the SBIR Phase II project, "A Model Management System for Numerical Simulations of Subsurface Processes".

To facilitate our reporting of our results, the Phase II technical objectives (bold font) are enumerated and below each objective (task) we describe advances made during the performance period.

1. Finalize the design of a framework for a Model Management System (MMS) that contains sufficient model metadata to support a wide range of queries and data mining exercises.

The metadata in place at the end of Phase I provided us the basic capability to successfully demonstrate that our Model Management System (MMS) can provide users with knowledge about the similarities and differences in model capabilities. In our Phase II proposal, we acknowledged that the Phase I metadata would not be sufficient to support a large, diverse set of subsurface numerical models. In the Phase II work we have to expanded our collection of metadata tags and reorganize the grouping and placement of the metadata on the MMS web site.

We have examined glossaries, metadata and ontologies employed by several organizations and agencies including

Semantic Web for Earth and Environmental Terminology (SWEET)
ASTM International
USGS, Groundwater Modeling Software
EPA, Center for Subsurface Modeling Support
EPA, Software for Environmental Awareness
International Groundwater Modeling Center, School of Mines, Colorado

We are continued to augment our initial set of metadata and reorganize the groupings of the metadata in response to comments and suggestions of beta-version testers of our MMS. We produced metadata that is consistent with existing data models, such as the links cited above.

Our work on Objective 9, Latent Semantic Analysis (LSA), has produced a semi-automated method for identifying metadata. The first step in performing LSA on a set of documents is to parse the document set and produce a master list of terms to be used in a term-document matrix. In the case that the documents are text containing model descriptions (e.g. abstracts, user manuals, journal articles, etc.), the term-document matrix has proved useful in assisting us in identifying additional metadata. Simply put, terms that appear with high frequency in several documents are good candidates for metadata and LSA exposes these terms.

2. Finalize the design of a relational database that will capture the information in the model metadata template.

In the performance period, we have continued to revise and improve the structure of the database that underlies the MMS. . As stated in the Phase II proposal Performance Schedule contained in Appendix C, improving the design of the database will continue over the entire course of the project and beyond.

3. Populate the database with a number of subsurface process simulation models from a variety of application areas with emphasis on models developed with DOE funding.

We have identified more than one hundred simulation models for inclusion in our MMS and have used several textual descriptions of these models in the semi-automatic method for identifying metadata tags described under Task 1.

4. Finalize the Phase I development of a web site through which users can access the models database and perform a wide variety of queries and data mining activities. Included in this objective are the construction of Help files and contextual help in the form of text in response to a mouse-hover event.

During the project we have made several modifications that improve the user-friendliness of the MMS. One improvement of note is that we have developed new methods for adding and removing MMS metadata tags.

5. Engage the subsurface sciences modeling community in the beta testing of the web site and revise the metadata template, database design and web interface design as required.

During this project, the PI attended several conferences and has described our MMS to specialists in subsurface science simulations. Contacts made at these conferences,

coupled with e-mail and phone calls, resulted in a core set of experts willing to help with the beta testing of the MMS.

6. Develop a computational subsurface sciences wiki and forum to enable the subsurface sciences modeling community to share results of models, data sets and to discuss model capabilities and possible improvements.

We made excellent progress in developing a working version of a computational subsurface sciences web site containing a wiki that supports community editing and a forum for community comments and discussions related to subsurface science problems. The wiki and forum can be accessed at

http://subsurface.vistacomputational.com/wiki/index.php/Main_Page

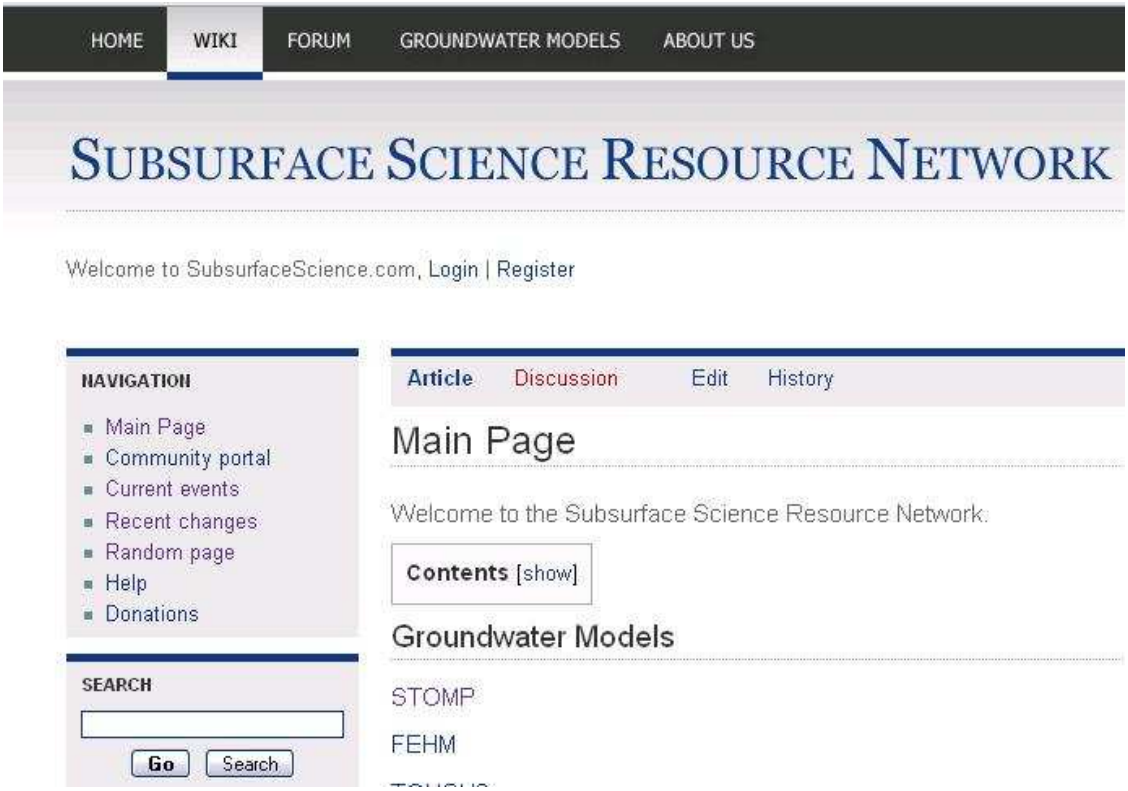

Figure 1: A fragment of the Subsurface Science Resource Network Wiki

To contribute to the wiki or to interact in a forum requires a user to be registered at our site. We have written scripts that support user registration and a password protected login for registered users. Also included is an automated password recovery mechanism to deal with forgotten passwords. Throughout the project, our development team and several beta users tested the wiki and forum for resilience and ease of use. All of the development of the wiki and forum has been done during the reporting period. The scripts that support the wiki and the forum appear to be stable at this point.

7. Add RSS and Atom feeds to the web site to provide an efficient means of keeping the user community informed about recent updates to the web site.

We have added RSS and Atom feeds to the wiki.

8. Replace some of the most heavily used PHP scripts with Java technologies (EJBs, JSPs, and Servlets).

During the project we replaced several placeholder PHP scripts with more robust Java based technologies on an "as-needed" basis to derive better performance of the MMS, resulting in improved user interfaces. In the computationally intensive Latent Semantic Analysis, all development has been carried out using Java.

9. Add a Latent Semantic Analysis feature to the model management system to facilitate enhanced key word search and to provide another model similarity metric to complement the current cosine-based similarity metric

During the project, we developed a powerful tool for performing Latent Semantic Analysis (LSA) on collections of text documents. Our working name for the LSA tool is "LSAer". After launching LSAer, a user is presented with the simple welcome screen shown in Figure 2.

Figure 2: Initial LSAer interface

This interface provides (1) a text field in which users enter search phrases; (2) a button that initiates a search of the document collection for the documents that best match the key works and (3) a link to a form that enables users to set preferences and view results – see Figure 3.

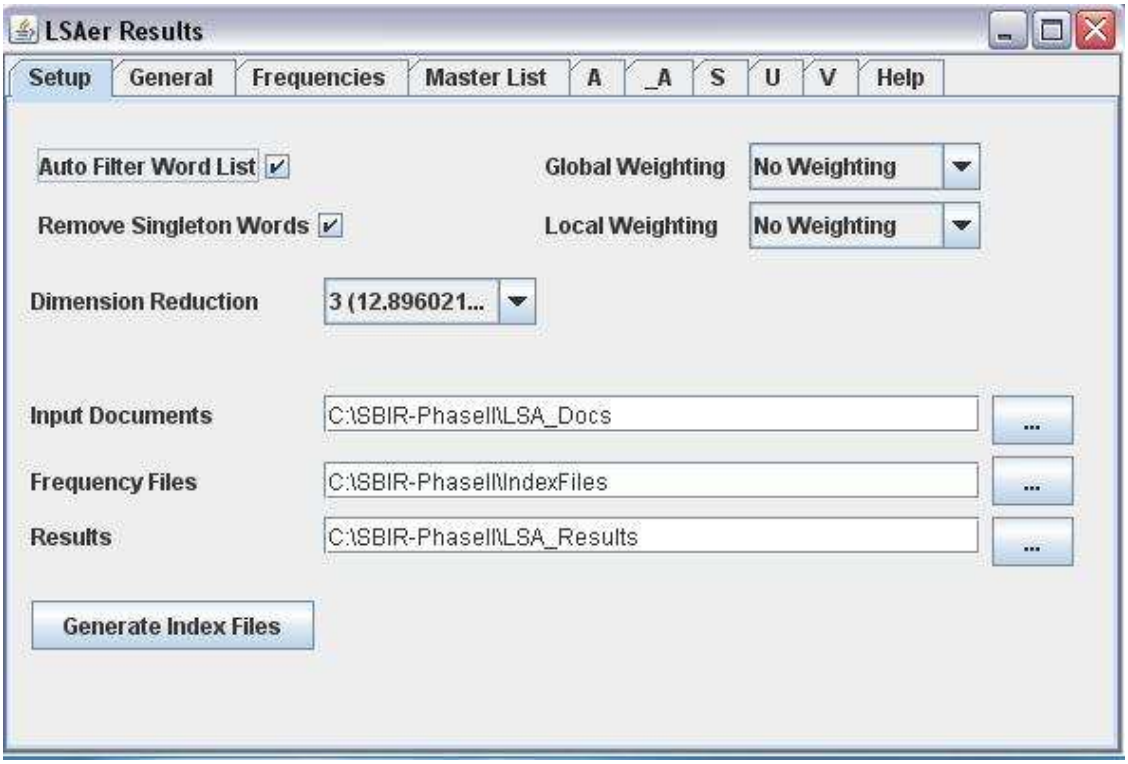

Figure 3: The preferences and LSAer results form with the Setup tab active

From the Setup form a user can check Auto Filter Word List to remove “stop words”, which are words that are typically ignored in a browser search action (e.g. the, this it so, etc.) Checking Remove Singleton Words prevents words that appear in only one document from being included in the Master List. The Dimension Reduction pull-down permits the user to zero out selected lower order singular values to view relationships between the documents in a concept space of any given dimension – full to one dimensional. The largest singular value to be replaced by zero appears next to the dimension choice.

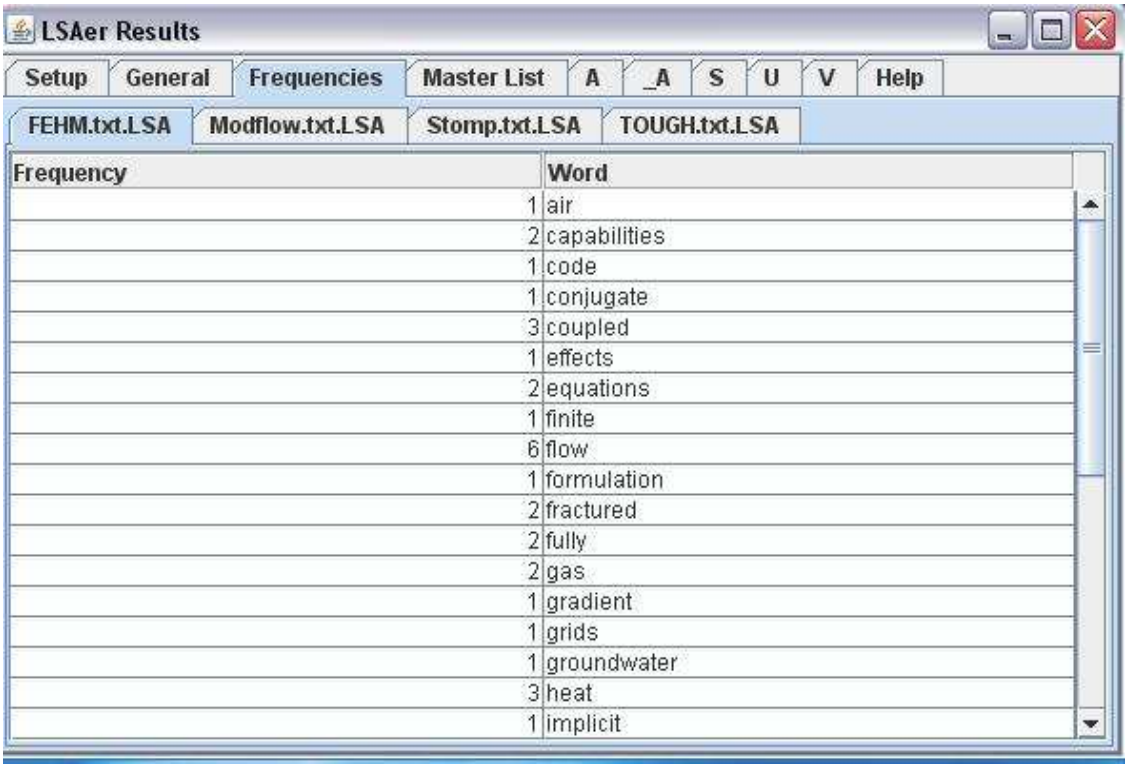

Figure 4: Terms associated with document FEHM.txt after filtering

Figure 4 displays the terms associated with the document FEHM.txt after filtering has been completed and also displays the frequency of occurrence of each term. The master list of terms for the document set is comprised of the union of the terms associated with each document in the solution. See Figure 5.

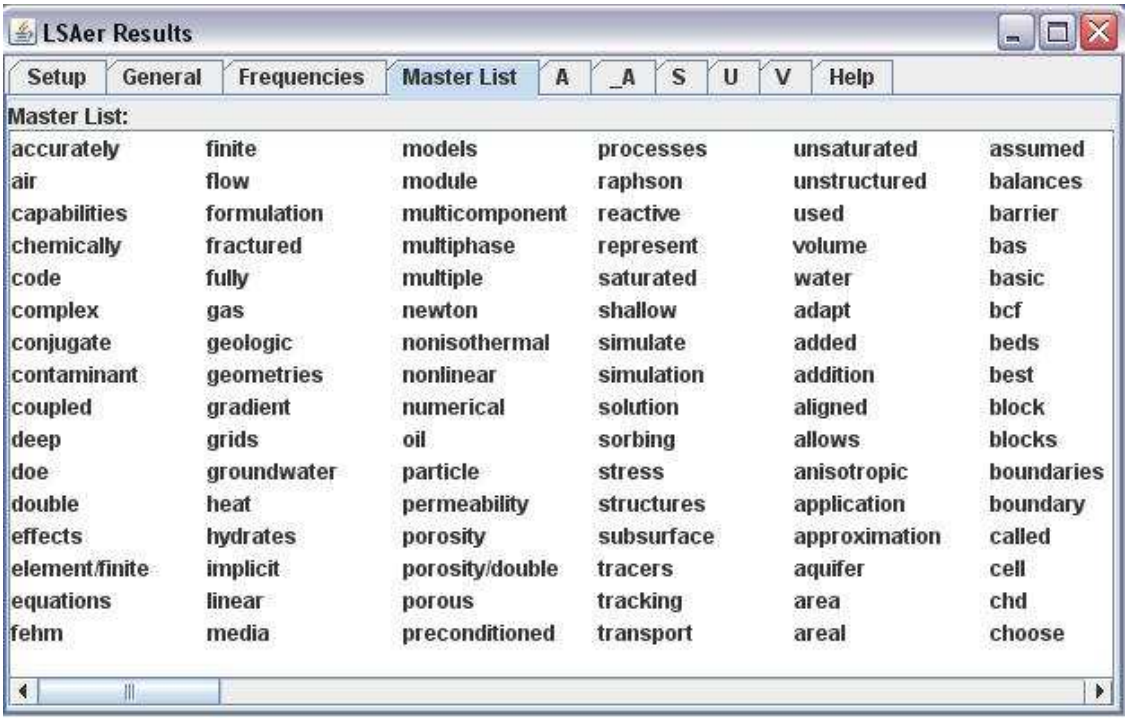

Figure 5: The master list from which the term-document matrix is formed

The master list shown in Figure 5 contains abbreviations bas, bcf, chd found in the document Modflow.txt. We developed a mechanism that allows a user to filter misspelled words and abbreviations from the master list.

The tab labeled "**A"** allows a user to view the term-document matrix. The tabs S, U and V provide views of the matrices in the Singular Value Decomposition of the term-document matrix, $A = USV^T$. The tab "**_A** "provides the approximate term-document matrix (_A) when dimension reduction is applied. The columns of _A are used to calculate correlations between the various model documents and to assess their respective similarities in term-document space and concept spaces of reduced dimensions. See Appendix A.

In either term document space or one of the reduced dimension concept spaces, the Search option of the LSAer main form allows users to enter a query (key words) and the program will return relevance ranks for each document in the collection. Figure 6 shows the rankings in three-dimensional concept space that are returned by the query "finite element". The model description document FEHM.txt (Finite Element Heat and Mass Transport) receives the highest rank with TOUGH scoring second highest. These are the only finite element models in this collection. Modflow and STOMP are both integrated finite difference models.

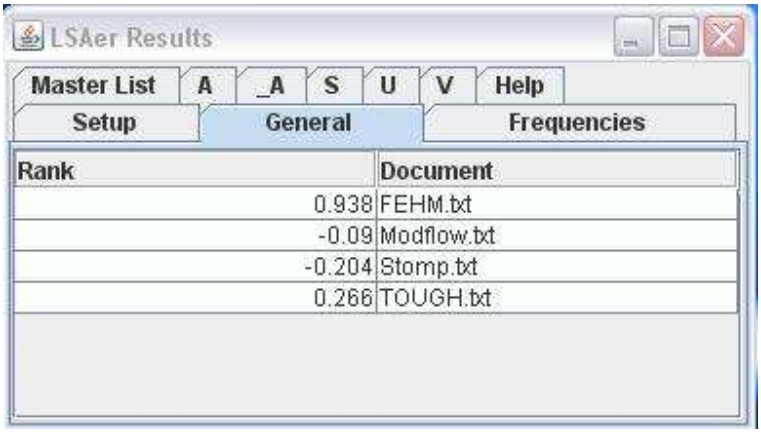

Figure 6: Rank values resulting from a search on "finite element"

In the development and testing of LSAer, we have performed several case studies using different sets of documents. One of these case studies used a collection of documents containing the research interests of five mathematics faculty members and can be found in Appendix B. That case study and other results in the literature indicate that there is great potential for LSA as a tool for matching referees to journal submissions, reviewers to proposals for funding and other similar activities.

10. Develop a marketable suite of metadata tools that can be used in a variety of Web 2.0 applications and will enable workers in industry, government and education to create access and manipulation high quality metadata with minimal impact on their normal course of work.

Automatic and semi-automatic generators of metadata are areas of potential commercial applications. We have found that the first steps of LSA – (1) parsing documents for key words; (2) elimination of stop words; (3) stemming; and (4) application of local and global weighting – have provided useful metadata in the two test areas cited under Task 9. We envision widespread opportunities to develop novel ways to harvest metadata from a variety of document sets, thus minimizing the need to create-by-hand word lists from which metadata can be extracted. In subsequent projects the PI has been able to use the tools developed in this project to obtain contracts in the private sector to provide services in the area of both surface and subsurface hydrology.

APPENDIX A
LATENT SEMANTIC ANALYSIS
OF FOUR SUBSURFACE MODEL DOCUMENTS

For this case study, we used text describing four popular models for simulating subsurface flow phenomena. The models used in the study are FEHM (LANL), MODFLOW (USGS), STOMP (PNNL) and TOUGH (LBNL). Without additional modules, MODFLOW simulates only groundwater flow. The other three DOE laboratory developed models all simulate groundwater flow, and have additional capacities for simulating transport of chemicals and heat.

The lengths of the documents describing the models varied considerable, with a small term list for FEHM of 87 to a large term list for MODFLOW of 380 terms. After merging the four term lists, we have a master list of 680 terms from the 4 documents. In the singular value decomposition (SVD) of the 680x4 term-document matrix, the singular values are

s1 = 152.26 s2 = 40.89 s3 = 24.39 s4 = 13.58

Below we display the (Pearson) correlations for this set of documents that describe their respective capabilities. Using the full term-document matrix to calculate the various correlations, we see that all four models have significant positive correlations – they all model the flow of groundwater. By a small margin the most closely correlated models are FEHM and TOUGH.

Replacing s4 by zero in the SVD representation of the term document matrix results in the best 3-Dimensional (Frobenius norm) approximation to the term-document matrix in what is called “concept space” in the LSA community. This reduction of dimension indicates that FEHM and TOUGH are strongly correlated according the text documents used to represent their capabilities.

A further reduction of dimension, replacing both s4 and s3 by zero results in the final set of correlations displayed. At this stage we see that all three DOE lab models are very highly correlated, with the USGS being a relative outlier in this small set of models.

	FEHM	MODFLOW	STOMP	TOUGH
FEHM	1.00	0.38	0.48	0.70
MODFLOW	0.38	1.00	0.63	0.59
STOMP	0.48	0.63	1.00	0.68
TOUGH	0.70	0.59	0.68	1.00

Correlations based on Full Term-Doc Matrix

	FEHM	MODFLOW	STOMP	TOUGH
FEHM	1.00	0.47	0.59	0.99
MODFLOW	0.47	1.00	0.63	0.59
STOMP	0.59	0.63	1.00	0.69
TOUGH	0.99	0.59	0.69	1.00

Correlations based on 3-Dim SVD Approx to Term-Doc Matrix

	FEHM	MODFLOW	STOMP	TOUGH
FEHM	1.00	0.54	0.98	0.99
MODFLOW	0.54	1.00	0.69	0.64
STOMP	0.98	0.69	1.00	.998
TOUGH	0.99	0.64	.998	1.00

Correlations based on 2-Dim SVD Approx to Term-Doc Matrix

APPENDIX B
LATENT SEMANTIC ANALYSIS
OF RESEARCH INTERESTS OF
FIVE CSU MATH FACULTY MEMBERS

For this case study, we harvested some "Research Interests" text from some faculty web pages found within the Colorado State University Mathematics Department web site. The individuals used in the study are Gerhard Dangelmayr, Paul DuChateau, Rick Miranda, Jennifer Mueller and Juliana Oprea. Four of these are applied mathematicians and Rick Miranda is an algebraic geometer. Gerhard and Juliana work closely together and have some joint publications. Paul and Jennifer do not work closely and have not published together, but they are both interested in groundwater flow and inverse problems.

After filtering out "stop words" (non-keywords e.g. the, a, this …), we have a master list of 167 terms from the 5 documents. In the SVD of the 167x5 term-document matrix the singular values are

s1 = 23.881 s2 = 8.645 s3 = 7.544 s4 = 3.935 s5 = 3.683

Below we display the (Pearson) correlations for this group's research interests, as stated on their respective web pages. The first set of correlations, using the full term-document matrix, do not reveal strong correlations, other than each individual's research interests are perfectly correlated with him/herself.

Replacing s5 by zero results in the best 4-Dimensional (Frobenius norm) approximation to the term-document matrix in what is called "concept space" in the LSA community. This dimension reduction step results in a marked increase in the correlations of the research interests of Paul and Jennifer – now .83 up from their previous .06. Gerhard's and Juliana's research interests remain to appear uncorrelated. Rick remains to be negatively correlated to the other four.

A further reduction of dimension, replacing both s5 and s4 by zero results in the final set of correlations displayed. The correlation between Paul and Jennifer increases to .93. With this 3-D "concept space" approximation, now co-workers and co-authors Gerhard and Juliana correlate at the .99 level! Rick remains to be negatively correlated to the other four.

	Gerhard	Paul	Rick	Jennifer	Juliana
Gerhard	1.00	-0.17	-0.17	-0.06	0.04
Paul	-0.17	1.00	-0.12	0.06	-0.13
Rick	-0.17	-0.12	1.00	-0.14	-0.14

Jennifer	-0.06	0.06	-0.14	1.00	-0.08
Juliana	0.04	-0.13	-0.14	-0.08	1.00

Correlations based on Full Term-Doc Matrix

	Gerhard	Paul	Rick	Jennifer	Juliana
Gerhard	1.00	-0.18	-0.16	0.21	0.05
Paul	-0.18	1.00	-0.12	0.83	-0.12
Rick	-0.16	-0.12	1.00	-0.20	-0.13
Jennifer	0.21	0.83	-0.20	1.00	-0.50
Juliana	0.05	-0.12	-0.13	-0.50	1.00

Correlations based on 4-Dim SVD Approx to Term-Doc Matrix

	Gerhard	Paul	Rick	Jennifer	Juliana
Gerhard	1.00	-0.18	-0.17	0.18	0.99
Paul	-0.18	1.00	-0.12	0.93	-0.24
Rick	-0.17	-0.12	1.00	-0.29	-0.06
Jennifer	0.18	0.93	-0.29	1.00	0.10
Juliana	0.99	-0.24	-0.06	0.10	1.00

Correlations based on 3-Dim SVD Approx to Term-Doc Matrix

The documents from which the terms were extracted follow. These documents were copied from the five researchers' CSU Math web sites. Note the large disparities in size and style.

Begin Gerhard.txt

* Geometrical theory of dynamical systems:
    - o Chaotic Dynamics
    - o Normal Forms and Unfoldings of Vector Fields and Maps
    - o Singularity Theory and Imperfect Bifurcations
* Dynamical Systems with Symmetries
* Algorithms for Center Manifold Reductions and Normal Form Transformations
* Perturbation Techniques:
    - o Averaging and Melnikov-Methods
    - o Multiple Time Scales
    - o Singular Perturbations
* Systems of Nonlinear Oscillators

Instabilities and Pattern Formation:

* Formation of Spatio-Temporal Patterns in Systems of PDE's:
    - o Analysis of Instabilities via Center Manifold and Normal Form Theory
    - o Spontaneous and Forced Symmetry Breaking
    - o Reduction of PDE's to Systems of ODE's
    - o Envelope- and Phase Diffusion-Equations
* Application to:
    - o Fluid Mechanics
    - o Reaction-Diffusion Systems
    - o Semiconductors and Superconductors
    - o Nonlinear Optics (optical bistability and laser)

Pattern Analysis and Neural Networks:

* Remodeling and Prediction of Dynamical Systems from Data via
    - o Topology Preserving Neural Network Algorithms
    - o Extraction of Invariant Manifolds
    - o Markov Analysis
* Dynamics and Modeling of Continuous Neural Networks
* Systems of Neural Oscillators
* Neural Learning Rules for Storing Patterns and Pattern Cycles

Methods of Mathematical Physics:

* Linear and Nonlinear Boundary- and Eigenvalue-Problems
* Variational Calculus and Optimization
* Asymptotic Expansions for Linear and Nonlinear Waves
* Geometrical Theory of Diffraction and "Singularity Optics"
* Semiclassical Methods of Quantum Mechanics
* Asymptotic Approach to Inverse Scattering Problems

End Gerhard.txt

Begin Paul.txt

Professor DuChateau's research interests lie in the area of partial differential equations, particularly in inverse problems arising in modeling flow through porous media. His research in these areas has been supported in the past by National Science Foundation Engineering and by the Office of Naval Research. This research has resulted in the publication of more than 50 papers on inverse
problems and partial differential equations.

End Paul.txt

Begin Rick.txt

Prof. Miranda's main field of interest is Algebraic Geometry, which is, broadly speaking, the study of curves, surfaces, etc. which are defined by the vanishing of one or several polynomials. He has written articles and/or directed research in the following areas.

* Elliptic Surfaces
* Geometric Invariant Theory
* Classification and Degenerations of Surfaces
* Finite Coverings of Algebraic Varieties
* Integral Quadratic Forms
* Resolutions of Singularities
* Toric Geometry
* Gaussian Maps for Curves
* Graph Curves
* Fano Threefolds
* Quantum Cohomology
* Linear Systems of Plane Curves

End Rick.txt

Begin Jennifer.txt

* Numerical algorithms for inverse problems
* Reconstruction algorithms for electrical impedance tomography (EIT)
* Contaminant transport in groundwater

End Jennifer.txt

Begin Juliana.txt

Hydrodynamic and Hydromagnetic Stability and Bifurcation: the dynamo problem, the electroconvection of nematic liquid crystals; Dynamical Systems, Pattern Formation, Mathematical Modelling;

End Juliana.txt

APPENDIX C
PHASE II PERFORMANCE SCHEDULE

Project Objectives:

1. Continue to develop a MMS framework and work with the subsurface sciences community to improve the ontology that underlies an optimal set of metadata tags.
2. Finalize the design of a relational database that will capture the information in the model metadata template.
3. Populate the database with a number of subsurface process simulation models from a variety of application areas with emphasis on models developed with DOE funding and to the extent possible provide links to sample input files.
4. Finalize the Phase I development of a web site through which users can access the models database and perform a wide variety of queries and data mining activities. This task includes the development of Help files and contextual tool-tip help in response to selected mouse-hover events.
5. Engage the subsurface sciences modeling community in the beta testing of the web site and revise the metadata template, database design and web portal design as required.
6. Develop and test a computational subsurface sciences wiki to enable the subsurface sciences modeling community to share results of models and data sets and to discuss model capabilities and possible model improvements.
7. Add RSS and Atom feeds to the web site to provide an efficient means of keeping the user community informed about recent updates to the web site.
8. Rewrite some of the most heavily used PHP scripts in Java to enable faster runtime and user interface responsiveness since the servlets and beans are compiled and not interpreted. It also allow for greater manageability of the code base and faster development.
9. Add a Latent Semantic Analysis feature to the model management system to facilitate “concept space” key word search and to provide another model similarity metric to complement the current cosine-based similarity metric
10. Develop a marketable suite of metadata tools that can be used in a variety of Web 2.0 applications.

Year 1 Schedule

Month

Task	1	2	3	4	5	6	7	8	9	10	11	12
1	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X
5												
6	X	X	X	X	X	X	X	X	X	X	X	X
7	X	X	X	X	X	X						
8	X	X	X	X	X	X	X	X	X	X	X	X
9	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X

Year 2 Schedule

Month

Task	13	14	15	16	17	18	19	20	21	22	23	24
1	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X
5	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X	X	X	X	X	X
7												
8						X	X	X	X	X	X	X
9	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X