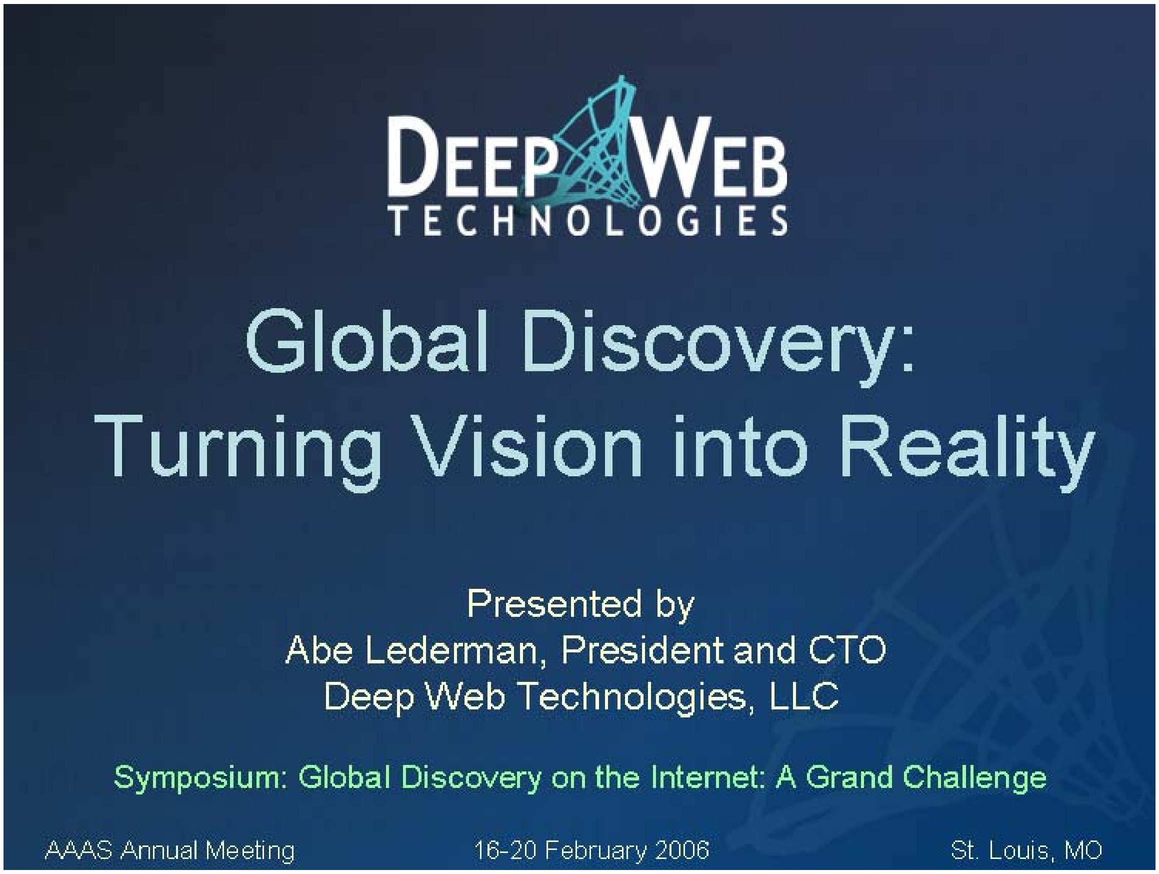
AAAS Annual Meeting, St. Louis, MO – Speaker Presentation Write-up



DEEP WEB TECHNOLOGIES

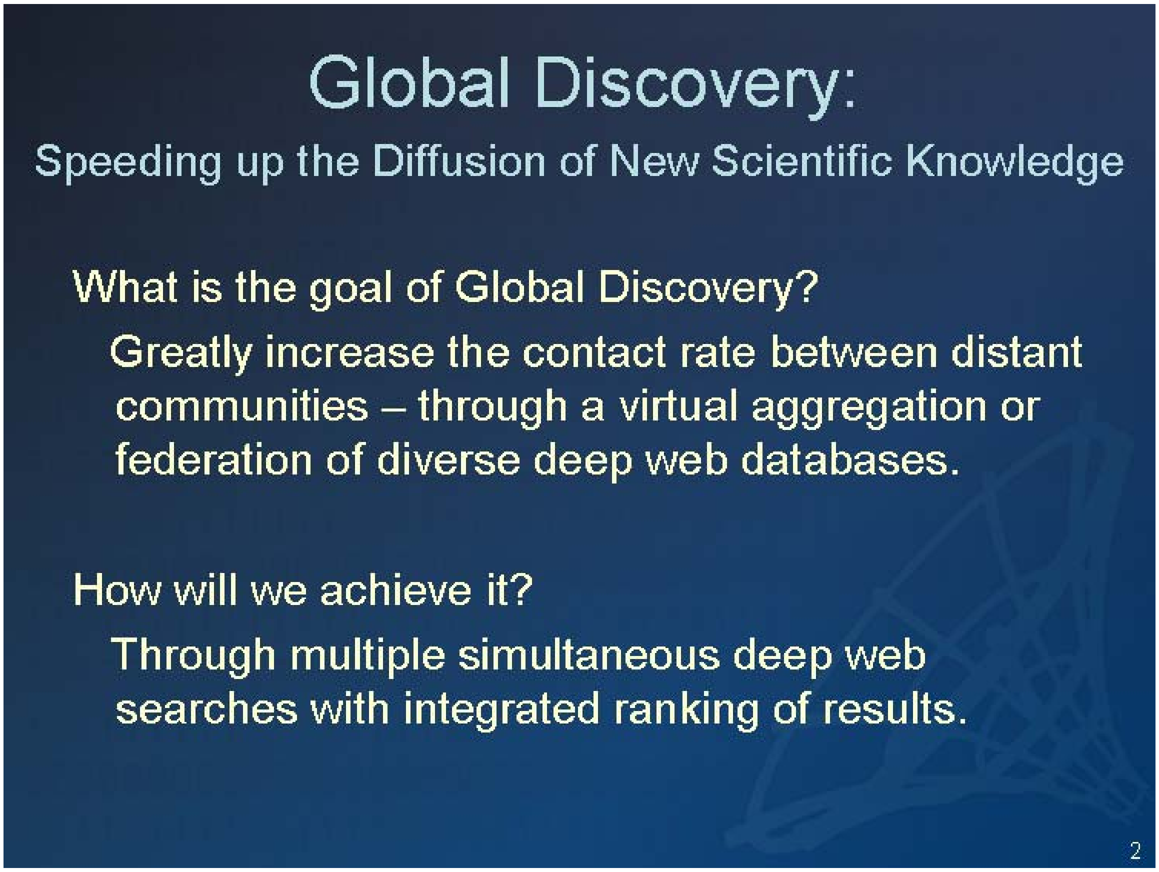
Global Discovery:
Turning Vision into Reality

Presented by
Abe Lederman, President and CTO
Deep Web Technologies, LLC

Symposium: Global Discovery on the Internet: A Grand Challenge

AAAS Annual Meeting 16-20 February 2006 St. Louis, MO

Presented as part of the Symposium: Global Discovery on the Internet: A Grand Challenge



Global Discovery:

Speeding up the Diffusion of New Scientific Knowledge

What is the goal of Global Discovery?

Greatly increase the contact rate between distant communities – through a virtual aggregation or federation of diverse deep web databases.

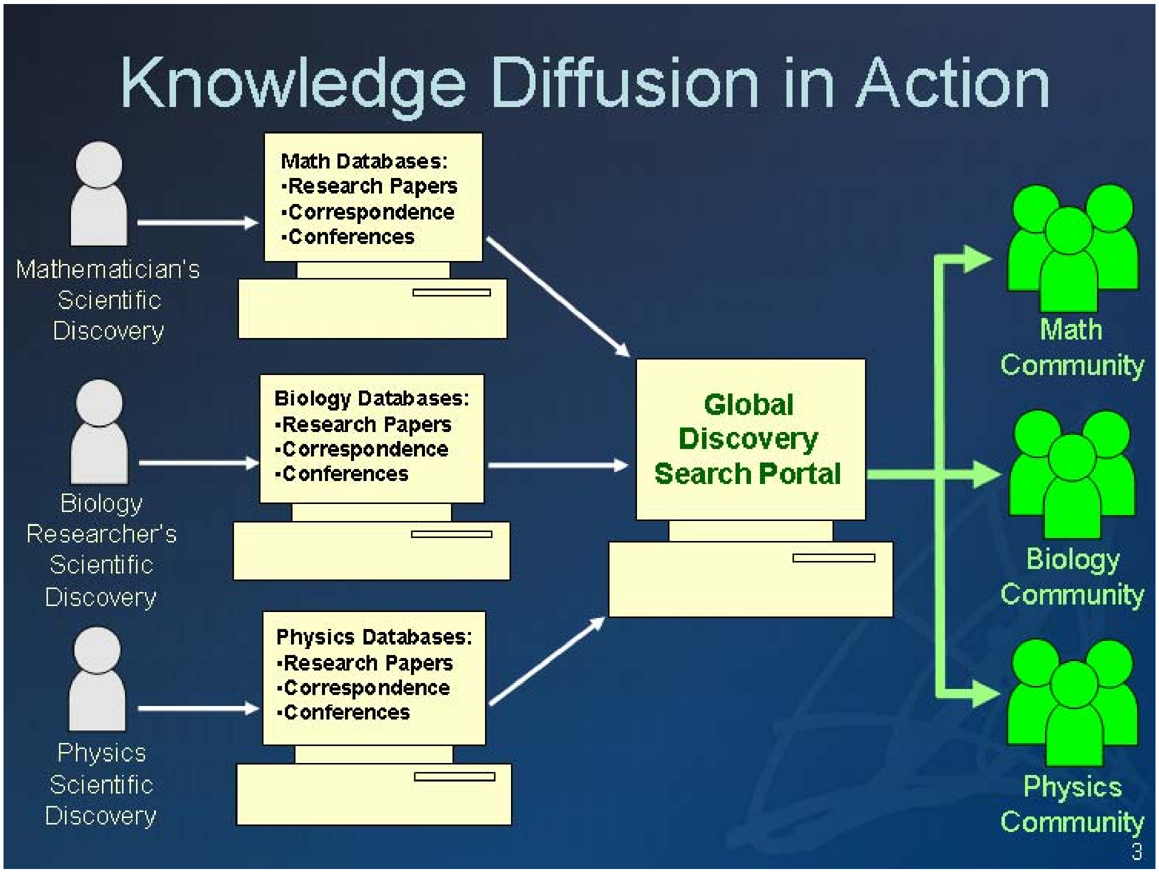
How will we achieve it?

Through multiple simultaneous deep web searches with integrated ranking of results.

Slide 2: Global Discovery: Speeding up the Diffusion of New Scientific Knowledge

What is the goal?
Greatly increase the contact rate between distant communities – through a virtual aggregation or federation of diverse deep web databases.

How will we achieve it?
Through multiple simultaneous deep web searches with integrated ranking of results.



Knowledge Diffusion in Action

Mathematician's Scientific Discovery

Math Databases:
- Research Papers
- Correspondence
- Conferences

Biology Researcher's Scientific Discovery

Biology Databases:
- Research Papers
- Correspondence
- Conferences

Physics Scientific Discovery

Physics Databases:
- Research Papers
- Correspondence
- Conferences

Global Discovery Search Portal

Math Community

Biology Community

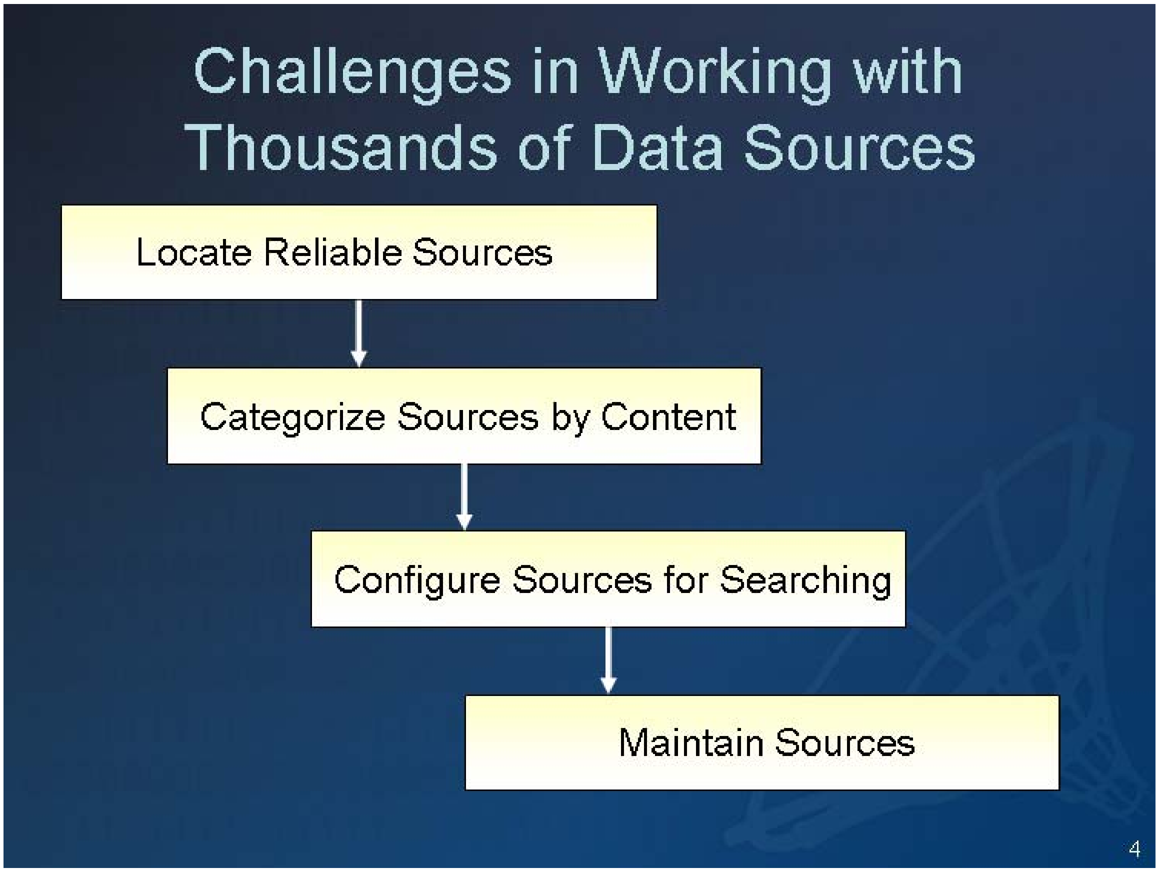
Physics Community

3

Slide 3: Knowledge Diffusion in Action

Knowledge diffusion is the essential challenge in addressing the vision of Global Discovery. On the right side of this slide we are representing different communities that would gain access to global information through the Global Discovery Search Portal.

The Portal aggregates scientific discoveries from a myriad of data sources covering information from a vast number of scientific disciplines. A researcher publishes the results of his work that become cataloged in databases specialized for his discipline. These databases are the sources of the information that becomes diffused.



Challenges in Working with Thousands of Data Sources

Locate Reliable Sources

Categorize Sources by Content

Configure Sources for Searching

Maintain Sources

4

Slide 4: Challenges in Working with Thousands of Data Sources

It is unknown at this time how many data sources would be searched in a comprehensive Global Discovery portal. This number is certainly in the thousands and very likely to be in the tens of thousands. Data sources include published journal literature, conference proceedings, report collections at university libraries and research laboratories, and many more.

Locate Reliable Sources

A major challenge of Global Discovery is to locate data sources worthy of inclusion. We are currently exploring the development of semi-automated tools that can mine the Web in search of candidate sources, and once potential sources are identified these sources can be presented for a researcher to determine whether the source contains useful content and should become a part of the Global Discovery portal.

One Web mining strategy is to start with web pages that are “hubs” (i.e. contain lots of useful links) within certain research communities and follow links from these “hubs” to other useful pages. For each web page that is examined, a determination is made whether the page is a part of the research community being explored and if it is links from that page are followed. For each web page that is examined a determination is made about whether the page contains a search form that should be further examined to determine whether it is a gateway to useful research results.

Categorize Sources by Content

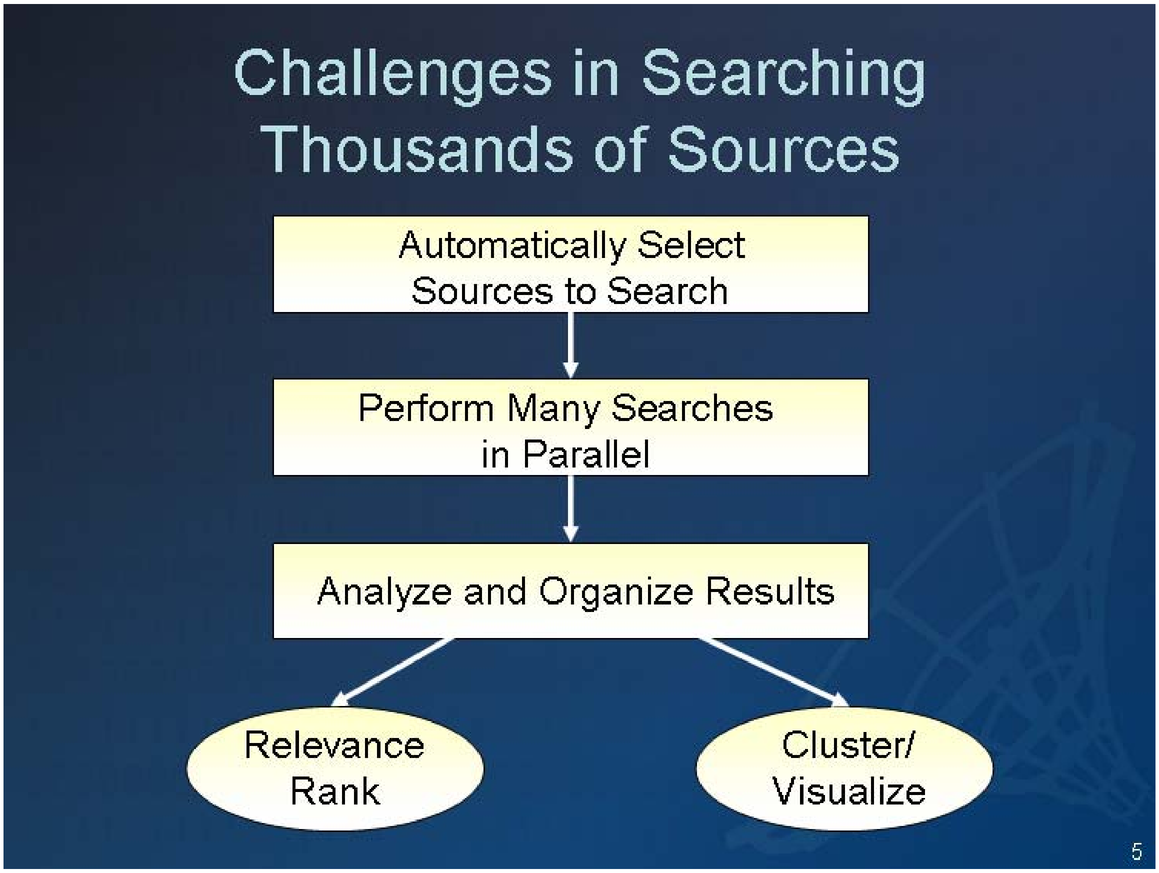
During the search for and evaluation of reliable sources to include in the Global Discovery portal, each data source can be categorized as to the scope of the source and type and size of the collection. Categorizing sources in this way will significantly improve the efficiency of searching the Global Discovery portal and lead to higher quality results.

Configure Sources for Searching

Each source that is searched by a federated search engine needs to be configured so that appropriate search parameters are passed to the data source, and so that results returned by data sources can be properly parsed and result data such as URLs, titles, authors and dates can be extracted. Configuring an access gateway between a federated search engine and each source is currently labor intensive. Developing semi-automated tools that can reduce (by 80% or more) the effort involved in configuring sources will be extremely important if thousands or tens of thousands of sources will need to be configured.

Maintain Sources

Already configured sources are subject to changes by the source administrators when they make changes (improvements) to meet their own internal needs. Monitoring the accessibility of sources requires periodic test searches to verify that a source is still searchable. Currently we have developed tools which on a daily basis perform searches of data sources and report to an application engineer any problems that are encountered. Development of new tools and improvements to existing tools will be required to significantly reduce the effort that will be involved in the maintenance of thousands or tens of thousands of sources.



Challenges in Searching Thousands of Sources

Automatically Select Sources to Search

Perform Many Searches in Parallel

Analyze and Organize Results

Relevance Rank

Cluster/ Visualize

Slide 5: Challenges in Searching Thousands of Sources

<u>Automatically Select Sources to search</u>

What is needed is a Source Selection Optimizer that orchestrates the selection of the right sources to search, with a feedback loop that can increase the breadth of sources that are searched if necessary to get relevant results. The Source Selection Optimizer is described in a later slide.

<u>Perform many searches in parallel</u>

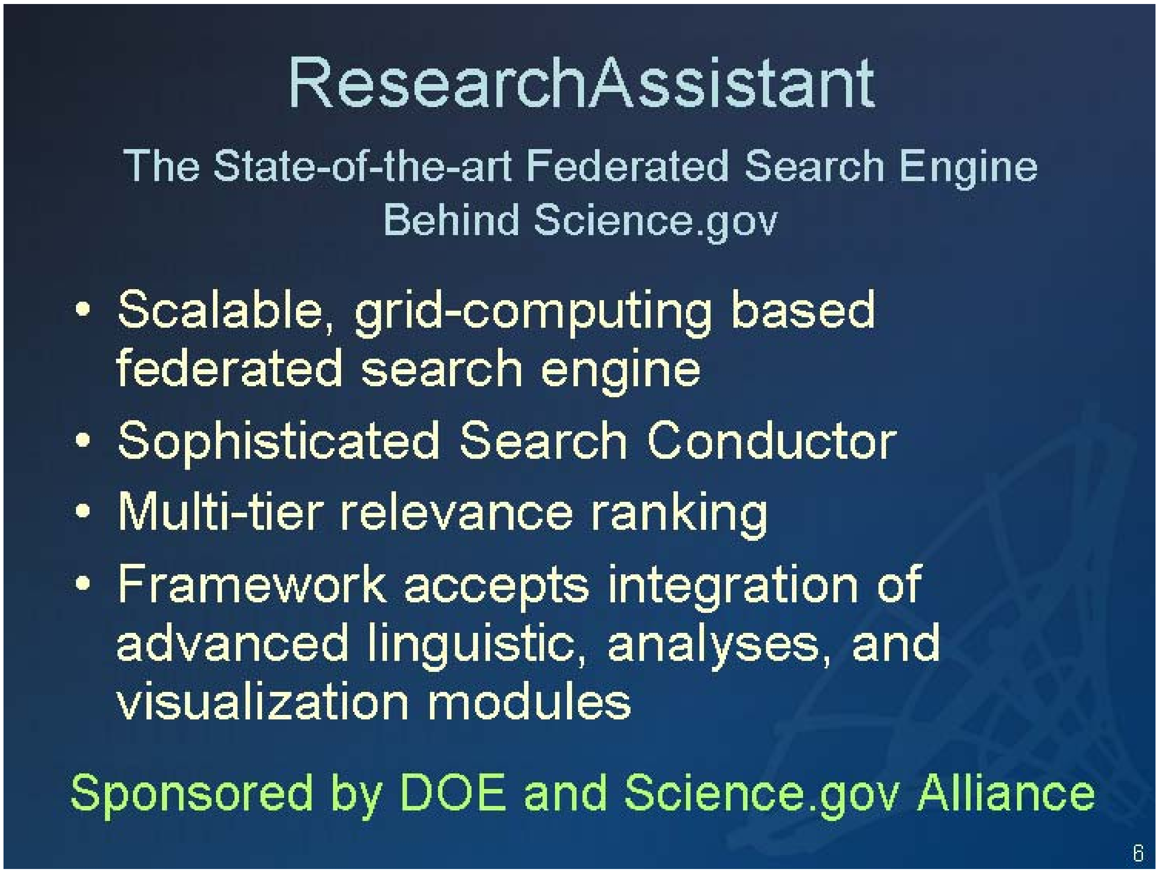
A key requirement and a key challenge of making Global Discovery into a reality is the ability to efficiently scale to search hundreds or possibly thousands of data sources simultaneously in response to a search request from a researcher.

This is an area on which DWT has focused significant R&D efforts, and has developed the ResearchAssistant, a next generation federated search engine that meets this challenge. Capabilities of the ResearchAssistant are discussed in a later slide.

<u>Analyze and Organize Results</u>

With only dozens of sources being searched, relevant results can be lost in a deluge of clutter. When Global Discovery allows a user to search hundreds or thousands of sources simultaneously it will be critical to present the user the most relevant results first. Only those systems with sophisticated relevance ranking will deliver the best quality information. ResearchAssistant's multi-tier Relevance Ranking algorithms have been designed to accomplish this goal even as the number of data sources searched significantly increases.

Cluster and/or visualization techniques in combination with Relevance Ranking can provide a user with additional ways of organizing search results so that the information most useful to a user is easy to find.



ResearchAssistant

The State-of-the-art Federated Search Engine Behind Science.gov

- Scalable, grid-computing based federated search engine
- Sophisticated Search Conductor
- Multi-tier relevance ranking
- Framework accepts integration of advanced linguistic, analyses, and visualization modules

Sponsored by DOE and Science.gov Alliance

6

Slide 6: Current State-of-the-Art: Science.gov is powered by ResearchAssistant

Science.gov, the portal developed by the Science.gov Alliance, provides “science attentive” citizens with access to the output of most of the R&D that is produced by the Federal government.

Scalable, grid-computing based federated search engine

The ResearchAssistant architecture is built to support applications that require extensive scaling. Grid-computing expands federated search capabilities with fault-tolerant load-sharing hardware nodes at geographically dispersed locations.

Sophisticated Search Conductor

The ResearchAssistant Search Conductor orchestrates the actions of our federated search engine. In real-time the Search Conductor determines 1) the sources to search, 2) which sources are returning “good” results, triggering additional search results requested from these sources and 3) which sources are returning “poor” results and should be abandoned. We have a flow-chart that describes at a very high-level the Search Conductor operation on a later slide.

Multi-tier relevance ranking

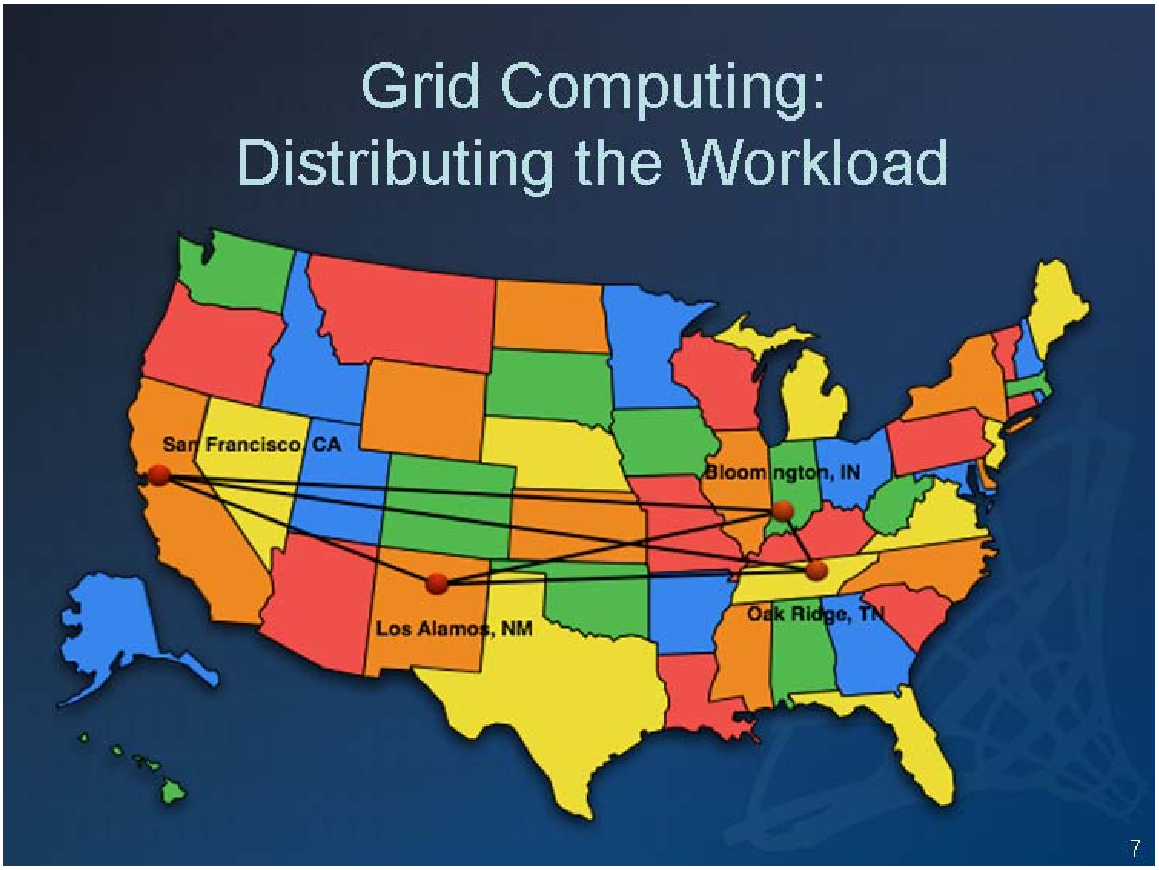
The ResearchAssistant implements a sophisticated multi-tier ranking approach that optimizes the load on sources searched and ensures that the most relevant results are found and returned to the user. ResearchAssistant relevance ranking is described in more detail in a later slide.

Framework for integration of advanced linguistic, analyses, and visualization modules

The ResearchAssistant open design architecture supports (through a Web Services interface) integration into the workflow of custom modules that can apply special filters to results returned, extract names of people, places, and dates, perform language translation and analyses and more.

Sponsored by DOE and the Science.gov Alliance

We are honored to have DOE and the Science.gov Alliance as our sponsors.



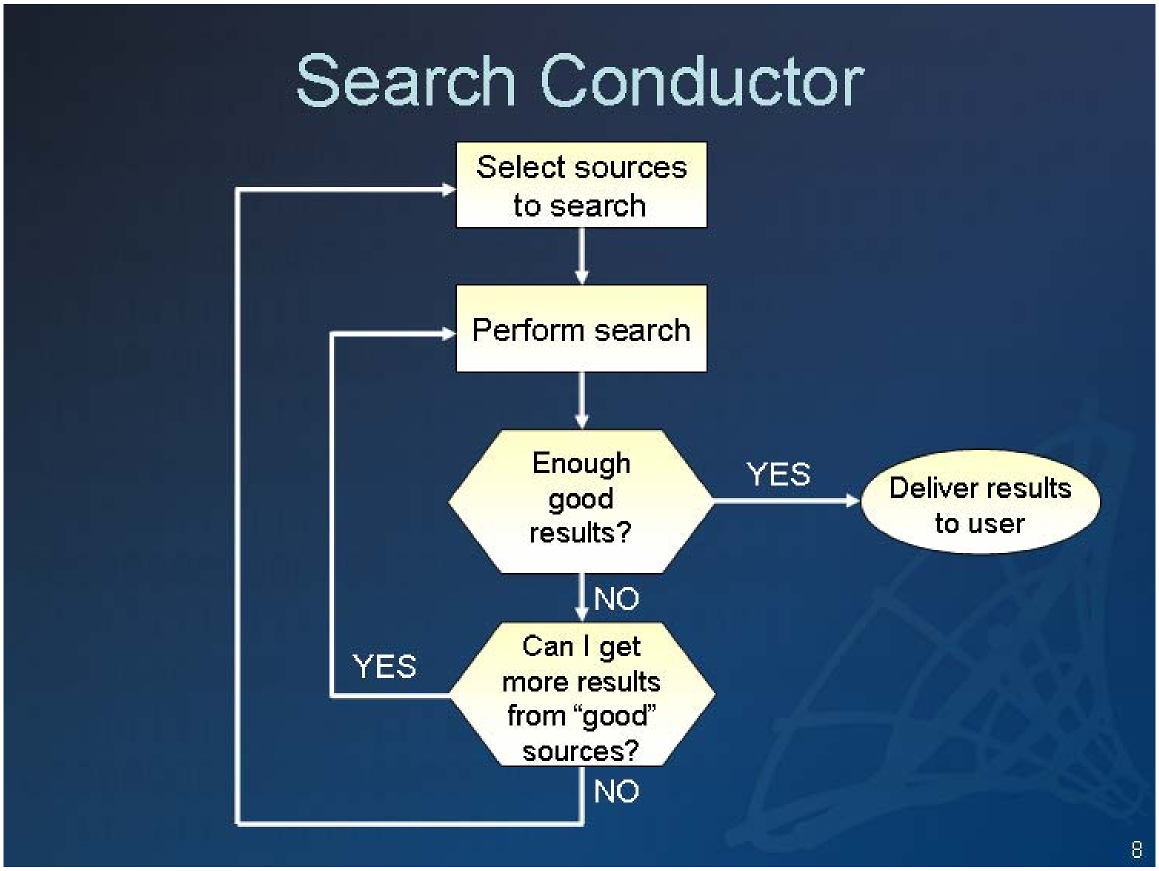
Grid Computing: Distributing the Workload

Slide 7: Grid Computing: Distributing the Workload

The ResearchAssistant is implemented as a number of Java Web Services whose execution can be distributed among available computing resources.

The ResearchAssistant can optimize federated search performance (primarily CPU and network bandwidth) and enable scaling to search of hundreds or thousands of sources through the use of grid-computing. The ResearchAssistant can be configured so that the search and ranking of heavily used data sources is performed by grid-nodes that are in proximity to the data source.

The nodes shown on the map illustrate one ResearchAssistant implementation with multiple grid-nodes deployed at the locations shown.



Search Conductor

Select sources to search

Perform search

Enough good results?

YES

Deliver results to user

NO

Can I get more results from “good” sources?

YES

NO

8

Slide 8: Search Conductor

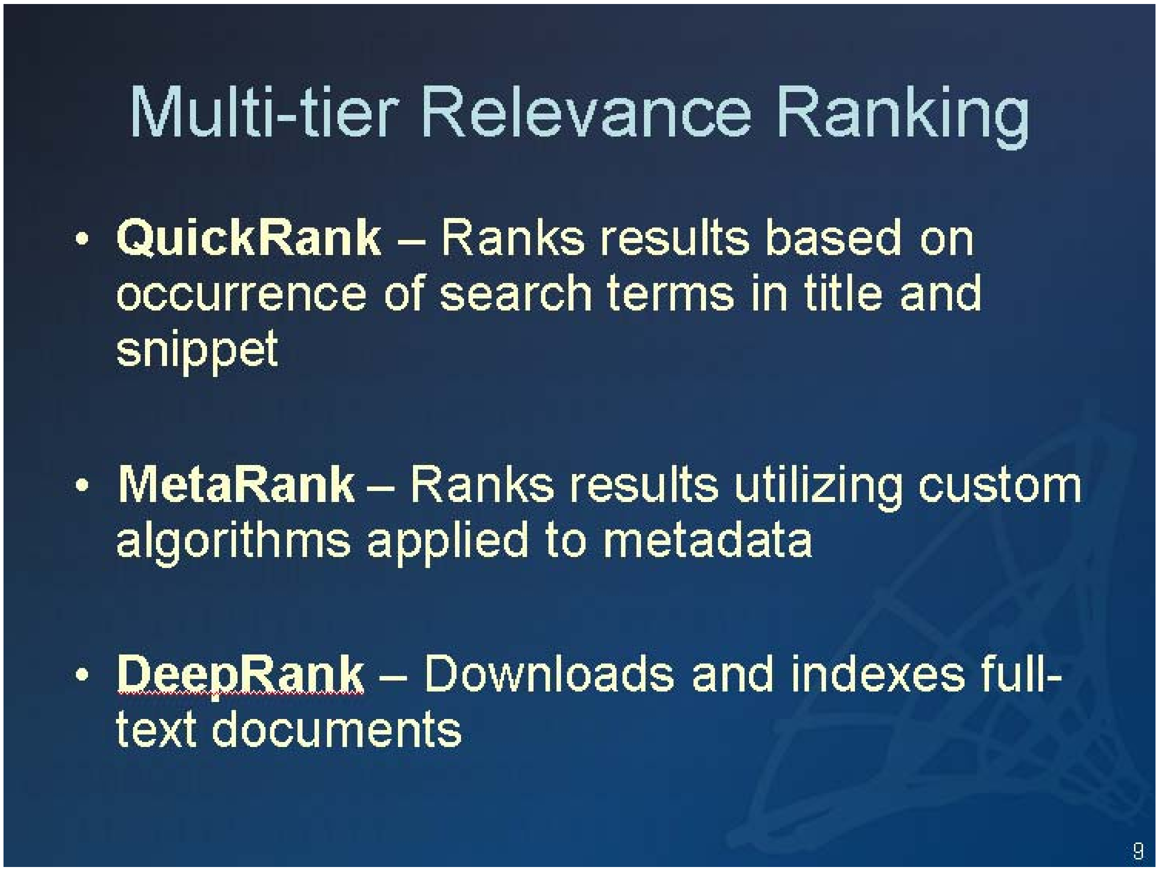
The ResearchAssistant utilizes a sophisticated Search Conductor designed to minimize system load on data sources searched while ensuring that no highly relevant results are missed.

The Search Conductor accomplishes these goals by monitoring in real-time how a search is progressing and evaluating the quality of search results returned by individual sources. Additional search results are requested from data sources that return a number of highly ranked results and no additional results are requested from data sources that initially return poor results.

The Search Conductor works in conjunction with the Source Selection Optimizer, described in more detail in a later slide, to determine the best sources to initially submit a search to. If the searches of initial data sources do not return sufficient high quality results then the Search Selection optimizer is requested to “recommend” additional data sources to search.

The Search Conductor also determines the appropriate level of use of the various ranking algorithms at its disposal (discussed in more detail in the next slide).

In the future an advanced user will be able to interact with the Search Conductor and provide feedback as to the usefulness of the search results returned and whether more comprehensive searching should be performed.



Multi-tier Relevance Ranking

- **QuickRank** – Ranks results based on occurrence of search terms in title and snippet
- **MetaRank** – Ranks results utilizing custom algorithms applied to metadata
- **DeepRank** – Downloads and indexes full-text documents

9

Slide 9: Multi-tier Relevance Ranking

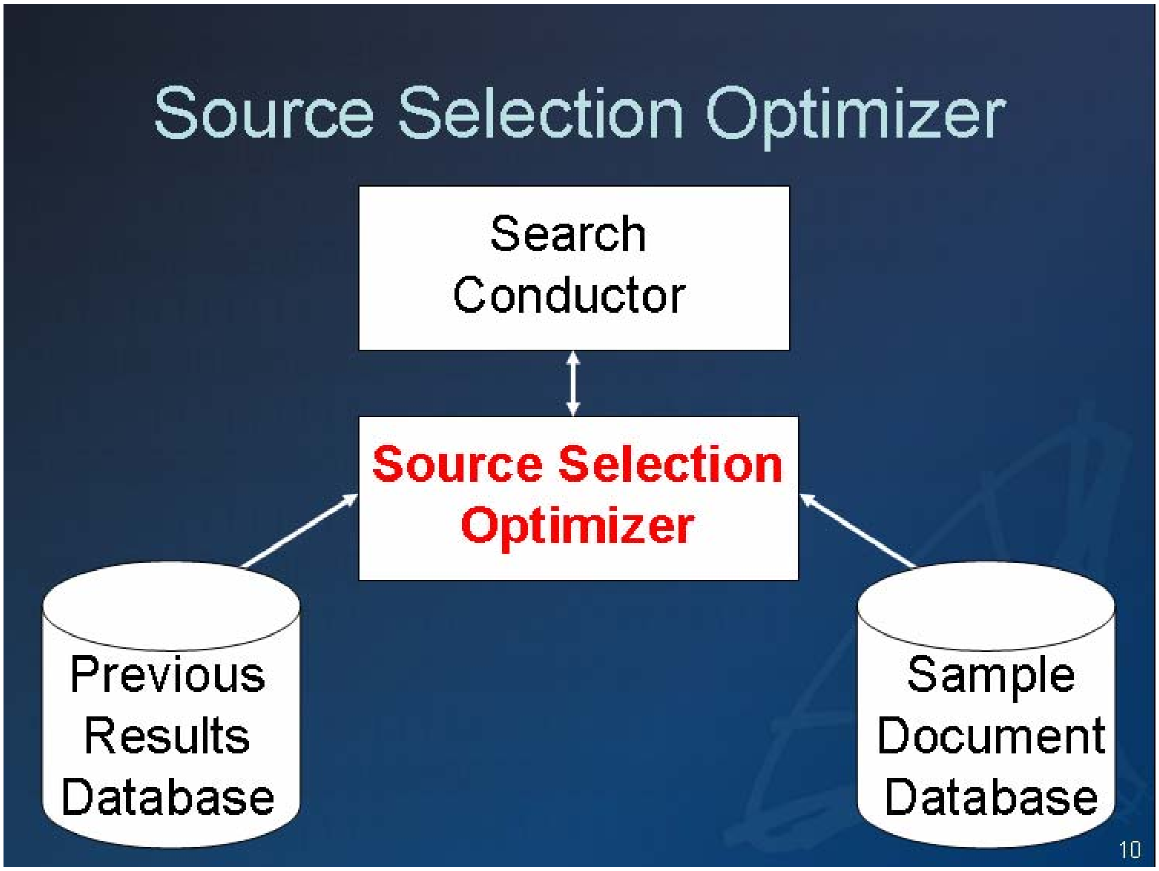
The ResearchAssistant and more specifically the Search Conductor has at its disposal three sophisticated ranking algorithms developed through extensive research.

Incorporated into these ranking algorithms is the assigning of a higher weight to results where there is: 1) a high density of search terms; 2) search terms occur earlier or later within a section of the document (e.g. title or abstract); and 3) proximity of search terms to each other.

QuickRank – Ranks results based on occurrence of search terms in title, snippet and other fields returned on a result list. QuickRank requires little overhead and is applied to all search results.

MetaRank – Ranks results utilizing custom algorithms applied to metadata. In order to perform MetaRank, the Search Conductor must request download of individual meta-data records for each result. There are performance costs and concern with not putting undue load on data sources being queried. MetaRank is performed only when QuickRank did not identify a sufficient number of highly relevant documents or a user has requested a comprehensive search.

DeepRank – Downloads and indexes full-text documents, including documents in PDF, Microsoft Word and other formats. DeepRank is typically only performed when QuickRank and MetaRank have failed to return relevant documents.



Source Selection Optimizer

Search Conductor

Source Selection Optimizer

Previous Results Database

Sample Document Database

10

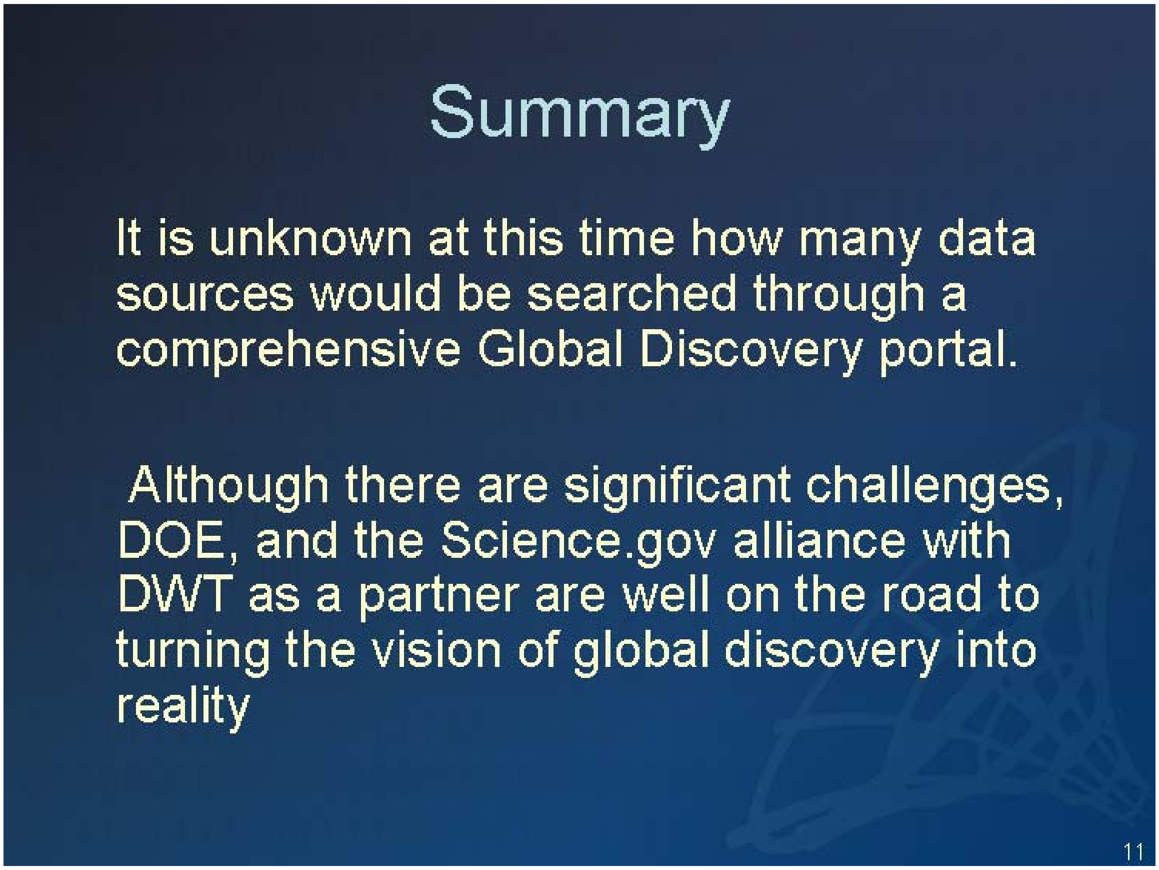
Slide 10: Source Selection Optimizer

The Source Selection Optimizer works in conjunction with the Search Conductor. The Source Selection Optimizer recommends to the Search Conductor which sources are the best to search in order to find and return to the user the most relevant results and avoid sources that are not likely to return relevant results.

The Source Selection Optimizer uses two approaches in determining which sources to search.

One approach is to initially perform a search of the Sample Document Database. This database is created by performing a number of representative queries against each data source that a ResearchAssistant application is able to search. The documents returned by each of these representative queries is indexed and stored in the Sample Document Database. When the Source Selection Optimizer is requested to provide a list of data sources to search, it first performs the requested search against the Sample Document Database and, those data sources identified by this search as containing the most number of results are the data sources recommended to the Search Conductor.

A second approach that the Source Selection Optimizer takes to determine the best data sources to search is to examine the Previous Results Database and determine which sources previously returned the most relevant results to the current query or a similar query (a thesaurus is used). The Previous Results Database maintains a comprehensive history of past search results that help optimize future searches.



Summary

It is unknown at this time how many data sources would be searched through a comprehensive Global Discovery portal.

Although there are significant challenges, DOE, and the Science.gov alliance with DWT as a partner are well on the road to turning the vision of global discovery into reality

11

Slide 11: Summary

It is unknown at this time how many data sources would be searched through a comprehensive Global Discovery portal.

Although there are significant challenges, DOE, and the Science.gov alliance with DWT as a partner are well on the road to turning the vision of global discovery into reality.

Turning Vision into Reality

Abe Lederman
122 Longview Drive
Los Alamos, NM 87544
abe@deepwebtech.com
www.deepwebtech.com

DEEP WEB TECHNOLOGIES

12

Slide 12: Turning Vision into Reality

Thank you for your attention. I will be available at the break for any questions.