JPL Media Search® Project

Multimedia Search Tool for Accessing Engineering 'Best Practice' Information

David Oberhettinger
Office of the Chief Engineer
NASA/Caltech Jet Propulsion Laboratory

David Tate
Managing Director
Owl Insight LLC

February 26-27, 2008

Background

Have you considered videotaping critical engineering meetings and important presentations, but knew you would never sit through 100 hours of videos to retrieve a key fact?

- Multimedia can help NASA effectively capture “best practice” information and forestall the loss of engineering know-how
- But the inability to index and search audio/video files, unlike text files, has limited their usefulness as technical references (or as engineering decision documentation)
- JPL and Owl Insight LLC have developed a system that can index/search collections and find/play the specific video snippet with the search term
- Our pilot-scale, browser-based, search engine can perform context-sensitive searches of audio-video files and graphics metadata, as well as all the text-based file types commonly used in an office environment
- Media Search® works by transcribing the audio track to text, performing semantic processing by applying a natural language model to determine the context in which engineering terms are being used, and indexing the information for access via a dedicated browser search page. *Semantic processing* permits a search on the word “gauss” to produce results containing terms like “magnetic contamination” that are not closely related

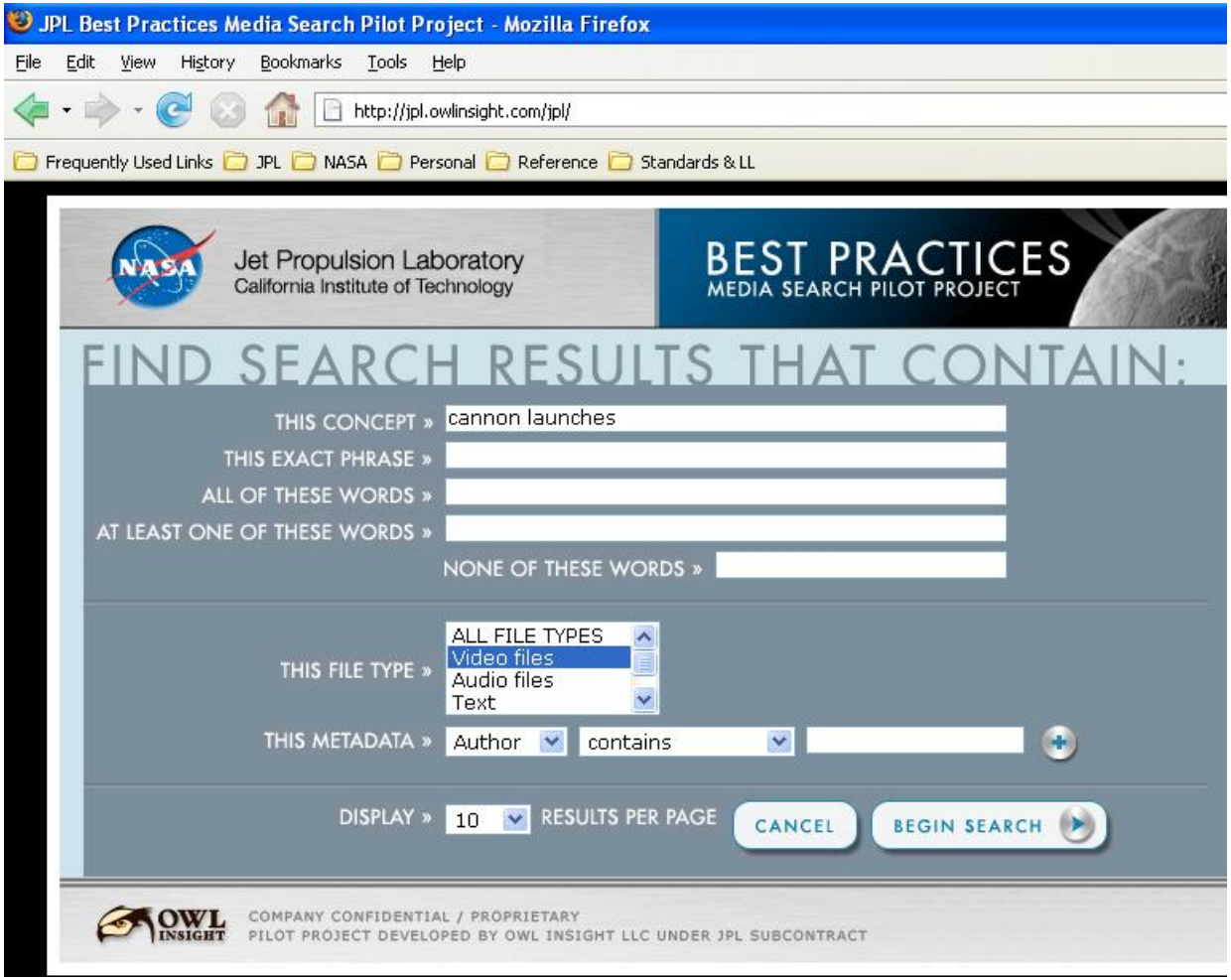
Semantic Processing

JPL
Jet Propulsion Laboratory
California Institute of Technology

- **Boolean search**: conventional search method that requires you to know exactly the right search terms in order to get successful search results.
- **Mathematical approaches to search**: represent documents as arrays (vectors) of numbers and use neural networks or statistical analyses to match documents with queries. These systems can recognize documents that more or less match the query based on the statistical regularities common to natural language. (That is, based on an assumption that words in a paragraph are not wholly independent of one another.)
- **OrcaTec® language modeling approach**: The OrcaTec® Information Retrieval Toolkit uses a Bayesian probabilistic model to estimate the likelihood that a document is relevant to a specific query. The software calculates the probability of one word being in a paragraph given the other words in a paragraph. For example, if the word "accelerate" is in a document, it is likely that such words as "force," "engine," or "velocity" will be in the document. Conversely, if the words “force," "engine," and " velocity" are in a document, then it is likely that this document is about "accelerate" even if that word does not happen to be in it. This enables the system to accurately predict which documents that lack the specific search terms are likely to be relevant to the search. The system progressively becomes more knowledgeable— or “learns”— as it is trained on additional documents.

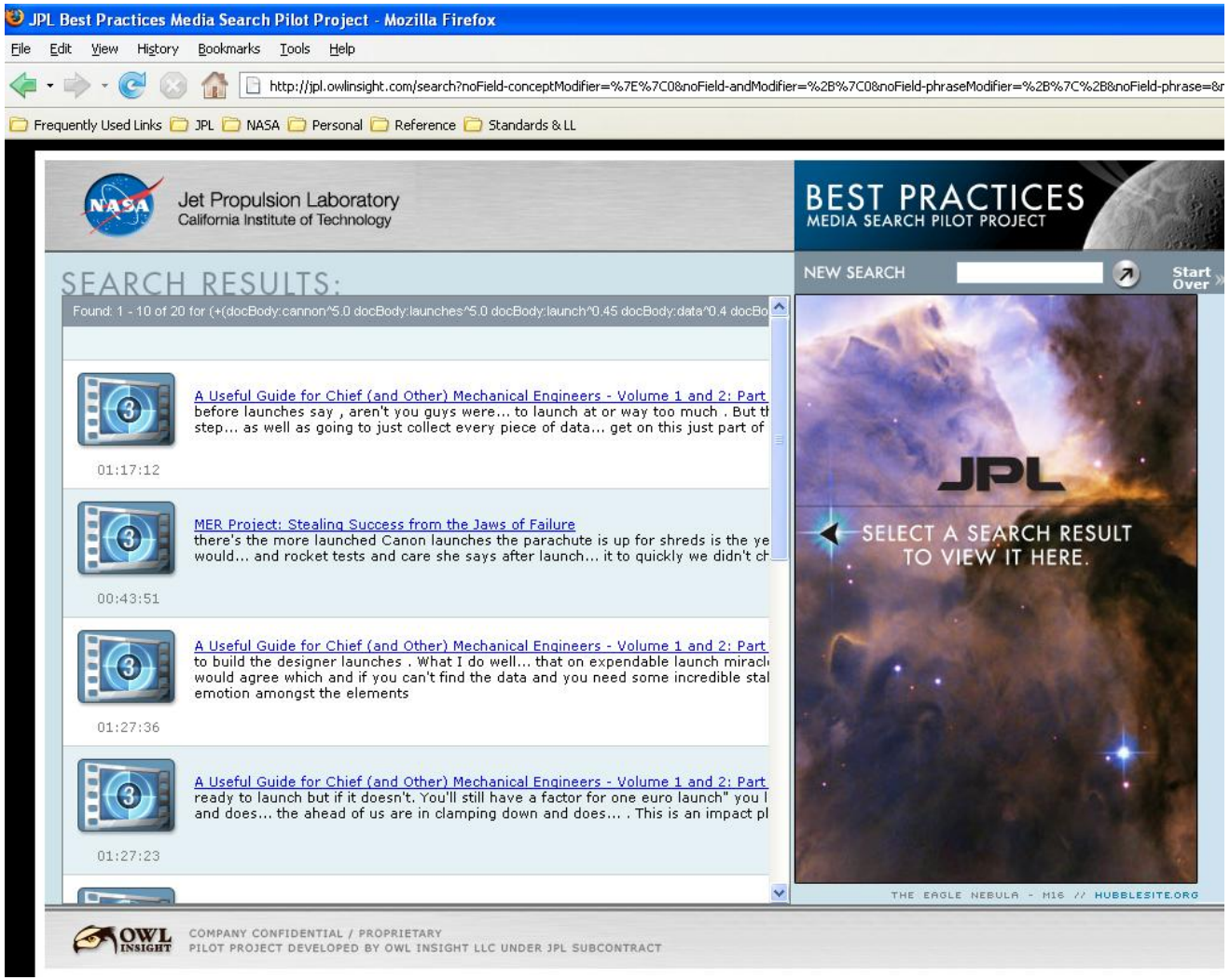
Media Search® Demonstration (#1)

Step 1. Online user enters search terms, and selects the file type “Video Files”



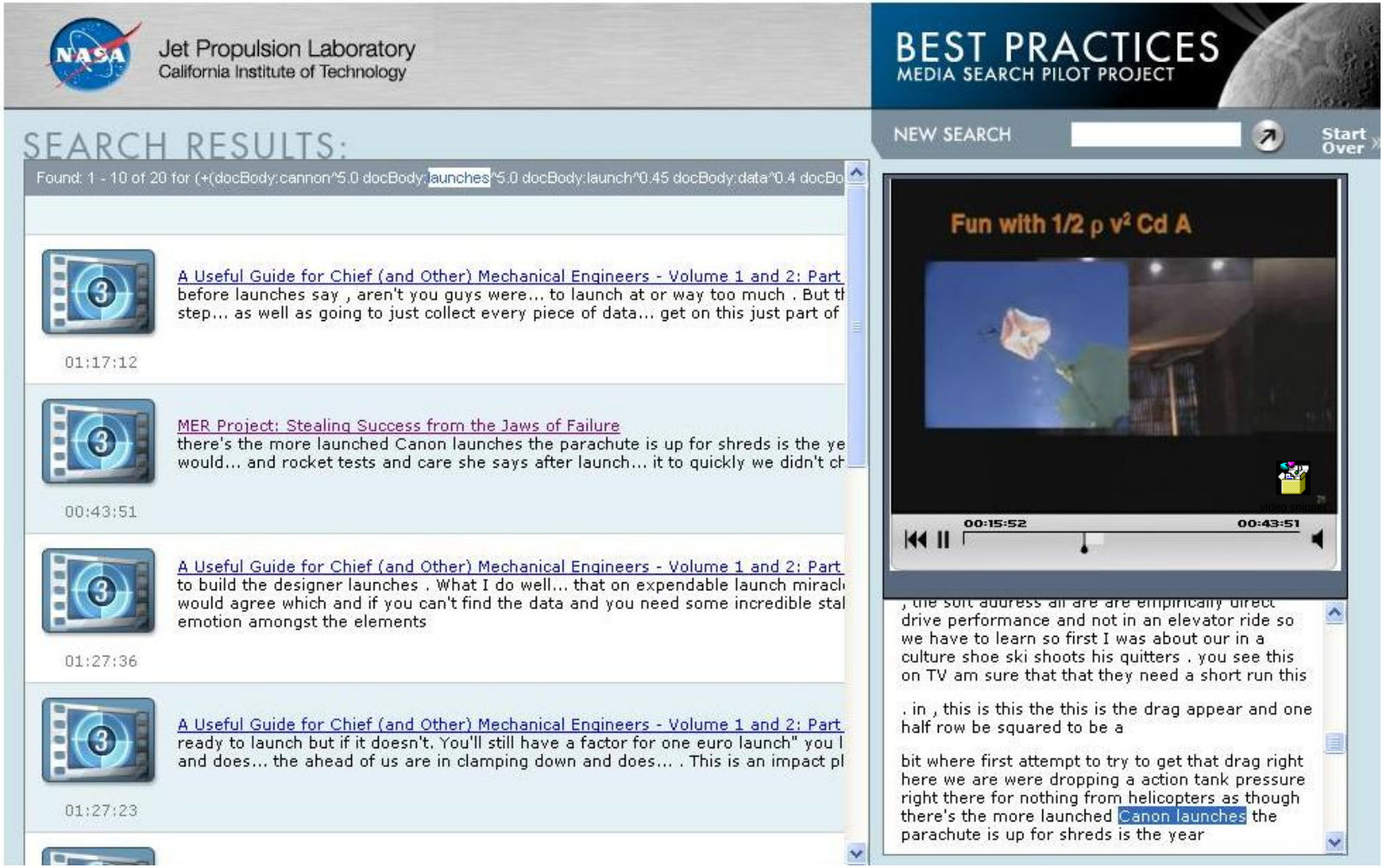
Media Search® Demonstration (#1)

Step 2. User selects a search result [selects "*MER Project: Stealing Success…*" video]



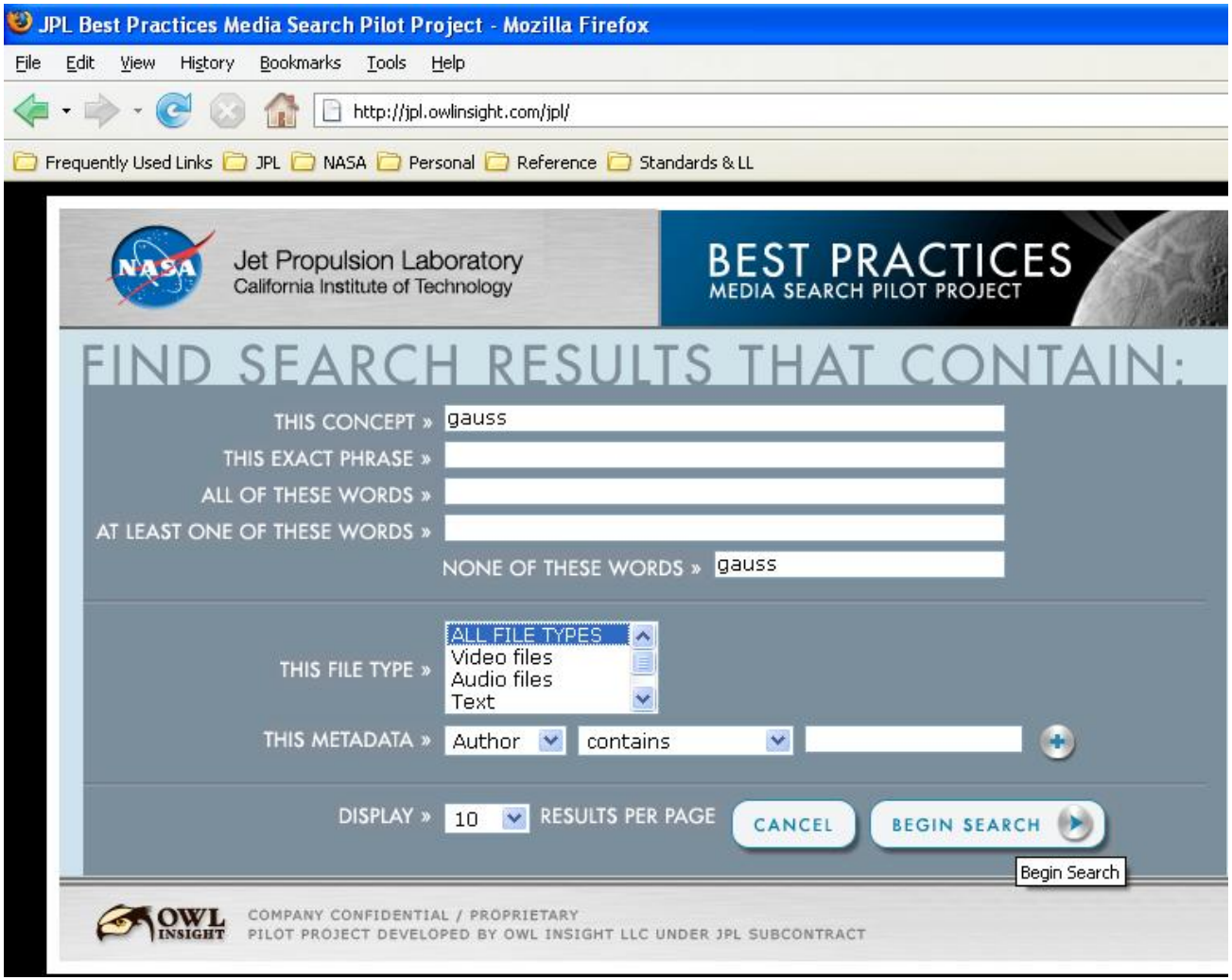
Media Search® Demonstration (#1)

Step 3. User plays short snippet from 43 min. video or reads transcript (note misspelling)



Media Search® Demonstration (#2)

Step 1. *All Files* concept search for “gauss” on files that lack the word “gauss”

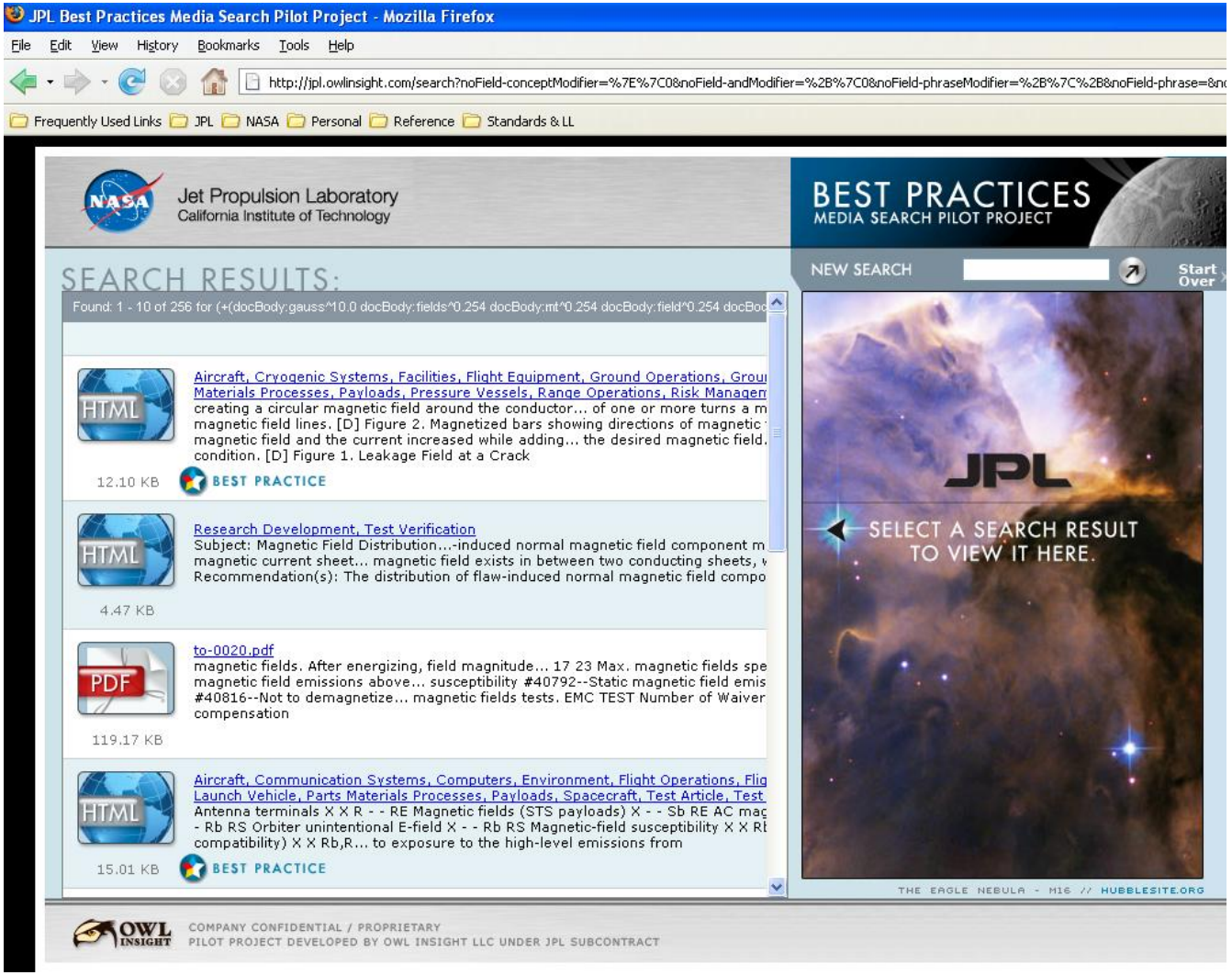


Media Search® Demonstration (#2)

Jet Propulsion Laboratory
California Institute of Technology

Office of the Chief Engineer

Step 2. These semantic-processed text files contain “magnetic field,” “magnetic susceptibility,” but not “gauss”



Plans for the 2nd Phase

JPL
Jet Propulsion Laboratory
California Institute of Technology

- The proof-of-concept pilot system now online allows users to search a set of 1700 files that were provided by JPL to the contractor, processed, and indexed on a server. Presently, administrators or users cannot add more files.
- A follow-on contract is planned for a Media Search® system that is scalable (i.e., additional servers and document collections can be added to the system) and that allows administrators to select and process additional files or collections so they will be available for subsequent searches.
- The initial application for the delivered Media Search® system may be the NASA Engineering Network (NEN), a NASA repository under development by JPL, which presently contains 1.3 million files. The system will be expandable; for example, it could be applied to indexing and quickly retrieving a key discussion from among hundreds of hours of videotaped spacecraft design meetings.
- The planned system will provide a tool and procedures for managing file ingestion and processing, allow users to perform a search-by-meaning (concept search) where they don't know the exact words used in a document or multimedia file, and incorporate the standardized NASA taxonomies.