Is Good Enough, Really Good Enough? Does algorithmic metadata search replace the need for discipline-oriented databases?

Mark Hyer
ProQuest, mark.hyer@proquest.com

Helen Ivy
College of Charleston, ivyh@cofc.edu

Bruce Pencek
Virginia Polytechnic Institute and State University, bpencek@vt.edu

Sharon Jordan
OSTI, jordans@osti.gov

Sharon Tahirkheli
American Geological Institute

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Library and Information Science Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Mark Hyer, Helen Ivy, Bruce Pencek, Sharon Jordan, and Sharon Tahirkheli, "Is Good Enough, Really Good Enough? Does algorithmic metadata search replace the need for discipline-oriented databases?" (2009). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284314744>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

IS GOOD ENOUGH, REALLY GOOD ENOUGH? DOES ALGORITHMIC METADATA SEARCH REPLACE THE NEED FOR DISCIPLINE-ORIENTED DATABASES?

Mark Hyer, VP STM Publishing, ProQuest, Moderator

Helen Ivy, Head Librarian, Marine Resources Library, College of Charleston

Bruce Pencek, College Librarian for Social Sciences, Virginia Polytechnic Institute and State University

Sharon Jordan, Assistant Director, Dept of Energy / OSTI

Sharon Tahirkheli, Director, Information Systems, American Geological Institute

Introduction: “Discipline-oriented databases: examining the shift from search to discovery”, Mark Hyer, VP STM Publishing, ProQuest

The role of abstracting and indexing has been to organize published research to dramatically increase the speed of access to information.

Historically named as Abstracting and Indexing (A&I) services, discipline-oriented databases remain relevant and a trusted source for subject navigation of a discipline with contextually-sensitive coverage of a field. Since 2000, new indexing processes and sophisticated linking to library collections via discipline-oriented databases have experienced exponential usage growth. Today, usage patterns of discipline-oriented databases remain steady, despite end users increasingly starting searches with “Mega-Metadata” services such as Google or Yahoo. Discipline-oriented databases support serious, comprehensive research by serving as discovery aids to the literature of that discipline. Subject-specific vocabularies aid users in pinpointing relevant literature and guide users to the source of the literature.

As models of scholarly communication, publishing, and information dissemination continue to transform, the use, value, and delivery of discipline-oriented databases is also evolving to a combination of “Search” and “Discovery”. ProQuest is publishing a white paper on this topic in 2010.

The panel formed in Charleston provided four unique vantage points:

- Subject specific librarian
- Interdisciplinary librarian
- Government aggregation publisher
- Discipline-oriented database publisher

Researchers Use of Databases: Discipline-Specific vs. Mega-Metadata, Helen Ivy, Head Librarian, Marine Resources Library, College of Charleston

The Marine Resources Library in Charleston, SC is a unique, scientific library serving researchers in the marine and aquatic sciences from three agencies—graduate/undergraduate students and faculty from the College of Charleston, research staff from NOAA / NOS / Center for Coastal Ecosystem Health and Biomolecular Research, and research staff from the South Carolina Department of Natural Resources / Marine Resources Division. Continuing budget reductions and a change in user behavior required that the librarian carefully study the value and effectiveness of the library’s subscription, discipline-specific databases to support the research interests of the scientists, faculty, and students. It appeared that all users were more likely to use Google Scholar as their first choice for a search engine to locate the available literature on their research topics. This raised doubt about the use of the subscription,

discipline-specific databases. Were researchers making effective use of the discipline-specific, subscription databases for their literature review resulting in cost-effective databases or did Google Scholar provide researchers the same search results at no cost to the library?

The librarian developed a survey using Survey Monkey and distributed the Database Evaluation Survey [see Fig. 1] to the graduate students, faculty, and scientists from the three agencies served by the library. The survey asked respondents to rank their choice, select their frequency of usage, and evaluate effectiveness for these databases: *ASFA: Aquatic Sciences and Fisheries Database*, *Environmental Sciences & Pollution Management Database*, *Oceanic Abstracts*, *PubMed Central*, *SciFinder Scholar*, *Web of Science—Science Citation*, *Zoological Record Plus*. [Note: PubMed Central was included because it was the only database to provide coverage of some new areas of research interest.] Next respondents were asked to list up to three topics that they routinely search in the discipline-specific databases. Respondents were then asked about the usage and value of the general databases *Academic Search Premier* and *InfoTrac Academic OneFile* to research their listed research topics. Finally respondents were asked about the usage and value of the mega-metadata databases Google Scholar and Science.gov to research their listed research topics. The survey was delivered by email to approximately 445 researchers with a 30% response rate including 11% faculty, 25% graduate students, 57% scientists, 7% technicians and 0.8% administrative staff completing the survey.

Survey Results for Scientific Databases:

The results for frequency of use for the scientific databases returned no surprises: Daily use – 24% of users; Weekly use – 47% of users; Bi-weekly use – 7% of users; Monthly use – 12% of users; Other – 10% of users.

Two questions asked users to indicate their first and second preference for the scientific databases. The top four databases were *Aquatic Sciences & Fisheries Database* (31%), *Web of Science* (30%), *PubMed Central* (22%), and *Zoological Record Plus* (5%). The top four databases for second preference were *Aquatic Sciences & Fisheries Database* (26%), *Web of Science* (26%), *PubMed Central* (17%), and *Zoological Record Plus* (10%).

Survey Results for General Databases:

Respondents (34%--generally students) reported using *Academic Search Premier* and/or *InfoTrac Academic OneFile* on a daily to monthly basis. Of those using the general databases, 5% felt the databases were extremely useful; i.e., they found all the sources they needed and never used the scientific databases; 15% felt the databases were very useful; i.e., they found key sources to get them started; 26% felt the databases were useful; i.e., they found one or two sources they could use. The databases were never used by 40% of the respondents.

Survey Results for the Mega-Metadata Databases:

For Google Scholar and/or Science.gov, 57% of the respondents used the databases on a daily to monthly basis. Twenty-eight percent of the respondents never use these sources. Google Scholar was preferred by 65% of those using the mega-metadata databases.

Comparison of Search Results from Four Databases

Selecting four topics that researchers routinely searched and using Joan Parker's methodology for evaluating bibliographic database overlap in marine science as detailed by her article in the

Spring 2005 issue of *Issues in Science and Technology Librarianship*, the librarian determined the number of unique sources cited for the period 2007-2008 in four databases. The four topics: Artificial reef [Reef], Coral health [Coral], Harmful algal blooms [HAB], Marine genomics [Gen]. The four databases: *Aquatic Sciences & Fisheries Database* [ASFA], *Zoological Record Plus* [ZR], *Web of Science* [WOS], Google Scholar [GS].

Findings:

Aquatic Sciences & Fisheries Database and *Zoological Record Plus* provided the most unique sources for the applied and biological sciences. *Zoological Record Plus* provided the best coverage of international publications. Google Scholar provided better coverage for health-related topics and for topics of popular interest. Google Scholar does cover from 50% to 90% of the grey literature cited in *Aquatic Sciences & Fisheries Database* and *Zoological Record Plus* for the four selected topics. *Web of Science—Science Citation* did provide better journal article coverage for some topics while it was equal to the coverage in Google Scholar for other topics.

Implications:

The above findings have direct implications for budgeting and for library instruction. The top three discipline-specific, subscription databases [*Aquatic Sciences & Fisheries Database*, *Web of Science—Science Citation*, *Zoological Record*] continue to be required if all of the library's researchers are to conduct complete literature reviews of their research topics. Instruction regarding database usage/coverage needs to be targeted to research groups so that all researchers understand the value and idiosyncrasies of databases specific for their research. Instruction for students must stress/demonstrate the value and advantages of the discipline-specific databases.

Maximizing the learning value of discipline-oriented databases for non-domain experts Bruce Pencek, Virginia Polytechnic Institute and State University

In our assumptions about research needs and behavior and in our allocation of research resources to connect students to scholarly information, librarians and teaching faculty risk privileging, at least implicitly, some modes of knowledge acquisition – and thus some users and some kinds of information – at the expense of others. Pervasive information technology exacerbates this risk. We educators have a pedagogical obligation to provide users cognitive infrastructures to frame their encounters with technologically constituted information. The low opportunity cost of finding *something* with a search engine on the free Web conceals the cognitive overhead that pervasive information imposes on researchers who enter unfamiliar knowledge domains. The information technology is at hand; search engines make enough sense, and quickly enough, to inspire confidence in users. On the other hand, for undergraduate university students, the substantive concerns, approaches, discourses, and technologies of most academic disciplines make sense only gradually. Most US universities concentrate this sense-making in academic majors. Core curricula and distribution requirements attempt to extend the sense-making into other disciplines, with different degrees of success.

I write from the standpoint of a onetime political science professor who, as an instructional librarian and investigator, has collaborated intensively with teaching faculty to wrestle with a fundamental question: *How can we induce students to operationalize their information needs, so that they can then choose the appropriate resources and use appropriate search techniques to fulfill specific research tasks?*

There are several factors that invite everyone – students, but also faculty and librarians – to beg that question:

The ready availability of search and retrieval systems (paradigmatically, Google) rewards both naïve searchers – who have a broad, sloppy notion of relevance – and domain experts – because they bring very precise notions of what they need and what words describe it. By their very different lights, both succeed. If both naïve undergrad and expert professor “know” how to find the information they need, each may assume the other thinks the same way and neither would feel compelled to rethink his/her own familiar process.

The disciplinary nature of the university creates and reinforces cognitive horizons at the same time as it differentially allocates information resources. Ambiguity is a vexing problem in any searching. This is further compounded by the often subtle varieties in denotations and connotations of significant terms across disciplinary lines, and sometimes across subfields and other fault lines within a discipline and between academic and lay communities whence the undergraduates come.

Even as they indulge the all-too-human inclination to clone themselves in their students, domain experts often forget that their own habits of mind as professional analysts and researchers are learned abilities. It is second nature for a couple professors chatting at a conference to turn a passing observation into an “Interesting question,” then hypothesize explanations, all in a few moments.

These factors reflect a common reality: there are knowledge insiders, and there are knowledge outsiders, separated by a cognitive chasm. The insiders know enough in their domains to recognize holes in their knowledge: like Socrates, they know that they do not know, and proceed to fill the gaps. Conversely, domain outsiders’ understanding is not developed enough to appreciate the contours of their ignorance – the known unknowns, both substantive and procedural. Moreover, their self-confidence as heretofore good-enough searchers makes them less attentive to signs of potentially useful knowledge that is simply beyond their perceptual or cognitive frameworks – their unknown unknowns.

Our experience and assessment data from political science courses at Virginia Tech suggest that students value instruction and exercises that develop the purposive habits of mind of insiders along with technical facility using information resources. We have developed a simple analytical framework that librarians and teaching faculty can use to identify tendencies in students’ research behaviors in order to guide interventions that reinforce desired information behaviors and redirect inappropriate ones. We have also developed a learning model whereby students follow a recommended sequence, which we call *word mining*, for working through different kinds of information sources to learn the “unknown unknown” terminology of unfamiliar domains and apply them in increasingly precise searches to master ever-more inclusive information environments. For the present discussion, I focus on the essential role played by discipline-oriented databases – crucially incorporating attributes of traditional online abstracting and indexing – in that larger process.

When we know what kind of in-domain knowledge researchers bring to the task and we identify the cues or topic-framing systems they can use to navigate knowledge space, we can expect different patterns or tendencies of search behavior. From these we can select appropriate kinds of intervention, from the way course instructors design assignments to the resources and techniques instructional and reference librarians counsel. Our typology should also serve as a

reminder to teaching faculty and librarians to reflect on their own assumptions and propensities as researchers.

Table 1. Tendencies of knowledge acquisition in library research

	Resource cues/frames		
		<i>Grounded / Explicit</i>	<i>Abstract / Self-generated</i>
Researcher's knowledge base	<i>Strong</i>	Recapitulate disciplinary norms	Retrieve
	<i>Weak</i>	Discover	Forage laterally

As experts, professors leverage their existing knowledge. They know the culture of their fields – the jargon, the canonical writers, the outlets for scholarly communication – and use them to structure their search for and application of information. They have firm ideas of what information they want and they have the wherewithal to look for it efficiently. As their expertise deepens, their opportunity cost decreases for retrieving more information in their domains. Full-text sources (the more, the better) suit their needs well. To the extent that searchers can bring a discipline's settled terminological conventions with them, explicit controlled vocabularies are less salient, and the evolving semantic web may well satisfy their sense of need. The risk is that experts' established paradigms become self-reinforcing within narrow knowledge horizons, potentially oblivious of relevant work and usages outside their own.

Undergraduate students are not (for the most part) stupid. But they are, by definition, relatively ignorant. Whatever they may learn in K-12, they come to college with at best a weak, and often distorted, acquaintance with academic disciplines, conventions, and mores. Confronted with assignments that assume insider perspectives, they gather information through brute empiricism, foraging for results that are good-enough to them because they do not know the more exacting standards by which experts teaching their classes will grade them. Open-ended, indiscriminate searching in unstructured information spaces like Google makes sense to them, however accidental their technique appears to us.

The antidote, of course, is to provide students explicit opportunities to discover how knowledge is structured and to situate their need within one or more structures. We must remember that these are structures of meaning, not simply of information. In this respect, traditional library discovery practices such as physical collocation of related works and controlled vocabularies may have the advantage over more recent, cybernetic discovery tools: physical objects may carry more meaning than abstract labels, and a rigorous controlled vocabulary might actually be a less abstract point of departure for a naïve user than a word cloud distilled from a Babel of natural language.

Pragmatically, undergraduates must learn to think and research like their professors. These naïve users need help and experiences to discover where the relevant work is done, who the experts are, and what terms they use to talk about their work. The core reality in US higher education is that both professorial expertise and curricula continue to be built around disciplines. Disciplines structure meaning, not merely information. The ubiquitous use of major fields, complemented by distribution requirements and core curricula to expose students to other disciplines, assumes that students cannot learn disciplinary knowledge on their own but rather must receive formal instruction. It is odd, then, that so many people in higher education seem to

assume that students will discover on their own how to acquire and use information in the way disciplines define. Given the time constraints built into US baccalaureate programs and the magnitude of the cognitive burden, it is both inefficient and irresponsible not to teach undergraduates the tools that help them discover information as yet unknown to them as well as those that help them retrieve information about which they already do know something.

Word mining

The word mining process does not try to overcome the habits and inclinations of naïve researchers. Instead, it directs them to where they can do the most good. As the name suggests, the process invites searchers to mine resources at one stage for potentially relevant search terms to use in later stages, moving from non-domain search terms (and concomitantly loose stories) to in-domain terms applied to in-domain sources. Disciplinary databases that include subject taxonomies are essential tools for crossing the domain-expertise boundary. Each stage presents instructional opportunities to introduce classes to appropriate resources and to issues about their significance. Ideally these instructional and reference interventions include hands-on exercises that lead to assessable outputs.

The first and most difficult step is to lead students very rapidly to articulate *stories* – plausible, provisional explanations of why something is the way it is, based on what they already know. Laying out the story is analogous to, and can work together with, the development of theories and hypotheses that guide empirical research. Students who must state their stories bring into the open what they must look *for* and why.

The emphasis upon telling a story is active and purposive, in contrast with the language of topics or even of research questions invites students to be passive, to wait until they have “all the information” that will mysteriously come together at some indefinite future time. In many respects, the story trope can serve at least to raise consciousness about the adequacy of their existing knowledge and understanding. Reflection and discussion about the initial story extends this self-awareness to the logic of the potential explanation. Articulating the story necessary lays out a relation of concepts – evidence and premises – which are expressed in a first round of potential search terms.

The first step of word mining proper is information environmental scanning. This is an explicit application of students’ inclinations to forage, to play around with sources and searches to see what happens. Like the story, the goal of this stage is not so much to be “right” or complete, but rather to form impressions about what other stories exist, the evidence they deploy, and how they square with their own provisional hunches. Habituated in their prior schooling to associating research with writing mere reports, undergraduates need to be shown that information is nested in contestable knowledge claims. Moreover, they be made of the paradox that widespread agreement in the literature on their initial story is a warning flag to think about new angles that will make their projects genuinely their own: pedagogically, reiteration does not develop analytical and synthetic habits of mind, and confirming the conventional wisdom is not “interesting” in academic culture.

Thus during the environmental scan, gray and interminable Google results lists, the Wikipedia discussion tab, the cranky reader replies to online newspaper stories, talk radio, faux television news are all within bounds, along with conventional background sources. Similarly, at this stage search strategies can be relatively loose: students can play with different combinations of terms and explore operators and filters. This is a time for exploration, which can be complemented by peer-to-peer conversation and formal instruction.

There is even a virtue to incompatibility and wackiness among the results of the environmental scan: the student is unlikely to be able to reconcile everything and will have to think through what parts fit with one another and what must be set aside. The odds are that the student will encounter solid sources that relate to at least some part of the story, which will of course provide additional search terms, including names of authors and authorities, possibly bibliographic references as well as knowledge claims.

Instructional and reference interventions by the library at this stage can appropriately address the nature, purpose, and use of reference collections and reference tools like subject encyclopedias; broad, general-purpose databases like Infotrac OneFile and Academic Search Complete; and the subject-based organization of the library and its services (eg, classification but especially subject resource guides and subject librarian contact). This is also the appropriate stage for instructors and/or librarians to discuss disciplines, evaluation of information, and rhetoric.

In the face of information glut, and the searcher's limited insider knowledge, less is more. Thus the central stage in word mining is to use the terms developed in the previous stages to search representations of information sources: catalogs, abstracting and indexing databases, mixed databases with searches limited to citation and abstract fields. The brevity of bibliographic records and abstracts tends to force precise, high-value words to the fore. More precisely: humans writing abstracts and assigning subject headings choose words that precisely as well as concisely describe the work. In disciplinary databases, these will be insider terms that will uncover additional records within and across other A&I databases and then unlock full-text sources in the next stage.

An ensemble of characteristics of disciplinary databases built around A&I principles make them especially valuable for bridging the cognitive gap. The parts are familiar search fields: article titles, text of abstracts, and descriptors. The contents of each field in a record are in themselves meaningful clusters of words, each more or less intelligible to the non-expert searcher. Taken together they present slightly different aspects of the content of the word the record represents. Each is likely to hook different terms from ordinary keyword searches, allowing multiple results sets to be compared and manipulated.

The process is recursive. Words discovered in the records from one search become fodder for further, often more refined, keyword and subject searches. The works behind the records produce more, and more in-domain, search terms – and of course, substantive content. Thus we use various pedagogical strategies to give students experiences in reading records on two levels. (Think bibliographic Kabbalah.) First there is the familiar, literal processing of the description of the work the record points to. By a process of mental triangulation from these fields, the students can get a pretty good idea of the work and decide its appropriateness to the project at hand. But we further ask naïve searchers to embrace their naiveté and interrogate the records: What *could* this unfamiliar word (or concept or argument) mean? How *might* it connect to another one in the record, in other records, and in the story research project at hand wants to tell? (Take that old favorite LC subject heading, *cooking*, add subdivisions, and stir briskly.) That is, we ask the students to reflect on potentialities of meaning and relevance embodied, however imperfectly, in a record. We want to cultivate in their search practices an attitude of openness toward unexpected connections and interpretations that reflects the openness and curiosity professionals bring to their research practices, whether textual or empirical, and make the research their own.

The juxtaposition of terms may be as significant as their simple denotation. For domain outsiders, parsimonious descriptions in conventional OPAC records (and databases that essentially digitize old print indexes without adding content to records) may make it difficult to use mental triangulation to decide if a work is worth pursuing. But records have grown richer over the past 30 years, and now they are mashed up with supplemental information such as social tags that bring unite contemporary argot with stable descriptions. Faceted displays increase the number of data points where a record could mesh with the student's existing understanding and geometrically increase the possibilities for finding connections. To mental triangulation, add connecting dots.

Subject headings are especially valuable. We all know the shortcomings of taxonomies and thesauri: their one-dimensionality, culturally and historically loaded language, the painful locutions, the all-too-human inconsistent application. As Clay Shirky says, ontologies are overrated – if understood as accounts of reality. But that does not mean they are not useful. For all of that, they *are* pointers to discover bounded universes of information within the catalog or database. Additionally, many times they reveal a technical, insider's vocabulary to mine for keyword searches in full-text resources.

Again, the juxtaposition of words may be more important than the words separately to help naïve searchers situate themselves within disciplinary scholarship. Fielded searches are key to the process of mixing and matching. The first step is explicitly contrary to many students' inclinations and to the advice of some librarians: keyword search for each term singly to see how well it works for one's story in the catalog or database at hand, look for appropriate subject terms, and only then combine the searches. Subject terms function to define chunks of related works, keywords represent aspects within them.

This combining and recombining of chunks and aspects can be done manually in any database that uses both subject and keyword searches, but it is much easier where well-designed search history functions (eg, in the CSA Illumina and OCLC FirstSearch platforms) make it easy to put these discrete searches into the search-history bucket from the results list, then combine searches with a few clicks. (The search history on the EbscoHost platform requires an inconvenient extra step: opening a record to run a fresh subject search across the database.) Our students say they love this technique as a way to let the database tell them what information exists rather than to try to think out comprehensive combinations of search terms a priori. Their biggest complaint is not with searching A&I architectures but rather with how many seemingly valuable journals are not available through our link resolver.

Analogous to our advice about searching terms singly rather than in combinations, we teach students that it is ultimately more efficient to search one database at a time rather than across a family or with a federated search tools. Cross-database searching blurs the differences among disciplines, makes it more difficult to discern what is in unfamiliar databases, identify what terms work, and diagnose why a search may fail. Crucially, they dilute the power of subject searches to resolve database content into smaller, manageable chunks.

The final two steps in word mining are straightforward: apply the searchable words discovered heretofore in full-text sources, first those the library subscribes to, then the free web. We recommend that order because the library resources provide cues to their value, along quality control regarding content, power of search features, and complementary with other resources. Library subscription full-text sources divide into those that add substantial full-text coverage and A&I functionality (eg, ABI/Inform, various EBSCO products) and those with little or no subject classification (eg, CIAO, LexisNexis Academic), and we recommend that students use them in

that order. The mixed A&I full-text databases in our library tend to have disciplinary emphases. Our pure-play full-text subscription sources are more likely to be distinguished by genre (journalism in Factiva or historical newspapers; journals in JSTOR) than disciplinary focus, though our subject guides (and reference/instruction interventions) help our users decide which ones are potentially appropriate to their stories.

The final, clean-up stage resembles the environmental scan: back to the open Web, but with a clearer idea of the story, what it requires, and a rich repertoire of domain-significant search terms. By now, students' confidence in their search ability is justified, and the task of librarian and course instructor in that regard is to point out less familiar places to look through, such as the invisible Web and conference paper servers.

Is Good Enough Really Good Enough? "Government R&D Results Distributor to Accelerate Science Discovery", Sharon Jordan, Assistant Director for Program Integration, U.S. DOE Office of Scientific & Technical Information

We at DOE's Office of Scientific and Technical Information are fulfilling our mission by making government-sponsored R&D information available via online databases - now with single-query portal interfaces that make DOE's R&D results easily accessible by many information seekers. These portals include (1) the Science Accelerator (www.scienceaccelerator.gov), which aggregates DOE databases of R&D information, (2) Science.gov, which enables you to find U.S. science information from multiple agencies, and (3) WorldWideScience.org, which finds global science information. Through these gateways, librarians, researchers, and the public can access a science page count comparable to, but not duplicative of, the entire science content of Google™.

For purposes of this discussion, we will focus on Science.gov, which integrates the databases and websites of 14 U.S. science agencies. Science.gov results are *uniquely different* from Google and Google Scholar. Science.gov Search Results Screen Shot

The screenshot shows the Science.gov search results page. At the top is the Science.gov logo with the tagline "USA.gov for SCIENCE" and "Version 5.0". Below the logo is a navigation bar with links: Home, Site Map, Index, Alerts, Help, Contact Us, About, Communications, and Alliance Only. The main search area includes a search bar with the query "solar", a "Refine Results" button, and a "New Search" button. To the right of the search bar are links for "Advanced Search" and "Deep Web". Below the search bar, it shows "Search: Full Record: solar" and "1,782 ranked results of 60,224,002 available". There is a "Create alert from this query" link and a "1432 more results found" notification. A "Summary of All Results" link is also present. The results are displayed in a table with columns for "Rank", "Limit to", and "All Sources". The first result is "Solar Lentiginos Treatment With the Triple Combination Cream" from ClinicalTrials.gov, dated 2009-09-10. The second result is "Solar Flares" from NASA Website, dated 2009-09-10. The third result is "Effect of Solar-Simulated Ultraviolet Radiation on Gene Expression in Unprotected and Sunscreen-Protected Skin of Healthy Adults With Fitzpatrick Skin Type II". On the left side, there is a "Clusters" section with "All Results (350)" and "Topics" including "Solar System (52)", "Energy (22)", "Solar Radiation (20)", "Activity (20)", and "Atmospheric (19)". On the right side, there is a "Wikipedia" section with a link to "Solar" and a brief description: "Solar means pertain to the Sun, or Sol, our planet's star. Solar also has other meanings. more... [exit federal site disclaimer]".

Through specialized federated search technology, Science.gov drills down to selected databases and websites in parallel, then presents ranked search results from over 40 databases, thousands of authoritative scientific websites within .gov domain/sponsorship, and more than 200 million pages of science information. Science.gov also helps users identify databases for more advanced searching as needed.

Recently, OSTI compiled 33 sample searches of scientific search terms typical of what a researcher would look for, i.e., "adenosine triphosphate", "influenza vaccine" transplant patient, and "laser cooling." We found that 95.8% of the time Science.gov results are *uniquely different* from Google and Google Scholar results. Science.gov results have only 1% overlap with Google and an approximate 3.2% overlap with Google Scholar.

Sample

	A	B	C	D	E	F	G	H	I	J	K	L
4	Search Term	Results	Results	Results	# Unique	% Unique	Google Overlap	%	G.Scholar Overlap	%	"Top 50" Overlap	%
21	gene expression biomass yield	421	956	968	968	100.0%	0	0.0%	0	0.0%		0.0%
22	high temperature superconductor	433	871	1144	1103	96.4%	15	1.3%	26	2.3%		0.0%
23	"influenza vaccine" transplant patient	487	807	522	509	97.5%	2	0.4%	11	2.1%		0.0%
24	laser cooling	520	875	1184	1084	91.6%	33	2.8%	67	5.7%		0.0%
25	magnetoresistance	303	898	1119	1107	98.9%	6	0.5%	6	0.5%		0.0%
26	"metabolic engineering" enzyme	566	909	886	863	97.4%	8	0.9%	15	1.7%		0.0%
27	nanotechnology drug administration	640	833	1024	1008	98.4%	12	1.2%	4	0.4%		0.0%
28	organic aerosol clouds	597	847	1207	1169	96.9%	22	1.8%	16	1.3%		0.0%
29	petroleum hydrocarbons	701	821	1900	1874	98.6%	5	0.3%	21	1.1%		0.0%
30	phytoremediation "waste management"	573	957	1078	1053	97.7%	8	0.7%	17	1.6%		0.0%

Because Science.gov uses an integrated search of individual databases, Science.gov users benefit from both platforms. There are benefits to using an integrated search tool as well as to using discipline-oriented databases. As an integrated search tool, Science.gov offers users the opportunity to quickly scan the environment relative to what they are seeking, find the relevant information they seek, and drill down for more. Results are typically current, as they are found in real time. Once users locate the databases they need, the benefit of presentation and depth of information of discipline-oriented search come into play.

Science.gov and other integrated search gateways such as Science Accelerator and WorldWideScience.org find information for users that generally cannot be found through Google Scholar/Google and other traditional crawlers.

GeoRef and Google Scholar: Some Comparisons, Sharon Tahirkheli, Director, Information Systems, American Geological Institute

As the publisher of GeoRef, the primary bibliographic service for the geological sciences, I was asked to provide some thoughts on the differences between a discipline-specific bibliographic service like GeoRef and a search engine like Google Scholar and, looking to the future, outline some potential steps that a traditional A&I service could take to enhance its perceived value to the user.

In 2007 and again in mid-2009 I examined the following questions: What are the characteristics of a discipline-specific bibliographic service like GeoRef and a search engine like Google Scholar™? And what are the strengths of each type of service?

The procedure used for comparing the two different services included examining the web sites and testing each system with three different types of searches – editor/authors, known-item and subject searches. Using some of the same searches in 2007 and again in 2009 also allowed changes in the systems to be observed.

Strengths

Google Scholar's continuous updating provides the user the opportunity to see recent materials and its broad coverage of science enhances interdisciplinary access to materials. Access is

easy and is simple for the user to use. Searches produce large numbers of hits. And, of course, it is free!

GeoRef has powerful search features enabling the user to limit searches to very specific queries. Coverage is limited to the geologic sciences eliminating the need to hunt through huge lists of results. A Controlled vocabulary enhances retrieval for specialists and the coverage includes grey literature, non-English language and print publications.

Full-Text Access

The relative strength of the availability of full-text access from results in Google Scholar and GeoRef was impossible to assess. While Google Scholar provides links to Web-based materials, many of the materials are not freely available and have access controls that prevent use except by users in institutions that license the content. Access to full-text through GeoRef is dependent on the user's library access system and the user's location (library, campus, remote). Potential Enhancements for Discipline-Specific Databases like GeoRef

While it is clear that traditional A&I Services like GeoRef continue to have value, there are some enhancements that could have a significant impact on use. More frequent updating of the database with improved currency can help to eliminate the traditional lag time for A&I databases. In 2008 GeoRef began making current records available in a file known as GeoRef in Process. These current records are made available before they have been fully edited and indexed for GeoRef. The records contain basic bibliographic information and provide the user the opportunity to find at least a preliminary reference without waiting for the complete version of the reference.

Enhanced indexing of figures, maps and tables can provide the user with additional access to geological publications. The illustrations in the geoscience literature can, at times, be the primary focus of a publication. The GeoRef database on Proquest has been enhanced with Deep Indexing of images from selected journals.

Map-based search and display can be attractive to users and has the potential for providing the user with a geographic way of coordinating results. The GeoRef database on Engineering Village now provides a Google mapping interface that allows geologists to see the geographic locations of their results.

Improving the utilization of discipline-specific Controlled Vocabulary in retrieval could reap significant benefits to users. Current implementations of GeoRef do not maximize the full potential of a Controlled Vocabulary by using the many inter-relationships between the vocabulary terms in any systematic way.

Is Good Enough Really Good Enough?

Discipline-specific databases continue to be necessary for comprehensive research and for locating non-web based and foreign-language materials; their powerful search features permit precise and flexible queries and eliminate false retrieval.