NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DEVELOPING INFORMATION STORAGE AND RETRIEVAL SYSTEMS ON THE INTERNET: A KNOWLEDGE MANAGEMENT APPROACH

by

Charles A. Fulmer

September 2011

Thesis Advisor:	Mark E. Nissen
Second Reader:	Eleanor Uhlinger

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704–0188) Washington DC 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2011	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE Developing Information Storage and Retrieval Systems on the Internet: A Knowledge Management Approach	**5. FUNDING NUMBERS**
6. AUTHOR(S) Charles A. Fulmer	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943–5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ___N/A__________.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public releases; distribution is unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Search is becoming the primary way in which people get information. In 2010, global Internet usage was over two billion people, with 92% of online adults using search engines to find information. Most commercial search engines (Google, Yahoo, Bing, etc.) provide their indexing and search services at no cost. The DoD can achieve large gains at a small cost by making public documents available to search engines. This can be achieved through the utilization of important design components and effective knowledge management. This thesis examines methods for making information available to search engines at the Naval Postgraduate School (NPS) and the Defense Technical Information Center (DTIC). In a large-scale project, over 200,000 documents were organized on the website dodreports.com. The results of this research revealed improvement gains of 8–20% for finding reports through commercial search engines during the first six months of implementation.

14. SUBJECT TERMS Information Systems, Knowledge Based Systems, Knowledge Management, Technical Information Centers, Information Retrieval, Department of Defense, Theses	**15. NUMBER OF PAGES** 91
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

DEVELOPING INFORMATION STORAGE AND RETRIEVAL SYSTEMS ON THE INTERNET: A KNOWLEDGE MANAGEMENT APPROACH

Charles A. Fulmer
Lieutenant, United States Navy
B.S., U.S. Naval Academy, 2006

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
September 2011

Author: Charles A. Fulmer

Approved by: Mark E. Nissen
Thesis Advisor

Eleanor Uhlinger
Second Reader

Dan C. Boger
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Search is becoming the primary way in which people get information. In 2010, global Internet usage was over two billion people, with 92% of online adults using search engines to find information. Most commercial search engines (Google, Yahoo, Bing, etc.) provide their indexing and search services at no cost. The DoD can achieve large gains at a small cost by making public documents available to search engines. This can be achieved through the utilization of important design components and effective knowledge management. This thesis examines methods for making information available to search engines at the Naval Postgraduate School (NPS) and the Defense Technical Information Center (DTIC). In a large-scale project, over 200,000 documents were organized on the website dodreports.com. The results of this research revealed improvement gains of 8–20% for finding reports through commercial search engines during the first six months of implementation.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

CFM	ColdFusion Markup
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DAU	Defense Acquisition University
DBMS	Database Management System
DTIC	Defense Technical Information Center
DoD	Department of Defense
DoE	Department of Education
ER	Entity-Relationship
FTP	File Transfer Protocol
GSA	Google Search Appliance
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
IM	Information Management
IT	Information Technology
IIS	Internet Information Services
KM	Knowledge Management
KMS	Knowledge Management System
KP	Knowledge Portals
NASA	National Aeronautics and Space Administration
NPS	Naval Postgraduate School
MB	Megabyte
OJT	On-the-Job training
OCR	Optical Character Recognition
PDF	Portable Document Format
POP	Post Office Protocol
ROI	Return on Investment
RSS	Really Simple Syndication

R&D	Research and Development
SEO	Search Engine Optimization
SQL	Structured Query Language
SMTP	Simple Mail Transfer Protocol
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Mark Nissen for his guidance during the work performed in this study. I have learned a great deal from you, and it was truly rewarding to apply my newly mastered knowledge management skills towards a real life problem.

Second, I would like to thank Eleanor Uhlinger for her guidance.

Lastly, I would like to thank my mother, Cassie Featherston. Mom, you have always inspired me to pursue opportunities and to not be afraid of challenges. Thank you for always being there for me.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Knowledge is a critical resource for all organizations, especially in the Department of Defense (DoD). Drucker (1993) pointed out that the most valuable assets of the 21st-century enterprise are knowledge and knowledge workers. As demonstrated by Nissen, in *Harnessing Knowledge Dynamics* (2006), knowledge enables direct action and competitive advantage, and is distinct from both data and information. Knowledge may be tacit or explicit and may reside in individuals, groups, documents or repositories. Knowledge Management (KM) is a strategy aimed at increasing organizational competitiveness (Bell & Jackson, 2001). It involves a “distinct but interdependent process of knowledge creation, knowledge storage and retrieval, knowledge transfer, and knowledge application” (Alavi & Leidner, 2001, p. 131). KM is the process “of generating, codifying, and transferring explicit knowledge within an organization” (Halawi et al., 2006, p. 388). One of the key benefits of KM is the ability to build and maintain a sustainable competitive advantage. Components of a competitive advantage involve lowering costs and enhancing differentiation. The path to achieve successful KM programs is not easy, as it requires the integration “of knowledge methods, technologies, and organizational forms to business strategy” (Halawi et al., 2006, p. 394).

In recent years, knowledge management systems (KMS) have become popular as organizations leverage new Information Technology (IT) systems and capabilities. IT is continually expanding the limits of what organizations can accomplish, as processing power exponentially grows and the costs associated with data storage are reduced. According to Moore’s Law, this trend will continue as the number of transistors on a chip doubles approximately every 18 months. The objective of KMS is to “support the creation, transfer, and application of knowledge in organizations” (Alavi & Leidner, 2001, p. 107). KMS can play an important role in large-scale KM projects through the effective use of the networks, browsers, databases and data mining. However, these technologies alone are not sufficient to enable knowledge flows. KM projects and KMS require more than just IT; the key is how the technology is applied and how it supports people and learning (Nissen, 2006).

The application of technology to knowledge portals (KP) can make them more accessible, improve knowledge flows and contribute to innovation. Incorporating search engine access into KMS can provide high impact at low cost. For example, millions of people depend on search engines (Google, Yahoo, Bing) to find the information that they are looking for at almost no cost. Our society has largely become a searching culture, with 92% of online adults using search engines to find information (Purcell, 2011). Search is becoming the primary way in which people get information. As our culture changes, it is important that all organizations recognize the power of Internet search and include it in their KM strategy.

To be found and indexed on the Internet, organizations must remove barriers that could block search engines from indexing websites or KPs. With a small investment, organizations can dramatically increase the quantity of knowledge stock that can be found using search engines. Organizations can evaluate their search strategy by asking the following questions:

- Has the organization thought about the ways people might be searching for documents?
- Can search engines find and extract the web pages and documents?
- Is the extractable information accurate?

Understanding these questions is important to ensure that the organization is not ignoring high impact solutions or creating barriers that will block potential.

In this report, the information storage and retrieval techniques used at the Naval Postgraduate School (NPS), a naval/defense-oriented research university, and the Defense Technical Information Center (DTIC), a distributor of authoritative DoD scientific research and engineering information to the defense community, are evaluated. Tests are conducted to determine how well search engines can index the KP content. At the end of the case study, a demonstration website located at dodreports.com is used to show the best practices and recommendations found during the research.

This research seeks to answer the following questions:

1. How well are NPS and DTIC harnessing the power of commercial search engines?
2. How can the DoD increase explicit knowledge presence through search engines?

The balance of this thesis is organized as follows. Chapter II includes a review of existing literature. Chapter III discusses the research methodology used to conduct the research. Chapter IV summarizes the results of the research. Chapter V is a discussion of the key results, conclusions drawn from this research, and recommendations of key interventions that the NPS Library and DTIC could act upon to advance their KM practices, as well as suggestions for follow-on research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

A. KNOWLEDGE

Drucker (1993) identified knowledge as a critical resource that has surpassed capital, labor and natural resources in the global market. By treating knowledge as a resource, organizations can create significant value and achieve long-term competitive advantages. Davenport and Prusak (1998) concluded that a knowledge advantage is a sustainable advantage because it generates increasing returns and continuing benefits. Bixler (2005) stated that organizations use knowledge to execute processes, to make decisions, and to improve efficiency and effectiveness. Given the value of knowledge, it is not surprising that organizations are investing in KM. However, before making this investment it is important for organizations to determine what knowledge is, how it differs from data and information, what forms of knowledge exist and how knowledge moves throughout an organization.

B. DEFINITION OF KNOWLEDGE

Many definitions of knowledge exist. Since the classical Greek era, the quest for a definitive definition of knowledge has continued without a clear consensus. Davenport and Long (1998) defined knowledge as “information combined with experience, context, interpretation, and reflection” that is ready to “apply to decisions and actions” (p. 43). Alavi and Leidner (2001) stated that “knowledge is information possessed in the mind of individuals: it is personalized information (which may or may not be new, unique, useful, or accurate) related to facts, procedures, concepts, interpretations, ideas, observations, and judgments” (p. 109). Within the DoD, knowledge is defined at the Defense Acquisition University (DAU) as “The ideas, understanding, and lessons that an organization has learned over time … knowledge is condensed information with context that has value for decision and action” (Pollock, 2002, p. 220). One important takeaway from the definition of knowledge is that knowledge enables action, and this property makes it unique from data and information.

C. KNOWLEDGE UNIQUENESS

In order to understand KM, it is important to understand that knowledge is unique and distinct from data and information. In order to visualize the differences between these, many scholars (Davenport & Prusak, 1998; Nissen, 2006) conceptualized a hierarchy of data information and knowledge, shown in Figure 1. Each level of the hierarchy builds upon the previous level to provide increasing actionability. Data is raw numbers and facts. Information is processed data and has context. Finally, knowledge is “personalized information” possessed in the mind of individuals that “may or may not be new, unique, useful, or accurate, related to facts, procedures, concepts, interpretations, ideas, observations, and judgments” (Alavi & Leidner, 2001, p. 109). Although knowledge is distinct from data and information, it is not independent; knowledge usually requires data in order to be actionable.

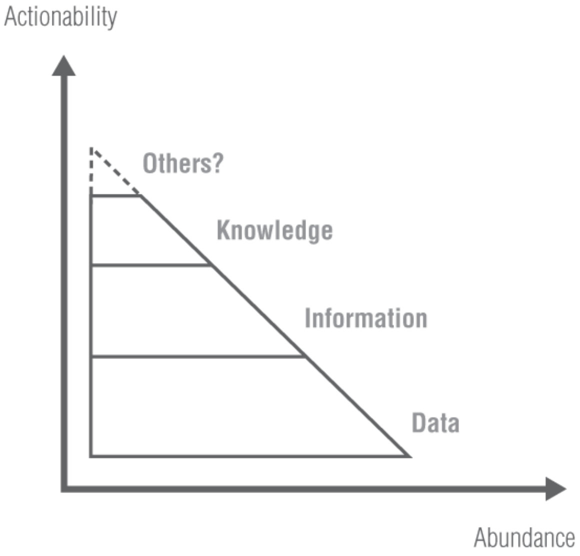


Figure 1. Knowledge Hierarchy (After: Nissen, Kamel, & Sengupta, 2000)

In order to “establish a semantic structure to represent information” some knowledge needs to exist first (Nissen, 2006, p. 20). In order to represent this preexisting knowledge an inverted hierarchy can be used along with the concept of directionality. As shown in Figure 2, existing knowledge can be used to create a structure for information and data that is later transmitted through signals. When the signals are interoperated, data can be collected and processed into information and eventually become knowledge.

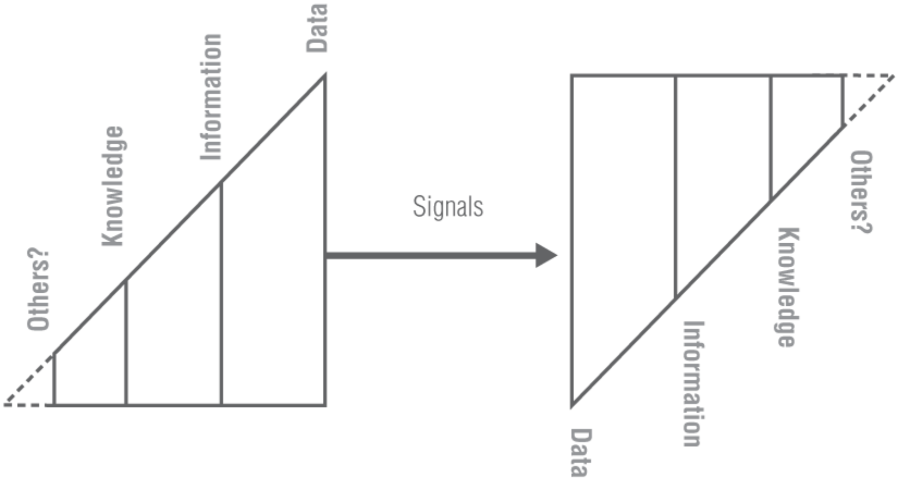
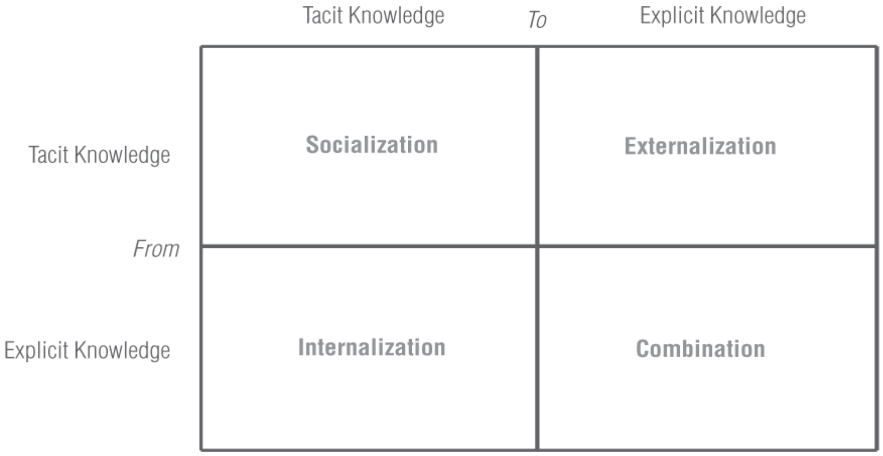


Figure 2. Knowledge Flow Directionality (After: Nissen, 2002a)

It is important to note that there are different types of knowledge, including tacit knowledge and explicit knowledge. Tacit knowledge "is subconsciously understood and applied, difficult to articulate, developed from direct experience and action, and usually shared through highly interactive conversion, storytelling and shared conversation" (Zach, 1999, p. 46). Explicit knowledge, on the other hand, "has been articulated through words, diagrams, formulae, computer programs, and like means" (Nissen, 2006, p. 247). Tacit knowledge is often "more difficult to mobilize in the activities of the organization, possibly resulting in difficulties in value creation" (Silvi & Cuganesan, 2006, p. 312). This is an important concept because different types of knowledge have different properties and they need to be treated differently. For example, tacit knowledge is more appropriable (e.g., can be captured for exclusive use and productivity) than explicit knowledge and thus offers more competitive advantage (Salviotti, 1998). Although tacit knowledge offers more competitive advantage, it is also more "sticky" and clumps in individuals or units, making it more difficult to move or transfer when compared to explicit knowledge (Nissen, 2006). When an organization needs to move knowledge quickly, it becomes important for knowledge workers to concentrate their efforts on explicit knowledge (e.g., IT support enables large amounts of information and data to be organized, aggregated, and disseminated broadly and quickly). When an organization needs greater appropriability for competitive advantage, it is important for knowledge workers to concentrate their efforts on tacit knowledge (e.g., knowledge that resides in the minds of employees). Knowledge workers therefore need to address the flow of knowledge and determine where and when knowledge is needed in order to determine the best course of action.

D. KNOWLEDGE FLOW

A knowledge flow represents the movement of knowledge across people, organizations, places and times. In Figure 3, the four modes of knowledge creation are shown: (1) from tacit to tacit, (2) from explicit to explicit, (3) from tacit to explicit, and (4) from explicit to tacit. The creation of knowledge can be accomplished through socialization, externalization, internalization and combination (Nonaka, 1994). Socialization allows tacit knowledge conversion and flow through the interaction of individuals. It does not necessarily require language as it can occur through "observation, imitation and practice" similar to what happens during on-the-job training (OJT) (Nonaka, 1994, p. 19). Combination allows the creation of explicit knowledge from other explicit knowledge. This can occur from reconfiguring existing information through "sorting, adding, recategorizing, and recontextualizing" (Nonaka, 1994, p. 19). Externalization is the conversion of tacit knowledge into explicit knowledge. Finally, internalization is the conversion of explicit knowledge into tact knowledge.



	Tacit Knowledge	Explicit Knowledge
Tacit Knowledge	Socialization	Externalization
Explicit Knowledge	Internalization	Combination

Figure 3. Modes of the Knowledge Creation (From: Nonaka, 1994)

Nonaka (1994) expanded upon the modes of knowledge creation and described a “spiral model” that defines the interactions between tacit and explicit knowledge. The spiral model shows how the knowledge creation process “can be viewed as an upward spiral process, starting at the individual level moving up to the collective (group) level, and then to the organizational level, sometimes reaching out to the interorganizational level” (Nonaka, 1994, p. 20). This model explains how knowledge is created by building both tacit and explicit knowledge, and how internalization and externalization are interchanged.

Nissen (2002, 2005, 2006) extended Nonaka’s model from two dimensions to four dimensions in order to better visualize dynamic knowledge flows. In Figure 4, the vertical axis represents explicitness and builds upon the Spiral Model (Nonaka, 1994) to measure the amount of tacit or explicit knowledge. The horizontal axis represents reach to measure the amount of social interaction; this axis builds upon the Spiral Model categories of socialization (individual, group, organization, and inter-organization). The third axis represents life cycle or the type of activity associated with the knowledge flows (create, organize, formalize, share, apply and refine). The axes are combined to show a three-dimensional space. Flow time represents the fourth dimension and the amount of time for knowledge to move from one coordinate to another. The amount of time is represented by the thickness of the arrows; for example, thick arrows represent long flow times. Starting at point A, tacit knowledge is created or learned by the individual. The flow from A to B represents socialization and tacit knowledge moves across the reach axis to the group category. The flow from B to C represents externalization and shows the movement of knowledge from tacit to explicit at the group level. The flow from C to D represents combination and shows the movement of explicit knowledge from the group level to the organizational level; this flow has a thinner arrow and represents how explicit knowledge moves faster than tacit knowledge. The flow from D to E represents internalization and shows the movement of knowledge from explicit to tacit form at the organizational level. Finally, the flow from E to B completes the loop and represents socialization, showing the movement of tacit knowledge moving from the organization level to the group level.

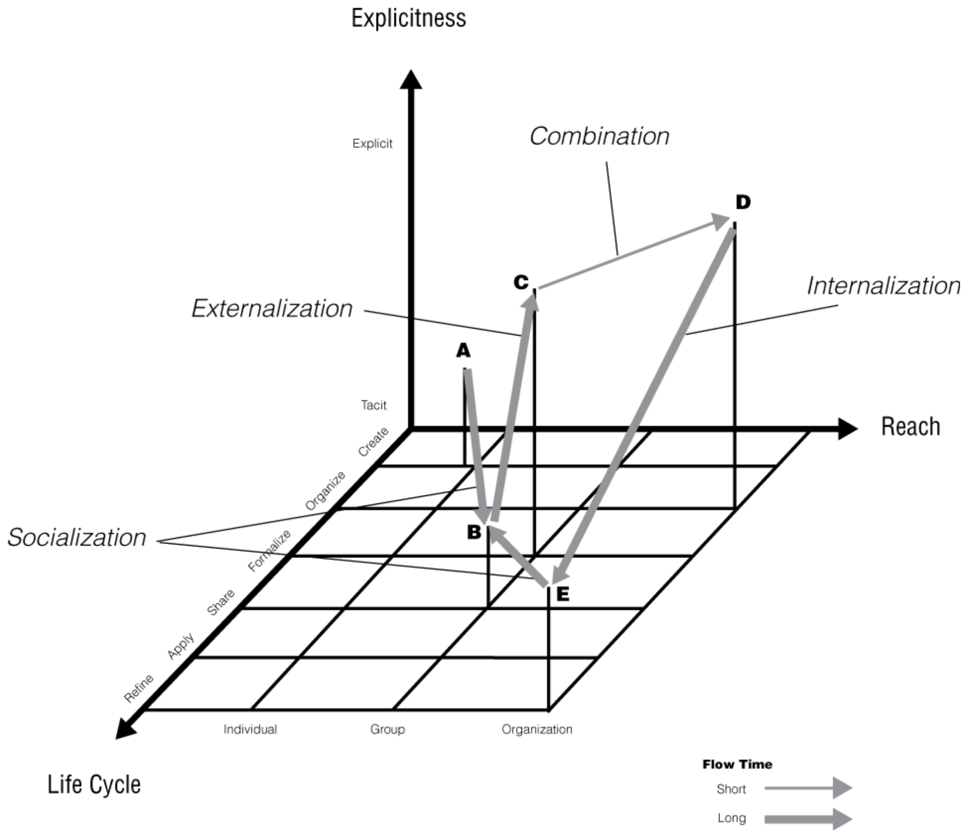


Figure 4. Multidimensional Knowledge-Flow Visualization (After: Nissen, 2005)

E. KNOWLEDGE AND INFORMATION TECHNOLOGY

Moore’s Law describes the long-term trend in the ability to double transistor capacity in computers approximately every 18 months. Similar improvements in communication, networking, and the Internet have allowed organizations to transform their supply chains, manufacturing processes, and distribution systems to save time and money. Reductions in the cost of obtaining, processing, and transforming information have changed the way organizations do business (Porter & Millar, 1985); this trend will continue into the foreseeable future. As a result of these large cost savings and productivity gains, IT is often used as a foundation of many KM projects.

IT (including storage technology, database management systems and query languages) is very good at “organizing, storing, manipulating, and facilitating the query and retrieval of data and documents” but it does not enable knowledge flows (Nissen, 2006, p. 51). In order to enable knowledge flows, IT needs “to be placed in context and used to enable direct action to become knowledge” (Nissen, 2006, p. 51). IT is an aid that can “systematize, enhance, and expedite large scale intra and inter-firm knowledge management” (Alavi & Leidner, 2001, p. 108). KM projects require IT, processes and people, with the latter being the most important. Many KM projects have failed because of over-reliance on IT. People are still required to read and understand documents and perform the majority of workflows requiring knowledge, particularly “those involving experience, judgment and like capabilities dependent on tacit knowledge” (Nissen, 2006, 50). IT plays an important supportive role, but it is the people in an organization who perform the majority of workflows.

The knowledge life cycle (Figure 5) shows where IT is best used. The knowledge life cycle has two different classes of knowledge: the localized view and the expanded view. The localized view includes knowledge organization, formalization, and sharing. These concepts are supported well by IT because the knowledge exists in explicit form. The expanded view includes knowledge creation, refinement and application. These are supported well by people and are not supported as well by IT. The steps in the life cycle do not need to happen in order, nor do they need to be unidirectional.

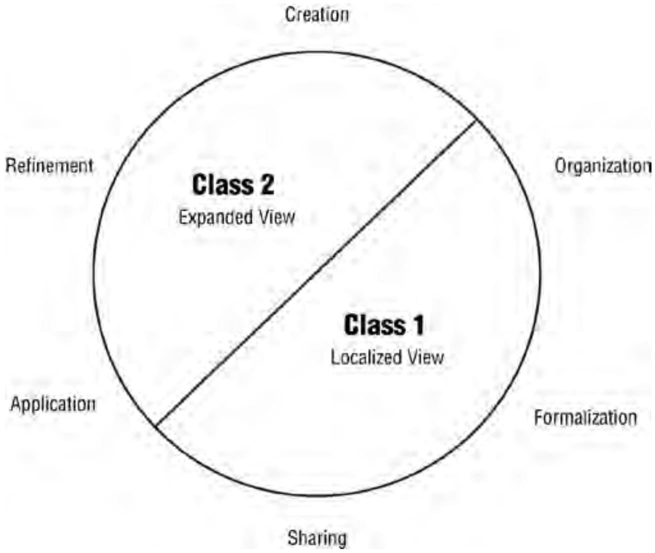


Figure 5. Knowledge Life Cycle (After: Nissen et al., 2000)

F. KNOWLEDGE MANAGEMENT

KM is the process of managing the intellectual capital of an organization (Silvi & Cuganesan, 2006, p. 310). Alavi and Leidner (2001) defined KM as “identifying and leveraging the collective knowledge in an organization to help the organization compete” (p. 113). Tirpak (2005) described KM as “the integration of people, processes, tools and strategy, to create, use and share knowledge, to accomplish an organization’s goals” (p. 15). Within the DoD, KM is defined at the DAU as “the process for effectively applying intellectual capital (human, social, and organizational) to enable faster, better organizational decisions” (Pollock, 2002). Managing intellectual capital can present a number of challenges, but if implemented properly can improve the organizational performance and competitive advantage by reducing costs and increasing profits. IT plays an important role in KM projects, but projects require more than just IT. Some preconditions for success include a requirement for senior management commitment, realistic expectations and empowered workers.

Alavi and Leidner (2001) identified the four basic IT processes of KM as knowledge creation, knowledge storage/retrieval, knowledge transfer, and knowledge application, shown in Table 1. The table also describes the role that IT plays in each process.

Table 1. Knowledge Management Processes and the Potential Role of IT (From: Alavi & Leidner, 2001)

KM Process	Knowledge Creation	Knowledge Storage / Retrieval	Knowledge Transfer	Knowledge Application
Supporting IT	Data mining Learning tools	Electronic bulletin boards Knowledge repositories Databases	Electronic bulletin boards Discussion forums Knowledge directories	Expert Systems Workflow Systems
IT Enables	Combining new sources of knowledge Just in time learning	Support of individual and organizational memory Inter-group knowledge access	More extensive internal network More communication channels available Faster access to knowledge sources	Knowledge can be applied in many locations More rapid application of new knowledge through workflow automation

1. Knowledge Creation

Pentland (1995) described knowledge creation as “developing new content or replacing existing content within the organization’s tacit and explicit knowledge” (Alavi & Leidner, 2001, p. 116). Knowledge can be created and shared through learning, reflection and social means. This is consistent with Nonaka’s four modes of knowledge creation: externalization, socialization, internalization, and combination. In Figure 6, the modes of knowledge creation are represented. Each arrow represents a form of knowledge creation and shows the relationships of knowledge creation between two individuals. IT systems that aid in knowledge creation include: data warehousing, document repositories, electronic mail, group support systems and the Internet (Alavi & Leidner, 2001).

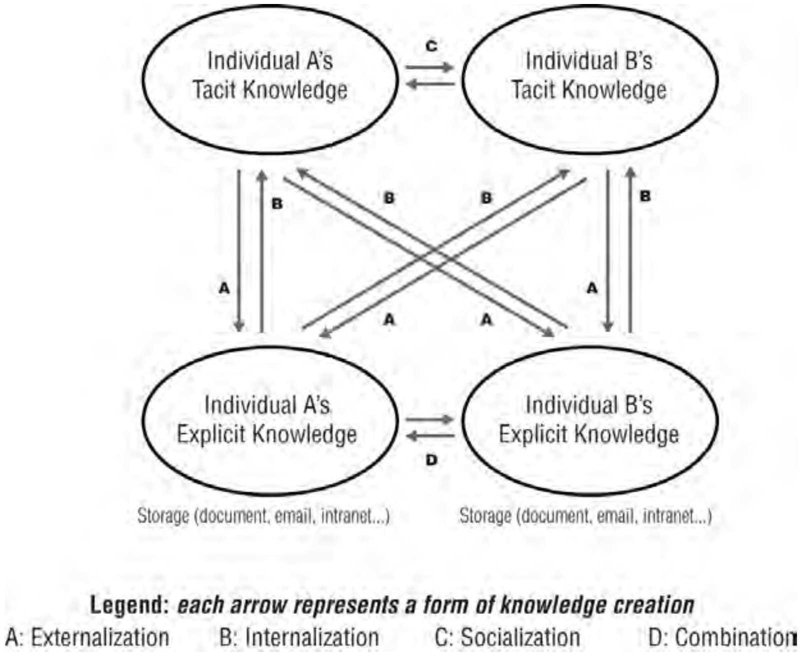

Individual A's Tacit Knowledge
C
Individual B's Tacit Knowledge
B
B
B
B
A
A
A
A
Individual A's Explicit Knowledge
D
Individual B's Explicit Knowledge
Storage (document, email, intranet...)
Storage (document, email, intranet...)
Legend: each arrow represents a form of knowledge creation
A: Externalization
B: Internalization
C: Socialization
D: Combination

Figure 6. Knowledge Creation Modes (From: Alavi & Leidner, 2001)

2. Knowledge Storage/Retrieval

The storage and retrieval ability of an organization is often referred to as organizational memory. Organizational memory includes “memory residing in various component forms, including written documentation, structured information stored in electronic databases, codified human knowledge stored in expert systems, documented organizational procedures and processes and tacit knowledge acquired by individuals and networks of individuals” (Tan et al., 1999; Alavi & Leidner, 2001, p. 118). IT systems that aid in organizational memory include storage technology, database management systems and query languages. These tools are able to increase the speed at which organizational memory can be accessed and help prevent organizational memory loss.

3. Knowledge Transfer

Knowledge transfer can occur at several levels: between individuals, from individuals to explicit sources, from individuals to groups, between groups, across groups and from the group to the organization (shown in Figure 7) (Alavi & Leidner, 2001). Knowledge transfers can be explained in terms of knowledge flows. Knowledge flows can be used to describe, explain, and predict the dynamics of knowledge (Nissen, 2006). IT aids in the transfer process and knowledge flows by increasing an individual’s network size to include a larger set of connections (a larger and more diverse social network) that is able to expose individuals to new ideas. IT systems that enable knowledge transfer include bulletin boards and discussion groups.

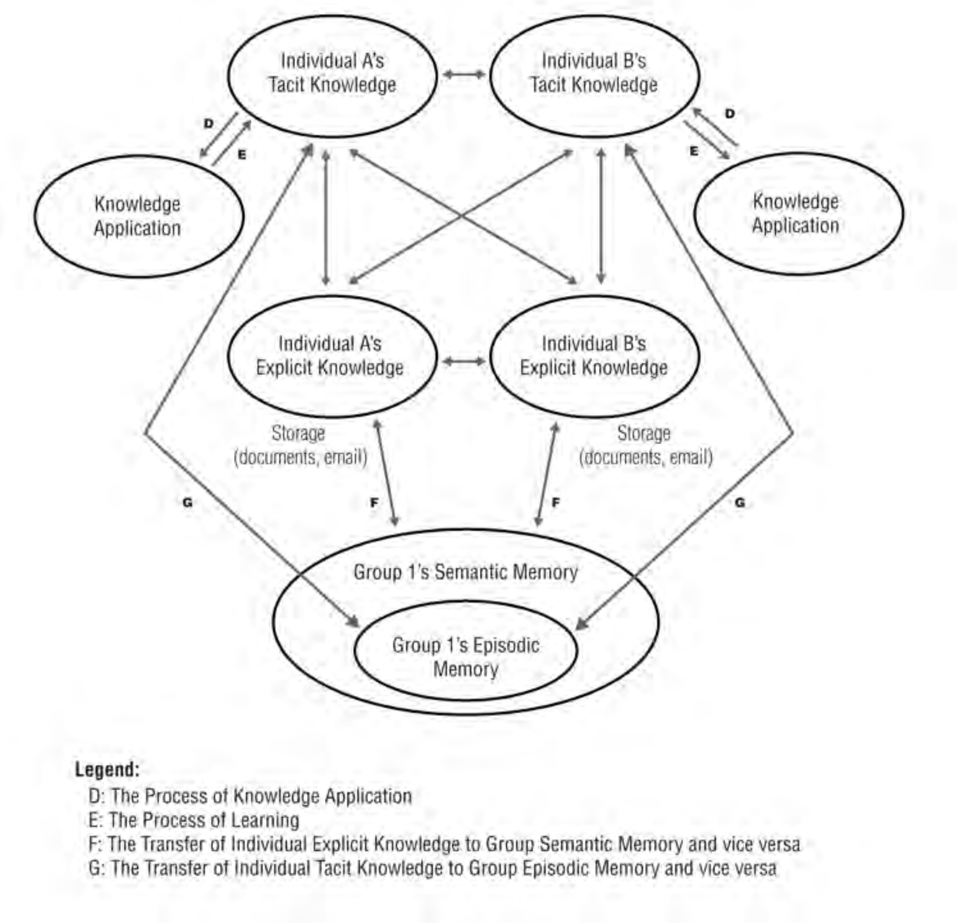

Individual A's Tacit Knowledge
Individual B's Tacit Knowledge
D
E
Knowledge Application
D
E
Knowledge Application
Individual A's Explicit Knowledge
Individual B's Explicit Knowledge
Storage (documents, email)
Storage (documents, email)
G
F
F
G
Group 1's Semantic Memory
Group 1's Episodic Memory
Legend:
D: The Process of Knowledge Application
E: The Process of Learning
F: The Transfer of Individual Explicit Knowledge to Group Semantic Memory and vice versa
G: The Transfer of Individual Tacit Knowledge to Group Episodic Memory and vice versa

Figure 7. Knowledge Transfer among Individuals in a Group (From: Alavi & Leidner, 2001)

4. Knowledge Application

Knowledge application is applying existing knowledge to work and decision-making. It leads to enhanced decision making and competitive advantages. IT can enhance knowledge application through embedding knowledge, codifying routines and automating organization routines. Examples of this include capturing, updating and making directives available, as well as increasing the efficiency of organization routines.

Nissen (2006) described seven preconditions that are key to KM implementation, shown in Table 2. The top three obstacles to implement change include: “(1) lack of sustained management commitment and leadership; (2) unrealistic scope and expectations; and (3) resistance to change” (Nissen, 2006, p. 94).

Table 2. Preconditions for KM success (After: Bashein et al., 1994; Nissen, 2006)

Precondition	KM Implication
1. Senior management commitment	Change of any magnitude requires commitment by senior managers. KM should be considered change of substantial magnitude.
2. Realistic expectations	Expecting too much, too fast, can deflate support for change. Change takes time to implement and refine in KM as in other areas.
3. Empowered and collaborative workers	People doing organizational work are the ones who will make KM effective or not. Knowledge workers need some empowerment for exploration and learning, not just exploitation and doing.
4. Strategic context of growth and expansion	Enthusiasm and optimism can pervade a change project and contribute toward its success, whereas negativity and pessimism can kill it. Setting goals for growth and expansion, through sustained competitive advantage, can facilitate KM change.
5. Shared vision	A vision of how knowledge flows can be enhanced must be conceived and shared broadly in order for empowered people to understand how to change.
6. Sound management processes	The better organized an enterprise is to begin with, the better its chances for successful change via KM.
7. Appropriate people participating full-time	Successful change requires talented people devoting their attention and effort toward enhancing knowledge flows. Assigning slack, part-time resources is unlikely to produce successful KM change.
8. Sufficient budget	Successful change costs money and requires time. Competitive advantage enabled by knowledge is not free. The KM budget should reflect this reality.

From 2000 to 2010, global Internet usage increased from 360 million to over 2 billion people (Farrell & Hutton, 2011). As Internet usage continues to expand, search will become increasingly important to people seeking access to information and data. The role search engines play in explicit knowledge organization, formalization and sharing will continue to grow. However, the use of search in knowledge management systems is not well defined in the literature. Nor is it well utilized in many government knowledge depositories. Finding published government reports through commercial search engines can be difficult, even when the full title of a report is known. A case study assessment at NPS and DTIC will identify areas where search engine integration at these two institutions can be improved, resulting in a positive impact in organizational memory.

III. METHODOLOGY

The purpose of this study is to provide a comparative assessment of KM programs at DTIC and NPS and to develop useful recommendations that can be applied generally. A case study design is used to examine each program. The focus is on the management of explicit knowledge, concentrating in particular on electronic documents that are stored, organized and made accessible through knowledge organizations. A discussion of the organizations, case study design, data collection process, and methods taken to establish quality are presented in this chapter.

DTIC collects and distributes authoritative Department of Defense (DoD) scientific research and engineering information to the defense community. The collection includes work from several U.S. government agencies (e.g., DoD, DoE, NASA) and military research received from a few foreign governments. As part of the acquisition, technology, and logistics community, DTIC is the DoD's primary information manager and provides:

> A wide range of data and information products on policy, scientific and technical planning, budget, research and development (R&D) descriptions, management, test and evaluation, research results, training, law, command histories, conference proceedings, DoD directives and instructions, foreign documents and translations, journal articles, security classification guides, technical reports, and summaries of works in progress. (Schwalb, 2005)

Benefits of the DTIC collection include preventing unnecessary or redundant research, getting scientific and technical information into the hands of the right people, and enabling the conversion of completed research into the production of mature technology to support the warfighter (Ryan, 2009). The DTIC study focused on the KM programs associated with unclassified/unlimited distribution technical reports published through the http://www.dtic.mil website.

As a naval/defense-oriented research university, NPS operates as a geographically distributed educational system that provides a broad range of high-quality graduate education in support of national and international security (NPS Strategic Plan, 2008). The NPS mission is to provide "high-quality, relevant and unique advanced education and research programs that increase the combat effectiveness of the Naval Services, other Armed Forces of the U.S. and our partners, to enhance our national security" (NPS Website, 2011). Over the course of the university's 100-year history, the NPS has established a superior level of academic excellence. Throughout the institution's four schools, its comprehensive institutes and several interdisciplinary centers and research groups are a wide breadth of relevant disciplines tailored to the direct needs of national and global security.

As America's national security research university, NPS delivers relevant and vital answers to real military, national security, and defense problems. This research is chronicled in the school's scholarly publications and degree-related products. Electronic versions of these papers—including NPS papers, NPS journals, proceedings, technical reports, dissertations, joint applied projects, MBA professional reports and theses are stored in the NPS Digital Archive. This archive is a tremendous knowledge resource representing the expertise of this community.

BOSUN is the Dudley Knox Library's online catalog. It acts as a KP, linking the publications and multimedia materials available in the Dudley Knox collections, including the Digital Archive, as well as providing access to external resources such as electronic journals and databases. This NPS study focused on the KM programs associated with documents published in the Dudley Knox Library's BOSUN catalog and the Digital Archive website.

Effective KM demands that existing knowledge be shared. KP facilitates this required sharing by providing simple single-point access to vast and diverse intellectual resources. Both BOSUN and DTIC act as KPs for their communities. This case study explores the possibility of expanding those communities by making these KPs more search engine accessible.

A. CASE STUDY RESEARCH

A case study is "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" (Yin, 2009, p. 18). Yin (2009) recommended using a case study when “how” or “why” questions are being proposed, the investigator has little control over events, and the focus is on contemporary phenomenon. This assessment involved analyzing KM programs (a contemporary phenomenon) that the DoD uses to manage intellectual capital. The KM programs were analyzed in their current state and there was no possibility of manipulation from outside forces. In addition to the “how” and “why” questions normally used in case studies, “what” questions were appropriate as well (Yin, 2009).

B. COMPONENTS OF CASE STUDY

The five components of case study research are: research questions, propositions, unit of analysis, logic linking data to the propositions, and criteria for interpreting the findings (Yin, 2009). A narrative of each component and how it relates to this study are provided below.

1. Research Questions

The case study method normally uses “how” and “why” and occasionally “what” questions in order to clarify the nature of the study precisely (Yin, 2009). The two investigative questions used in this study are stated below.

> RQ 1. How well are NPS and DTIC harnessing the power of commercial search engines?
>
> RQ 2. How can the DoD increase explicit knowledge presence through search engines?

2. Propositions

A proposition is a statement that “directs attention to something that should be examined within the scope of study” (Yin, 2009, p. 28). The proposition statement is used to identify the scope and establish boundaries. The propositions used in this research are: Proposition 1: KM is important to the DoD. Proposition 2: Each organization has a different approach to KM and can be compared through similarities and differences.

Proposition 3: KM practice in the DoD can be improved by identifying critical explicit knowledge factors. These propositions helped identify what was addressed in this research.

3. Unit of Analysis

The unit of analysis is related to "defining what the case is" (Yin, 2009, p. 29). In this case study, the units of analysis are the KM programs at NPS and DTIC. In order to compare and contrast the programs, each case was studied individually. The unit of analysis narrowed the amount of relevant data and ensured the case study stayed within feasible limits.

4. Data Collection

The sources of evidence in a case study can be quite extensive and common items include documentation, archival records, interviews, direct observations, participant-observation, and physical artifacts (Yin, 2009). Data was collected about the KM programs in each organization (DTIC and NPS). Data collection included gathering documentation, archival records and exploring the capabilities of KM portals. In order to limit the scope of the data collection, the collection of archival records was limited to 10,000 documents at NPS and 250,000 documents at DTIC.

5. Data Analysis

Yin (2009) explained several ways of linking data to propositions; these include "pattern matching, explanation building, time-series analysis, logic models, and cross-case synthesis" (p. 34). Pattern matching is a technique for linking data to the propositions, where several pieces of information from the case may be related to the proposition. This case study accomplishes data analysis through supporting or refuting the propositions with the pattern matching technique. This method shows where the organizations share similarities or diverge in KM approaches and allow conclusions to be drawn about the data.

C. QUALITY AND RELIABILITY METRICS

It is important for case study research to establish quality through validity and reliability. The following tests were performed to check for design quality: construct validity, internal validity, external validity, and reliability (Yin, 2009).

Table 3. Case Study Tactics for Four Design Tests (After: Yin, 2009)

Tests	Case Study Tactic	Phase in Research
Construct Validity	• Use multiple sources of evidence • Establish chain of evidence • Have key informants review draft of case study report	• Data collection • Data collection • Composition
Internal Validity	• Do pattern matching • Do explanation building • Address rival explanations • Use logic models	• Data analysis • Data analysis • Data analysis • Data analysis
External Validity	• Use theory in single-case studies • Use replication logic in multiple-case studies	• Research design • Research design
Reliability	• Use case study protocol • Develop case study database	• Data collection • Data collection

1. Construct Validity

Construct validity "is used to identify the correct operational measures for the concepts being studied" (Yin, 2009, p. 40). The primary objective of this research is to provide a comparative assessment of KM programs at NPS and DTIC. Multiple sources of evidence are used in order to establish good construct validity. Three sources of evidence were collected, including: documentation, archival records and KM portals. The sources for all data collected are listed in the appendices to provide a clear chain of evidence.

2. Internal Validity

Internal validity "seeks to establish a causal relationship, whereby certain conditions are believed to lead to other conditions, as distinguished from spurious relationships" (Yin, 2009, p. 40). This case study uses pattern matching and explanation building, and addresses rival explanations, to improve internal validity.

3. External Validity

External validity is used to determine "whether a case study's findings are generalizable beyond the immediate case study" (Yin, 2009, p. 43). This case study used replication logic in two case studies. This ensures the results obtained from NPS and DTIC can be applied more generally. Each case study was evaluated for the same elements and compared to identify patterns.

4. Reliability

The goal of reliability is to minimize the errors and biases in a study (Yin, 2009). This can be accomplished by demonstrating that the operations of a study can be repeated. Yin (2009) stated that research is reliable if another investigator can use the same procedures on the same case and arrive on the same findings and conclusions. Data collection procedures were documented through a case study database in order to maintain a chain of evidence.

D. CASE STUDY LIMITATIONS AND BIASES

Yin (2009) stated it is important to understand and openly acknowledge the limitations of case study research and to reduce and control for bias. This study admits bias on the part of the investigator as a member of the Navy. By relating preliminary findings to several critical colleagues the data collection and analysis phases, this bias was reduced as much as possible. DTIC's collection is very large and includes more than two million Technical Reports (Ryan, 2009). As a result of the large collection size and the limitation of having one researcher the scope of analysis was limited. The documents, tools and websites evaluated were limited to publically available sources. The data collection process concentrated on explicit knowledge sources and as a result no data was collected from interviews from current DoD KM practitioners. KM is an immature discipline in the DoD as a result each service has different definitions, expectations, and levels of experience with KM. In order to mitigate these challenges the data collection focused on gathering recent documentation and archival records as well as exploring the new capabilities of KM portals through the publically available DTIC and NPS collections.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS

Contrary to systems of the past, today's search engines work hard to adapt to the user rather than asking the user to adapt to them. They have traveled light years past just locating keywords. Today, Google uses more than 200 signals including contextual clues, bigram analysis and a user's personal search history and location to find and rank query results (Levy, 2010). The hundreds of millions of people that use Google are constantly helping the company develop new signals. By monitoring their user's searches, how they modify their queries and which results they click on, Google can move closer to understanding not what their users say but what they mean:

> This is the hard-won realization from inside the Google search engine, culled from the data generated by billions of searches: a rock is a rock. It is also a stone, and it could be a boulder. Spell it "rokc" and it's still a rock. But put "little" in front of it and it's the capital of Arkansas. Which is not an ark. Unless Noah is around. The holy grail of search is to understand what the user wants," Singhal says. "Then you are not matching words; you are actually trying to match meaning." (Levy, 2010)

Given this ever-increasing ease of use, the breadth of the information available and the unsurpassed speed with which queries are answered, the number of knowledge seekers turning to commercial search engines first for their information needs will continue to grow. Repositories must become accessible or risk being ignored.

NPS Digital Archive is a well-controlled collection of relatively homogenous documents. As such, using the BOSUN catalog search function, documents are easily found—as long as the user knows what he is looking for. However, if a knowledge seeker does not search the website using the BOSUN catalog, is unsure what he is searching for or misspells the search term, the information located there is very difficult to find. In effect, the current organization of this important document repository and its existing KP allow for knowledge transfer and transmission within the community, but it exists as a largely unknown stockpile to the global community of knowledge seekers.

DTIC is a less homogenous collection of documents and its KP user interface is more search engine like (more forgiving of misspellings and poorly selected queries), but it too hinders commercial search engines from effectively indexing its pages. Making both of these repositories more commercial search engine-accessible is a low-cost, high-impact method of supporting the knowledge flow that will contribute to innovation.

The old-school method of search involved the use of directories. Directories are subject-tree style catalogs that organize the information into topics. Of the three big search engines, only Yahoo still operates a directory structure (although web search is always available). Using the Yahoo directory, a user clicks on a subject (e.g., *Health*) and is presented with a sub directory of common topics. Here, the user chooses *Diabetes* and is given basic information and further sub-topics. On choosing *Diabetes Mellitus, Type 2*, further subtopics are presented and the user chooses *Prevention*. Here, hopefully, the user finds the information they are seeking.

Today, most people use web search rather than directory browse. Using Google for the same query as above, a user might key *type 2 diabetes prevention* into a search box. Google will return 3,140,000 results in 0.26 seconds. More important than the speed or the volume of those results (which are both very important) is the ranking. Rarely do users venture past the first page of search results. The search engine's goal is to rank the results presented in order of relevance to that particular user. Each search engine company has its own secret formula for results ranking.

The big commercial search engines (Google, Yahoo, Bing) search their own indexes, not the Internet, when queried. A search engine's index is built using bots, also referred to as spiders or crawlers. A bot is an automated software program that collects data from webpages and adds it to the search engine's index. That data is then used in highly proprietary algorithms to attempt to rank the most relevant search results at the top of any results page. An engine's effectiveness is measured by how fast it can return the results the user wants.

When bots crawl a webpage, they are looking for clues that can be used to determine the relevance of a particular page to a given query. One clue is keyword usage. Bots pull and index words that seem to be important. Words used in the title, mentioned near the beginning of the document, used frequently or in headings are given more weight than other words. As they visit a page, the bots copy the content and then follow the links from that page to the pages linked to it where they repeat the process, crawling billions of pages.

For a website, webpage or document, the first steps to being “found” are to allow the bots in and then to feed them as much information as they will accept. To be well ranked, particularly for highly competitive keywords, involves the very complicated and ever changing “science” of search engine optimization (SEO). While SEO is beyond the scope of this discussion, suffice it to say that observing the best practices for bot accessibility provides the necessary foundation for SEO.

For this analysis, a total of eighteen best practices were selected from Google’s Webmaster Guidelines—nine that relate to websites and nine related to webpages (Google Webmaster Tools, 2011). The NPS and DTIC websites are evaluated for compliance to these eighteen best practices and then each site is tested to determine what percentage of the hosted documents can be found in the first page of Google results when searching by document title. Additionally, general observation about the websites and their HTML code are recorded. Finally, a prototype site is built to host NPS and DTIC documents while following the best practices. The prototype site is tested to determine what percentage of the hosted documents can be found in the first page of Google results when searching by document title.

A. BEST PRACTICES

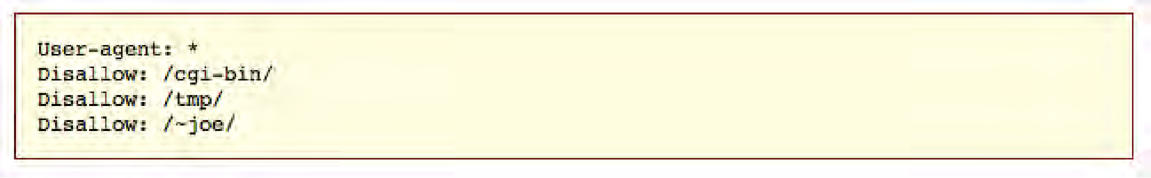
A total of eighteen best practices were selected for this analysis, nine that relate to websites and nine related to webpages.

1. Websites and Search Engines

Search engine bots need to be able to crawl and extract content from webpages. If there are barriers to the bots’ successful crawl, the knowledge will not be findable by a simple query using the search engine. If users cannot find the explicit knowledge they are looking for, an opportunity is wasted. Websites that contain explicit knowledge should be designed to work with search engines. The key website components necessary for successful search engine integration are outlined in the following section.

a. Robots.txt

A robots.txt file (Figure 8), also known as the robots exclusion protocol, is used to block search engines from crawling webpages that do not need to be indexed. Blocking pages that do not need to be indexed helps search engines concentrate on the information that should be indexed. The robots.txt file is placed in the website’s root folder (e.g., example.com/robots.txt) and contains directories that should be ignored. Websites that have multiple domains such as example.com and extra.example.com require their own robots.txt file. The robots.txt file does not guarantee security because it is a voluntary standard. Some search engines support looking for a sitemap.xml link inside the robots.txt file. A robots.txt file helps search engines by reducing the number of pages that they crawl.

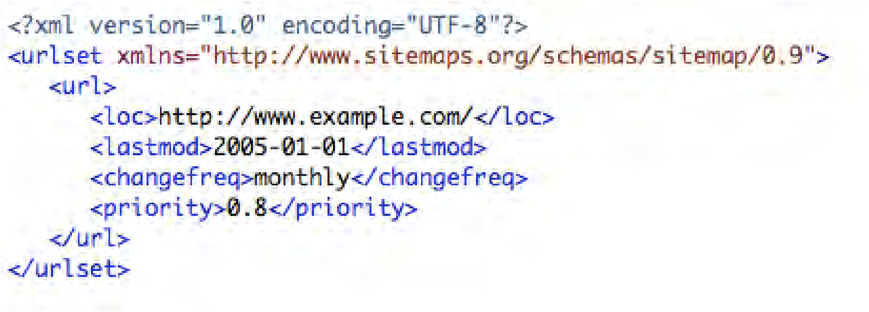


```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tmp/
Disallow: /~joe/
```

Figure 8. Example Robots.txt File

b. Sitemap.xml

A sitemap.xml (Figure 9) is a file that contains a list of all of the URLs on a website that are available for crawling. It helps search engines find pages to index. The file also includes optional information such as when the URL was last modified, how often the URL is updated, and its priority compared to other URLs. The sitemap.xml file can be no larger than 10 Megabyte (MB) and contain no more than 50,000 URLs. A website can have multiple sitemap.xml files.



```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

Figure 9. Example Sitemap.xml File

c. Extractable Text and Links

Most search engines look for extractable text to index and extractable links to follow. Search engines are not able to extract links and text contained in JavaScript, images and Adobe Flash files. If it is not possible to include extractable text, alternative or alt text may be used to describe files. For example, an alt image tag may be used to describe an image.

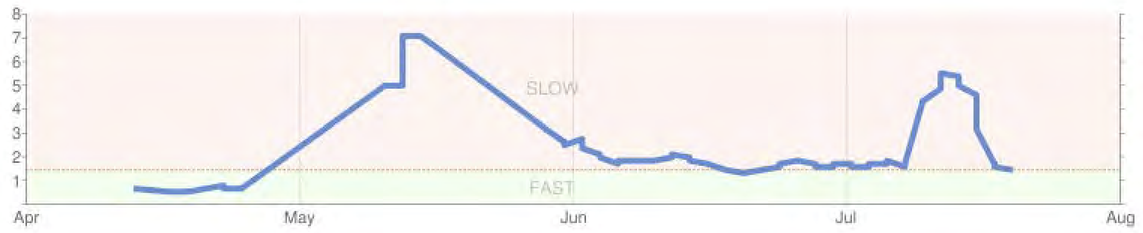
d. Accessible URL Structure

Accessible URLs contain valid characters and valid extensions; they do not include temporary information such as the current time or session number. A good URL structure should be easy for a user to read and share with others. It should include a valid extension when the file type is something other than HTML.

e. Fast Response/Load Times

A search engine measures the response and load times to determine the speed to crawl a website. If a website is slow to respond, most search bots will reduce their crawl rate to avoid overloading a website. On the demonstration website for this report, the Google Search Bot crawled at a rate of 12,000 pages a day with response time of 1.4 seconds per page load (Figure 10). When the server was slowed to a response time of 6 seconds or greater, the Google crawl rate was reduced to 1,500 pages a day. During this slow loading period, other search bots (Yahoo and Bing) stopped crawling.

Figure 10. Google Webmaster Tools (Response Time in Seconds)



f. Stable/Permanent Links

Stable or permanent URLs allow links to work over a long period of time, even if the website design and structure changes. If links are not stable, a search engine is likely to index the same content multiple times, or save URLs to links that have moved. When a user clicks on unstable links, the webserver is likely to deliver “page not found” errors.

g. Duplicate Content

Duplicate content can occur from poorly designed URL structures or saving content to a website multiple times. Duplicate content is undesirable because it reduces reputation and page rank in search results (Google Webmaster Tools, 2011). Duplicate content can be avoided by using canonicalization and server redirects.

h. Canonicalization

Canonicalization is the process of preventing multiple versions of a URL from being indexed. Canonicalization is implemented by picking the best URL for a set of URLs that resolve to the same content. For example, consider the following URLs:

http://example.com/item

http://www.example.com/item

http://www.example.com/index.php?ID=item

To prevent a search engine from indexing each link as a different page, the following code can be placed in the heading of an html file: <link rel=“canonical”

href=“http://www.example.com/item”>. Canonicalization tells search engine bots what URL to save when duplicate content exists on a website.

i. Server Redirects

A server redirect can be used to enforce valid URLs and can automatically direct a browser to a web page that has moved. In order to avoid duplicate pages, a server-side “301 redirect” can be used to send traffic from duplicate URLs to the preferred URL. For example, using a 301 redirect on the “non-www” version of the URL to the “www” version of a URL is a way to “permanently” redirect users to the preferred URL.

2. Webpages and Search Engines

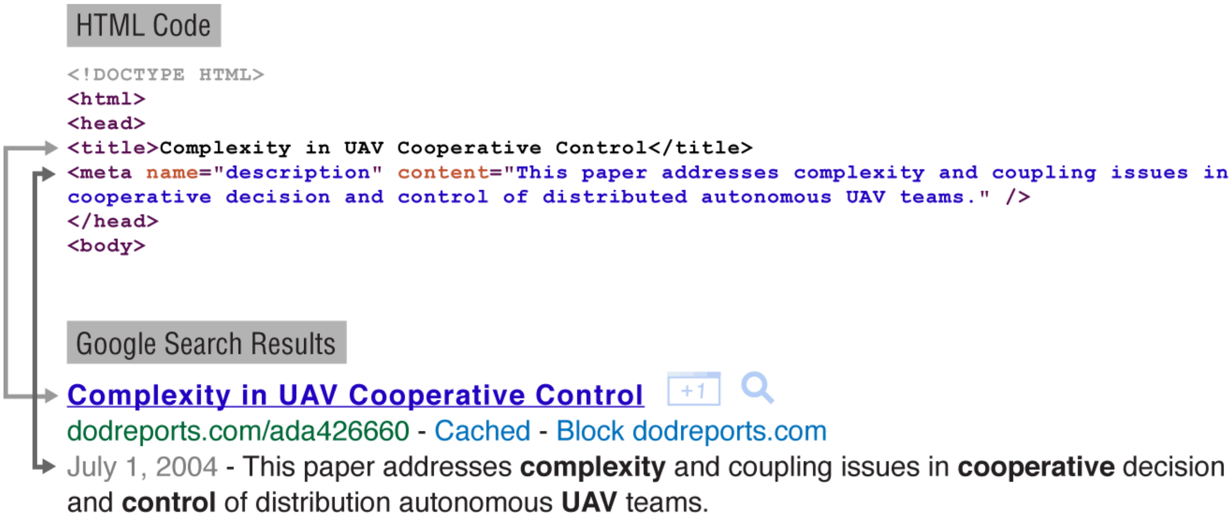
Search engines work best when there is extractable text on standard HTML pages. When a search engine extracts text, it assigns a value based on html tags. For example, a heading tag carries more weight than a paragraph tag. Text that is improperly tagged can create problems for search engines. Important HTML tags include the page title, meta description, headings, paragraphs, tables and alt text.

a. Standard HTML

Standard HTML markup follows World Wide Web Consortium’s (W3C) standards. The W3C standards define how pages should be structured to avoid coding errors. If a webpage contains errors, the page may not be indexed properly. Search engines look for text that is contained in html elements such as headings, paragraphs, lists and tables.

(1) Page Title. The title tag (Figure 11) helps users understand the focus of the page. It is used to describe what the page is about. This text is important because it is generally used as the search link and title.

(2) Meta Description. The meta description element (Figure 11) is a snippet of text that search engines use to display under the page title. The meta description is used to describe, to a user, the content of the page in one or two sentences.

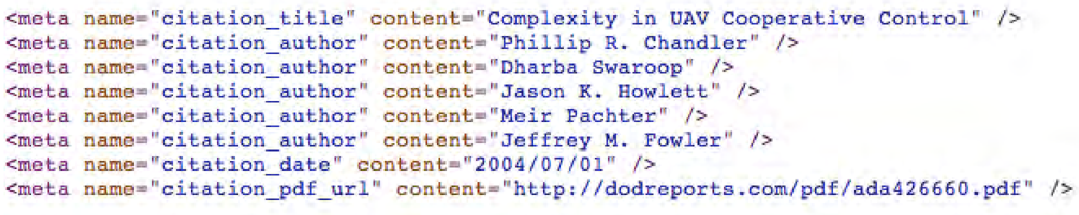


```
<!DOCTYPE HTML>
<html>
<head>
<title>Complexity in UAV Cooperative Control</title>
<meta name="description" content="This paper addresses complexity and coupling issues in
cooperative decision and control of distributed autonomous UAV teams." />
</head>
<body>
```

Complexity in UAV Cooperative Control
dodreports.com/ada426660 - Cached - Block dodreports.com
July 1, 2004 - This paper addresses **complexity** and coupling issues in **cooperative** decision and **control** of distribution autonomous **UAV** teams.

Figure 11. Title and Description Elements and Google Search Results

(3) Meta Tags. HMTL tags (Figure 12) are generally used to describe additional information about a webpage like the author's name, publication data and a link to the PDF version of a page.



```
<meta name="citation_title" content="Complexity in UAV Cooperative Control" />
<meta name="citation_author" content="Phillip R. Chandler" />
<meta name="citation_author" content="Dharba Swaroop" />
<meta name="citation_author" content="Jason K. Howlett" />
<meta name="citation_author" content="Meir Pachter" />
<meta name="citation_author" content="Jeffrey M. Fowler" />
<meta name="citation_date" content="2004/07/01" />
<meta name="citation_pdf_url" content="http://dodreports.com/pdf/ada426660.pdf" />
```

Figure 12. HTML Meta Tags

(4) Headings, Paragraphs, Tables and Alt Image Text. It is important for webpages to contain text that is tagged according to type. This allows search engines to give some text higher priority. Generally, search engines cannot understand images; it is important that if images are used an alternative text be used to describe the image.

b. Use of Internal Links

Internal links allow users and bots to navigate in a website. Users and search engines need internal links to explore other parts of the website. Generally, pages that are linked to multiple times have a greater authority on a website.

B. HOW WELL ARE NPS AND DITC HARNESSING THE POWER OF COMMERCIAL SEARCH ENGINES?

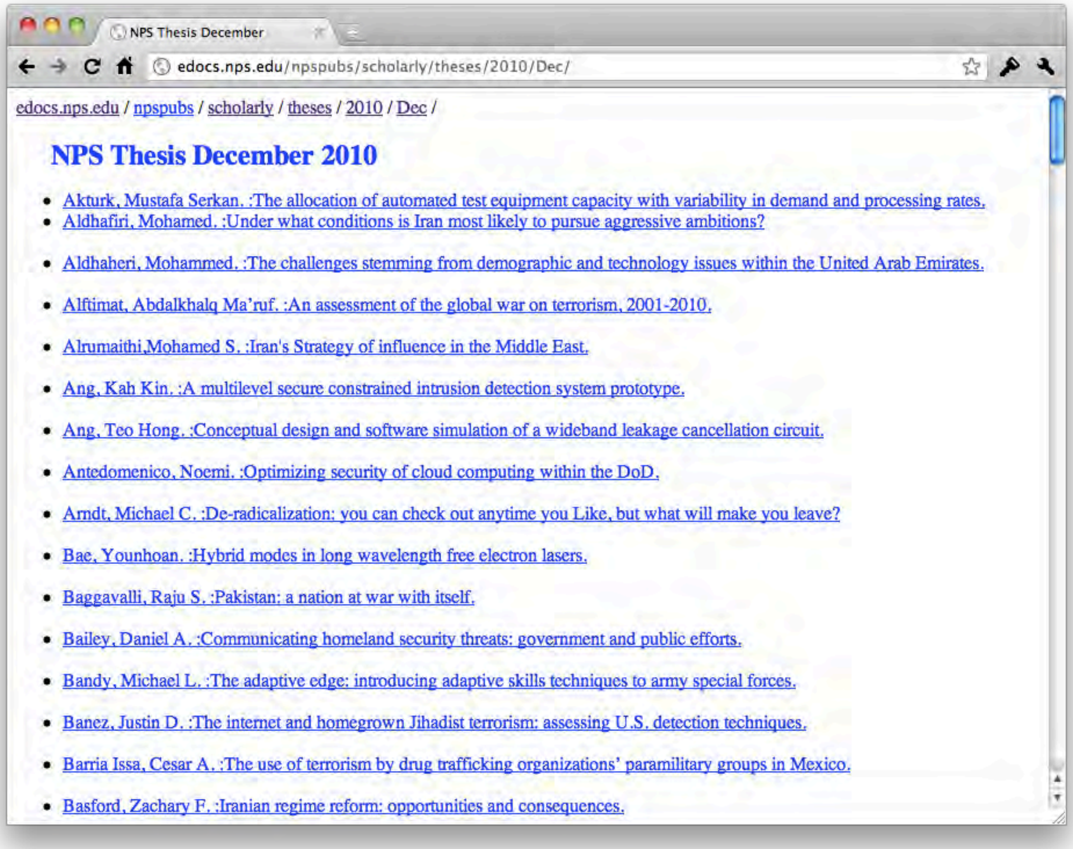
The evaluation at NPS and DTIC was conducted using three tools, a custom search robot, evaluation of title search results at http://google.com, and a demonstration website at http://dodreports.com. The search robot was built to discover webpages by following webpages through links, index pages that contained extractable content, and ranking extracted content based on tags. The evaluation of title search results was conducted on the Google search by searching for the full title of a particular report. If the report title was on the first page or first ten results and had the correct Uniform Resource Locator (URL) to the NPS or DTIC websites, the test was successful.

1. Evaluation of Naval Postgraduate School

The NPS evaluation consisted of gathering documents from the NPS Digital Archive located at http://edocs.nps.edu/ and the NPS BOSUN catalog located at http://bosun.nps.edu/. During the period from October 2010 to December 2010, 10,000 Portable Document Format (PDF) files and 236 webpages were downloaded from the NPS Digital Archive. An attempt was made to automatically download webpages and documents from the NPS BOSUN catalog. However, this effort proved impractical due to the website’s propriety software. Data was manually gathered from the NPS BOSUN catalog by randomly choosing webpages through the internal search interface. During October 2010, 50 Hypertext Markup Language (HTML) files were downloaded from the NPS BOSUN catalog.

In December 2010, a measurement at NPS was conducted to determine how well documents (Figure 13) on the Digital Archive could be found through commercial search engines. This test involved searching for the title of published reports without quotes on the Google search engine. If the report was found in the first page (top 10 links) and linked to the NPS Digital Archive, a passing score was recorded. If the report was not found or linked to a location other than the NPS Digital Archive, a failing score was recorded. Three hundred and seventy randomly selected titles were tested. With a 95% Confidence Level, the test confirmed that the percentage of NPS documents that are found on the first page of Google search results is between 15.3% and 25.3%. This percentage is lower than expected and indicates that search engines cannot easily locate documents on the NPS Digital Archive.

In January 2011, the NPS Digital Archive was evaluated (Figure 13 and Table 4). This evaluation involved gathering general observations about the website and HTML code and an evaluation of key website components. The following are observations found in January 2011: The NPS website was found to primarily store PDF files of NPS publications. The website contains very little text and contains links to publications organized into folders. The webpages are generally easy to navigate, although some search engines will not be able to follow the “back links” at the top of the page because these links are coded in JavaScript. The website is missing several key components including robots.txt, sitemap.xml, server redirects, meta description, html tags, and headings. The links to approximately 1% of the publications do not work; this usually occurs when there is a missing apostrophe in a file name. The HTML files are organized in folders according to publication type, year and month. There were no individual webpages for individual reports.



edocs.nps.edu / npspubs / scholarly / theses / 2010 / Dec /

NPS Thesis December 2010

- Akturk, Mustafa Serkan. :The allocation of automated test equipment capacity with variability in demand and processing rates.
- Aldhafiri, Mohamed. :Under what conditions is Iran most likely to pursue aggressive ambitions?
- Aldhaheri, Mohammed. :The challenges stemming from demographic and technology issues within the United Arab Emirates.
- Alftimat, Abdalkhalq Ma'ruf. :An assessment of the global war on terrorism, 2001-2010.
- Alrumaithi,Mohamed S. :Iran's Strategy of influence in the Middle East.
- Ang, Kah Kin. :A multilevel secure constrained intrusion detection system prototype.
- Ang, Teo Hong. :Conceptual design and software simulation of a wideband leakage cancellation circuit.
- Antedomenico, Noemi. :Optimizing security of cloud computing within the DoD.
- Arndt, Michael C. :De-radicalization: you can check out anytime you Like, but what will make you leave?
- Bae, Younhoan. :Hybrid modes in long wavelength free electron lasers.
- Baggavalli, Raju S. :Pakistan: a nation at war with itself.
- Bailey, Daniel A. :Communicating homeland security threats: government and public efforts.
- Bandy, Michael L. :The adaptive edge: introducing adaptive skills techniques to army special forces.
- Banez, Justin D. :The internet and homegrown Jihadist terrorism: assessing U.S. detection techniques.
- Barria Issa, Cesar A. :The use of terrorism by drug trafficking organizations' paramilitary groups in Mexico.
- Basford, Zachary F. :Iranian regime reform: opportunities and consequences.

Figure 13. NPS Website (From: NPS Digital Archive, 2011)

Table 4. NPS Website Evaluation (From: NPS Digital Archive, 2011)

Item	Passing %	Grade
Website Search Engines		
Robots.txt	0 % [not found]	Needs improvement
Sitemap.xml	0 % [not found]	Needs improvement
Extractable Text and Links	Text: less than 1% Links: 100%	Text: Needs improvement Links: Excellent * Navigation links were inside JavaScript
Accessible URL Structure	99%	Excellent * Special characters were missing in some author's names (these reports could not be downloaded)
Fast response / load times	100%	Excellent
Stable / Permanent Links	100%	Excellent
Server Redirects	[not found]	Needs improvement
Canonicalization	[not found]	Not needed (no duplicate content was found)
Avoiding Duplicate Content	100%	Excellent
Webpage Search Engines		
Standard HTML (W3C Validation)	20% [valid lines/total lines]	Needs improvement
Page Title	100%	Excellent
Meta Description	[not found]	Needs improvement
Tags	[not found]	Needs improvement
Headings	[not found]	Needs improvement
Paragraphs	[not found]	Needs improvement
Alt Image Text	N/A	N/A
Use of Internal Links	100%	Excellent
Clear Site Navigation	100%	Excellent

In January 2011, the NPS BOSUN catalog was evaluated (Figure 14 and Table 5). The NPS BOSUN catalog is used primarily as a research tool at the NPS Library. The BOSUN catalog uses a temporary link structure that cannot be accessed by external search engines and links to pages cannot be shared. In order to enter the website, the internal search engine must be used. The website has a valid robots.txt file but this file references a sitemap.xml that does not exist. A second sitemap.xml.gz exists but the links contained in the file did not work. The website is missing several key components including stable or permanent links, server redirects, canonicalization, unique page titles, meta descriptions, tags and headings. It should be noted that there is little value in updating these key components until the temporary link structure is changed. There were individual webpages for each report and each page contained helpful navigation links to the author and subject area. The website contained a helpful “ask a librarian” tool (Figure 14) that allows users to get in contact with a human to find a resource.

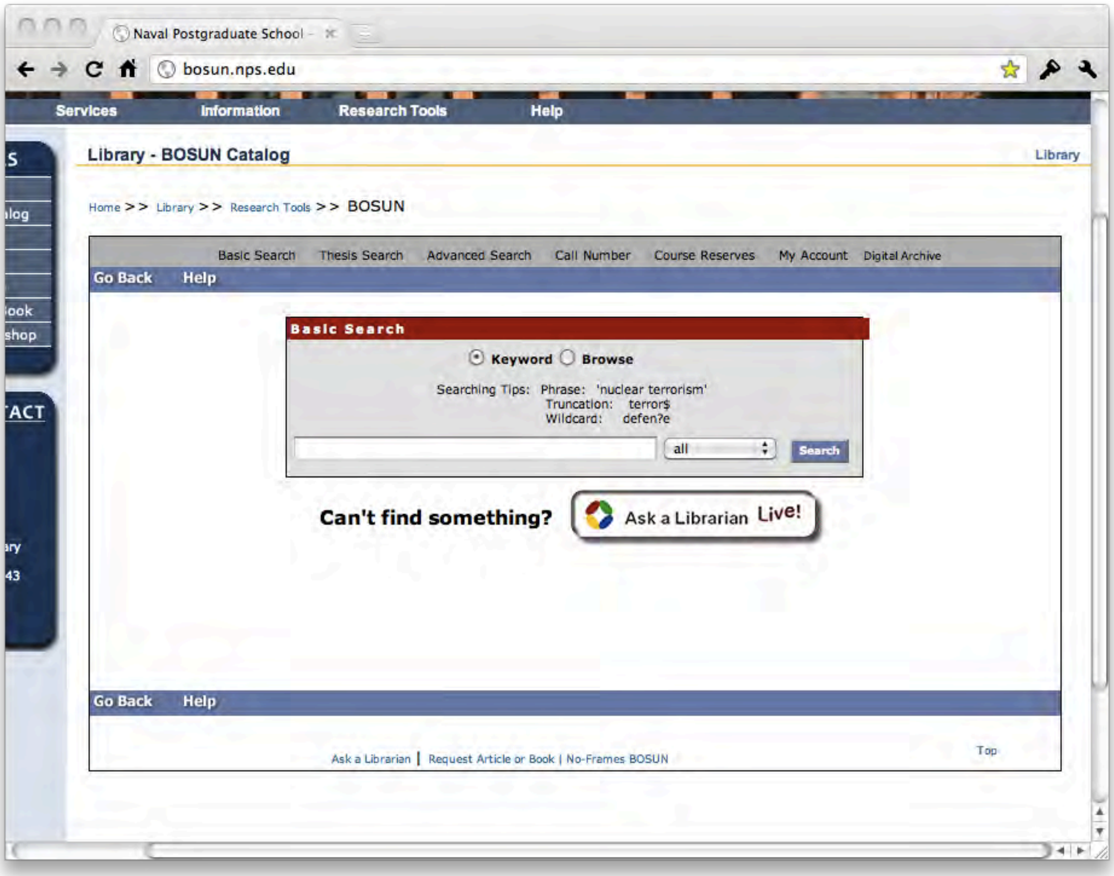
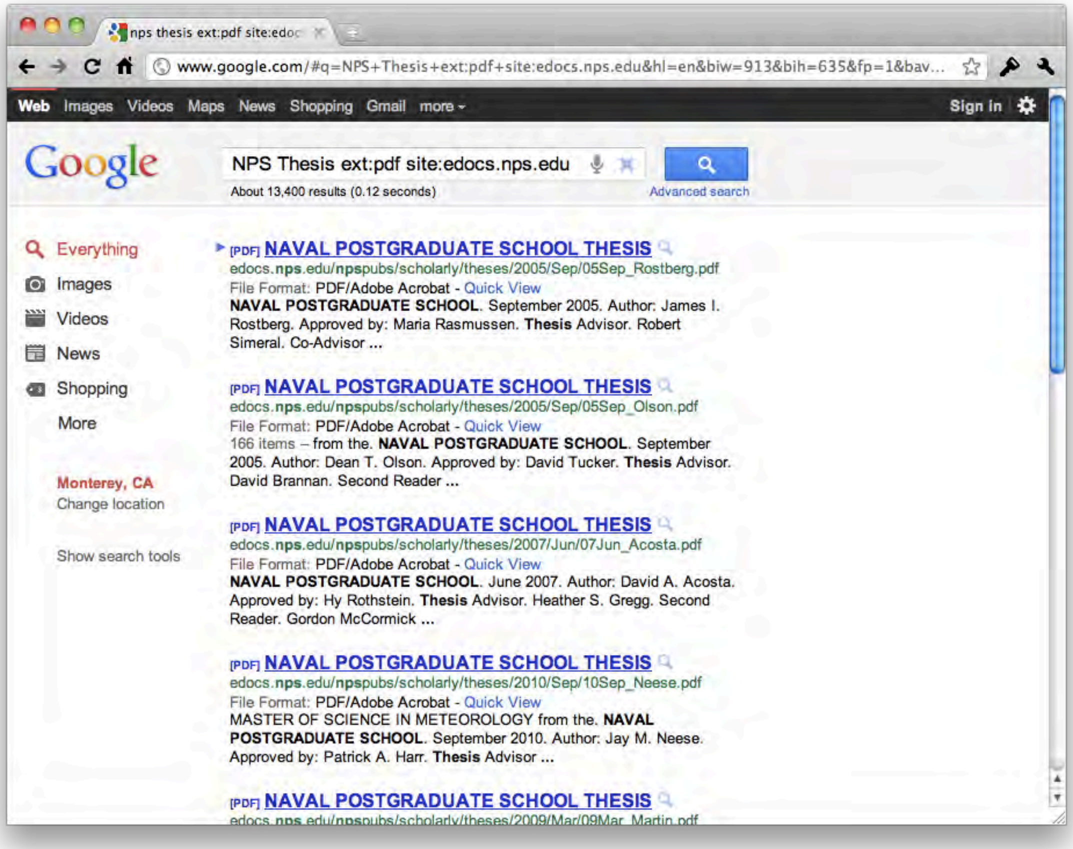


Figure 14. BOSUN Catalog Front Page (From: NPS BOSUN Catalog, 2011)

Table 5. NPS Website Evaluation (From: NPS BOSUN Catalog, 2011)

Item	Passing %	Grade
Website Search Engine		
Robots.txt	80% (4/5) [1 reference to sitemap that does not exist]	Satisfactory
Sitemap.xml	0% [found but does not contain working links]	Needs improvement
Extractable Text and Links	100%	Excellent
Accessible URL Structure	100% * contained within iframe	Excellent
Fast response / load times	100%	Excellent
Stable / Permanent Links	0%	Needs improvement
Server Redirects	[not found]	Needs improvement
Canonicalization	[not found]	Needs improvement
Avoiding Duplicate Content	Unknown	Unable to measure
Webpage Search Engine		
Standard HTML (W3C Validation)	99%	Excellent
Page Title	[not found]	Needs improvement
Meta Description	[not found]	Needs improvement
Tags	[not found]	Needs improvement
Headings	[not found]	Needs improvement
Paragraphs	[not found] * Text is inside table or divs	Needs improvement
Alt Image Text	N/A	N/A
Use of Internal Links	100%	Excellent
Clear Site Navigation	100%	Excellent

In January 2011, the NPS Digital Archive PDF collection located at http://edocs.nps.edu/ was evaluated (Figure 15 and Table 6). Approximately 60% of the PDF files were text based, with the majority of PDF files published before September 2001 being image based. Very few of the PDF files contained a title tag or were optimized for fast web view. The title page on the NPS thesis PDFs is structured with the school name as the largest and uppermost part of the page. With no title tag, most search engines improperly index the School’s name as the report title, shown in Figure 15.



NPS Thesis ext:pdf site:edocs.nps.edu

About 13,400 results (0.12 seconds)

[PDF] **NAVAL POSTGRADUATE SCHOOL THESIS**
edocs.nps.edu/npspubs/scholarly/theses/2005/Sep/05Sep_Rostberg.pdf
File Format: PDF/Adobe Acrobat - Quick View
NAVAL POSTGRADUATE SCHOOL. September 2005. Author: James I. Rostberg. Approved by: Maria Rasmussen. **Thesis** Advisor. Robert Simeral. Co-Advisor ...

[PDF] **NAVAL POSTGRADUATE SCHOOL THESIS**
edocs.nps.edu/npspubs/scholarly/theses/2005/Sep/05Sep_Olson.pdf
File Format: PDF/Adobe Acrobat - Quick View
166 items – from the. **NAVAL POSTGRADUATE SCHOOL**. September 2005. Author: Dean T. Olson. Approved by: David Tucker. **Thesis** Advisor. David Brannan. Second Reader ...

[PDF] **NAVAL POSTGRADUATE SCHOOL THESIS**
edocs.nps.edu/npspubs/scholarly/theses/2007/Jun/07Jun_Acosta.pdf
File Format: PDF/Adobe Acrobat - Quick View
NAVAL POSTGRADUATE SCHOOL. June 2007. Author: David A. Acosta. Approved by: Hy Rothstein. **Thesis** Advisor. Heather S. Gregg. Second Reader. Gordon McCormick ...

[PDF] **NAVAL POSTGRADUATE SCHOOL THESIS**
edocs.nps.edu/npspubs/scholarly/theses/2010/Sep/10Sep_Neese.pdf
File Format: PDF/Adobe Acrobat - Quick View
MASTER OF SCIENCE IN METEOROLOGY from the. **NAVAL POSTGRADUATE SCHOOL**. September 2010. Author: Jay M. Neese. Approved by: Patrick A. Harr. **Thesis** Advisor ...

[PDF] **NAVAL POSTGRADUATE SCHOOL THESIS**
edocs.nps.edu/npspubs/scholarly/theses/2009/Mar/09Mar_Martin.pdf

Figure 15. Incorrect PDF Title Tag in Google Search Results

Table 6. NPS PDF Evaluation (From: NPS Digital Archive, 2011)

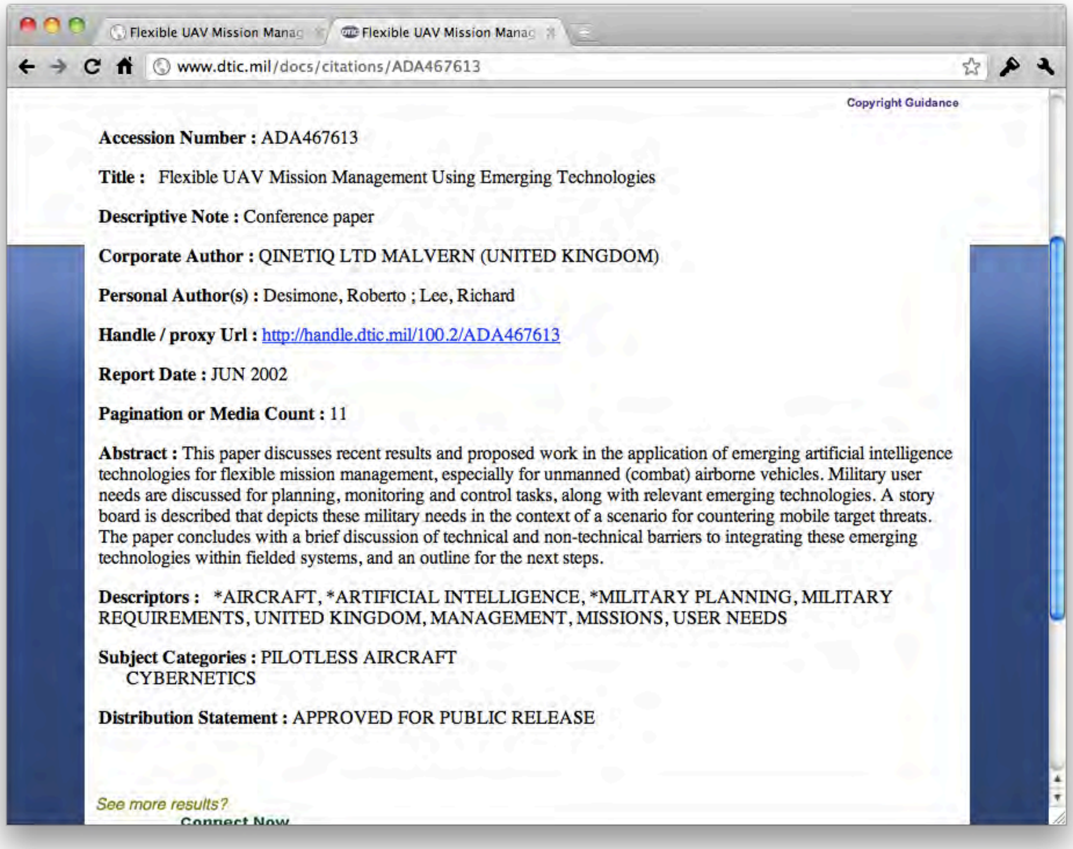
PDFs	Passing %	Grade
Text Based PDF Files	60% * Documents published before September 2001 were scanned with no extractable text	Satisfactory
Document Properties (Title)	Less than 1% *Very low percentage of accurate titles	Needs improvement
File Size (under 10MB)	99% *262 files over 10MB	Excellent
Fast Web View	Less than 1% *Very low percentage	Needs improvement

2. Evaluation of Defense Technical Information Center

The DTIC evaluation consisted of gathering documents from three DTIC websites—the main website located at http://www.dtic.mil/, the citation website located at http://oai.dtic.mil/, and the handle service website located at http://handle.dtic.mil. During a period from January 2011 to May 2011, 200,000 PDF files and 250,000 webpages were downloaded from the DTIC websites. The DTIC study was limited to unclassified/unlimited distribution technical reports.

In January 2011, the DTIC website located at http://www.dtic.mil was evaluated (Figure 16 and Table 7). The first test was used to determine how well documents (Figure 16 and Figure 17) on the website could be found through commercial search engines. This test involved searching for the title of published reports without quotes on the Google search engine. If the report was found in the first page (top 10 links) and linked to a DTIC website, a passing score was recorded. If the report was not found or linked to a location other than the DTIC websites, a failing score was recorded. A total of 383 randomly selected titles were tested. With a 95% Confidence Level, the percentage of DTIC documents that can be found on the first page of Google search results is between 49.3% and 59.3%.

Next, general observations about the website and HTML code is gathered and an evaluation of key website components is performed. The following observations were recorded in January 2011 (see Table 7): The DTIC website located at http://www.dtic.mil/ is the largest central resource of government funded research. The webpages were generally easy to navigate and webpages were well linked. The sitemap.xml file contained links to PDF files but did not contain links to the HTML citation files. The HTML citation files were duplicated on two websites (http://www.dtic.mil and http://oia.dtic.mil) and there was no preferred canonicalization method. The PDF files were duplicated across three locations and there was no preferred canonicalization method. The HTML pages contained no headers or meta-description information.


Flexible UAV Mission Manag
Flexible UAV Mission Manag
www.dtic.mil/docs/citations/ADA467613
Copyright Guidance
See more results?
Connect Now

Accession Number : ADA467613

Title : Flexible UAV Mission Management Using Emerging Technologies

Descriptive Note : Conference paper

Corporate Author : QINETIQ LTD MALVERN (UNITED KINGDOM)

Personal Author(s) : Desimone, Roberto ; Lee, Richard

Handle / proxy Url : http://handle.dtic.mil/100.2/ADA467613

Report Date : JUN 2002

Pagination or Media Count : 11

Abstract : This paper discusses recent results and proposed work in the application of emerging artificial intelligence technologies for flexible mission management, especially for unmanned (combat) airborne vehicles. Military user needs are discussed for planning, monitoring and control tasks, along with relevant emerging technologies. A story board is described that depicts these military needs in the context of a scenario for countering mobile target threats. The paper concludes with a brief discussion of technical and non-technical barriers to integrating these emerging technologies within fielded systems, and an outline for the next steps.

Descriptors : *AIRCRAFT, *ARTIFICIAL INTELLIGENCE, *MILITARY PLANNING, MILITARY REQUIREMENTS, UNITED KINGDOM, MANAGEMENT, MISSIONS, USER NEEDS

Subject Categories : PILOTLESS AIRCRAFT
CYBERNETICS

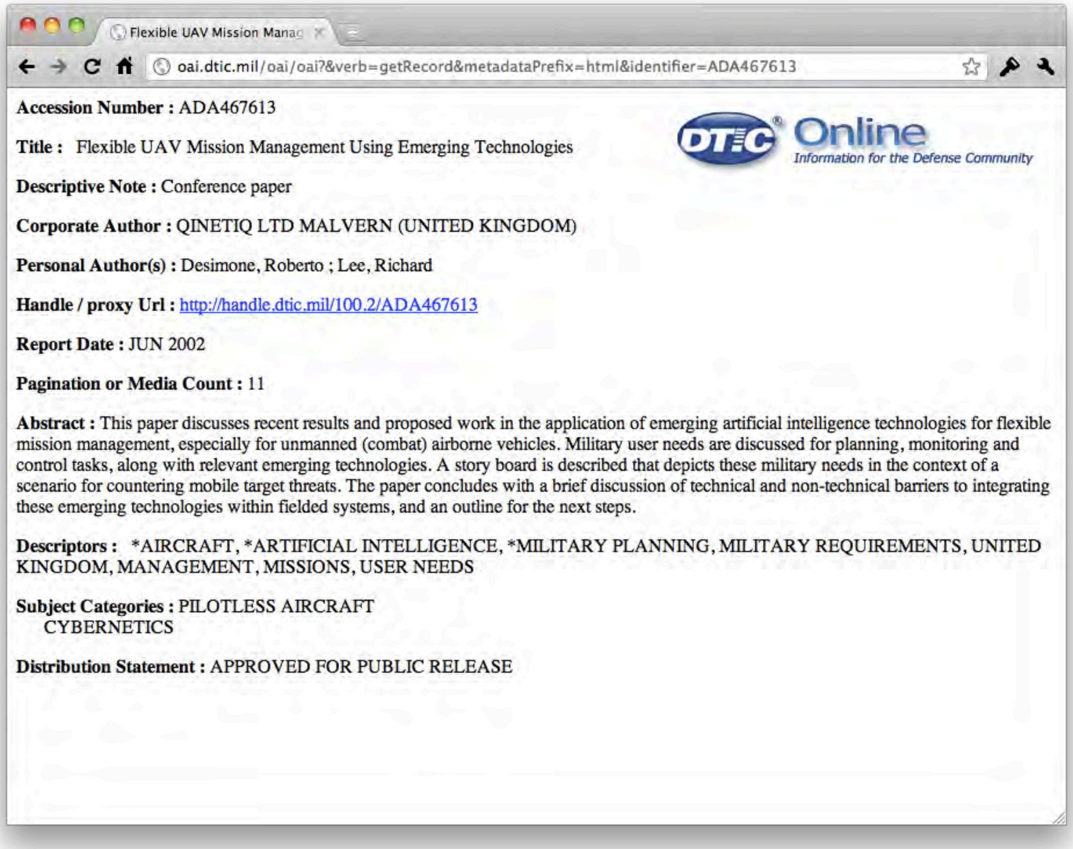
Distribution Statement : APPROVED FOR PUBLIC RELEASE

Figure 16. Individual Report Page (From: DTIC, 2011)

Table 7. DTIC Website Evaluation (DTIC, 2011)

Item	Passing %	Grade
Website Search Engine Components		
Robots.txt	100%	Excellent
Sitemap.xml	50%	Needs improvement
Extractable Text and Links	100%	Excellent
Accessible URL Structure	100%	Excellent
Fast response / load times	100%	Excellent
Stable / Permanent Links	100%	Excellent
Server Redirects	Mixed use * Temporary redirects are used with the handle service	Needs improvement
Canonicalization	0%	Needs improvement
Avoiding Duplicate Content	97% [content] *based on close title and abstract data 33% [pdf files x3] 50% [html citations x2]	Excellent Needs improvement Needs improvement
Webpage Search Engine Components		
Standard HTML (W3C Validation)	96%	Excellent
Page Title	100%	Excellent
Meta Description	[not found]	Needs improvement
Tags	100%	Excellent
Headings	[not found]	Excellent
Paragraphs	100%	Excellent
Alt Image Text	N/A	N/A
Use of Internal Links	100% * Missing author, publisher, and tag links	Excellent
Clear Site Navigation	100%	Excellent

In January 2011, the DTIC website located at http://oia.dtic.mil was evaluated (Figure 17 and Table 8). The DTIC website located at http://oai.dtic.mil/ is used primarily to store PDF citations in HTML format. The webpages were very difficult to navigate because the linking structure on the home page is out of date (last linked item was June 13, 2007) and there are no navigation links on individual report pages. The website was missing several key components including server redirects, canonicalization and headings. The URL structure works with upper- and lower-case characters and does not redirect to a preferred page. The citations are duplicates to the pages found at the http://www.dtic.mil/ website and there is no canonicalization.


Flexible UAV Mission Manag
oai.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA467613
DTIC Online
Information for the Defense Community

Accession Number : ADA467613

Title : Flexible UAV Mission Management Using Emerging Technologies

Descriptive Note : Conference paper

Corporate Author : QINETIQ LTD MALVERN (UNITED KINGDOM)

Personal Author(s) : Desimone, Roberto ; Lee, Richard

Handle / proxy Url : http://handle.dtic.mil/100.2/ADA467613

Report Date : JUN 2002

Pagination or Media Count : 11

Abstract : This paper discusses recent results and proposed work in the application of emerging artificial intelligence technologies for flexible mission management, especially for unmanned (combat) airborne vehicles. Military user needs are discussed for planning, monitoring and control tasks, along with relevant emerging technologies. A story board is described that depicts these military needs in the context of a scenario for countering mobile target threats. The paper concludes with a brief discussion of technical and non-technical barriers to integrating these emerging technologies within fielded systems, and an outline for the next steps.

Descriptors : *AIRCRAFT, *ARTIFICIAL INTELLIGENCE, *MILITARY PLANNING, MILITARY REQUIREMENTS, UNITED KINGDOM, MANAGEMENT, MISSIONS, USER NEEDS

Subject Categories : PILOTLESS AIRCRAFT
CYBERNETICS

Distribution Statement : APPROVED FOR PUBLIC RELEASE

Figure 17. Individual Report Page (From: DTIC Citation Website, 2011)

Table 8. DTIC Website Evaluation (DTIC Citation Website, 2011)

Item	Passing %	Grade
Website Search Engine Components	-	
Robots.txt	100%	Excellent
Sitemap.xml	100%	Excellent
Extractable Text and Links	Text: 100% Links: [not found on report pages]	Text: Excellent Links: Needs improvement
Accessible URL Structure	100%	Excellent
Fast response / load times	100%	Excellent
Stable / Permanent Links	100%	Excellent
Server Redirects	[not found]	Needs improvement
Canonicalization	[not found]	Needs improvement
Avoiding Duplicate Content	50% [html citations x2]	Needs improvement
Webpage Search Engine Components		
Standard HTML (W3C Validation)	88%	Excellent
Page Title	100%	Excellent
Meta Description	100%	Excellent
Tags	100%	Excellent
Headings	[not found]	Needs improvement
Paragraphs	100%	Excellent
Alt Image Text	N/A	N/A
Use of Internal Links	[not found]	Needs improvement
Clear Site Navigation	[not found]	Needs improvement

In January 2011, the DTIC PDF collection was evaluated (Table 9). The PDF storage system located at http://handle.dtic.mil uses a handle system to assign persistent identifiers and URLs to electronic resources. The handle service prevents broken links and ensures long-term access to electronic resources. Approximately 70% of the PDF files collected were text based. Very few of the PDF files contained a title tag or were optimized for fast web view. The handle service did not reduce the amount of duplicate content on the DTIC website. For example, there are at least three ways to download identical files from DTIC: the handle service, the cgi-bin and the full text folder. As stated before, duplicate content is undesirable because it reduces page authority, increases the time it takes search engines to crawl a website, and creates confusion over which document should be used.

Examples of duplicate PDFs through URLs at DTIC:

- URL 1 (handle): http://handle.dtic.mil/100.2/ADA501360
- URL 2 (cgi-bin): http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ada501360&Location=U2&doc=GetTRDoc.pdf
- URL 3 (full text): http://www.dtic.mil/dtic/tr/fulltext/u2/a501360.pdf

Table 9. DTIC PDF Evaluation (From: DTIC Handle Service Website, 2011)

PDF Documents	Passing %	Grade
Text Based PDF Files	70% * Estimation based on document range collected Documents published before ADA3xxxxx were not collected and will likely lower this value	Satisfactory
Document Properties (Title)	Less than 1% * Very low percentage of accurate titles	Needs improvement
File Size (10MB)	94% * 11,789 / 200,000 files oversized * Estimation based on document range collected Documents published before ADA3xxxxx were not collected and will likely lower this value	Good
Fast Web View	Less than 1% *Very low percentage	Needs improvement

C. HOW CAN THE DOD INCREASE EXPLICIT KNOWLEDGE PRESENCE THROUGH SEARCH ENGINES?

Knowledge management systems (KMS) are IT systems developed to support “knowledge creation, storage/retrieval, transfer, and application” (Alavi & Leidner, 2001, p. 116). These systems are designed to provide knowledge to the correct personnel at the appropriate time in order to improve knowledge flows.

Information technologies like the Internet, combined with the use of websites and search engines, can be used to enhance KM. When building a KMS, it is vital to understand the importance of search and to ensure the organization is not creating barriers to block the KP’s potential. In addition to adhering to the best practices discussed in the previous sections, these key matters should be noted.

1. PDFs and Search Engines

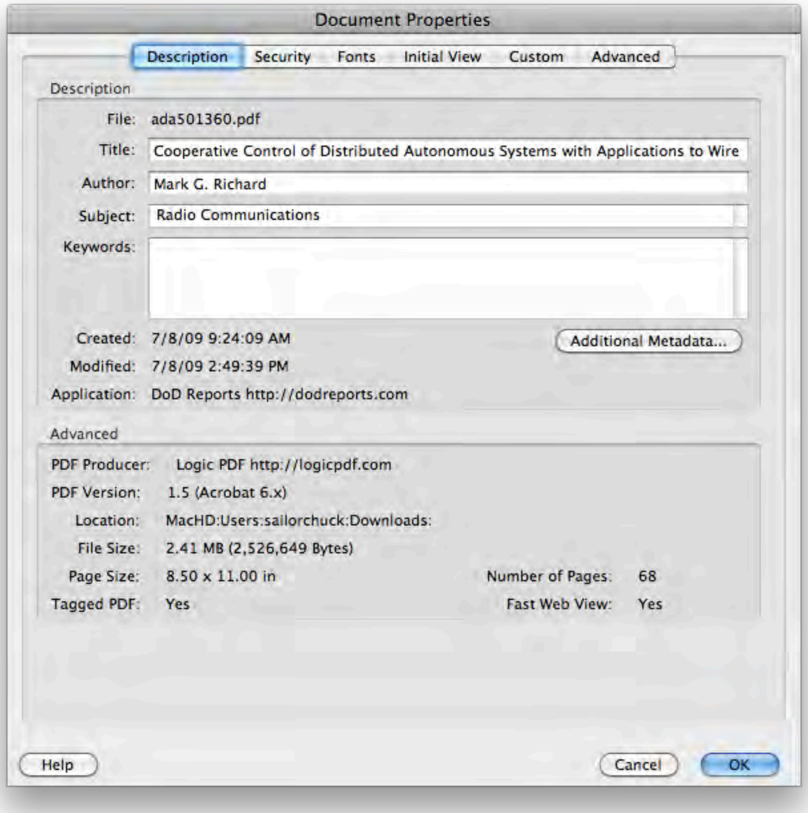
A PDF file is a simple way to share content over the Internet. Most PDF viewers are free and can be integrated with web browsers. As with the website and the HTML components listed before, there are key PDF components necessary for search.

a. Text Based PDF Files

Normally, there are two types of PDF files, text based and image based. Image-based PDF files have images of text that cannot be indexed. Text-based PDF files are better for search engines because content can be extracted and indexed. Text-based PDF files have a smaller file size, enabling faster downloads for end users. Image-based PDF files can be converted to text-based PDFs by using Optical character recognition (OCR) technology.

b. Document Properties

Similar to HTML files, PDFs have meta tags that can be used by search engines. Basic description tags include title, author, subject, keywords, and creation and modification dates (Figure 18).



Description

File: ada501360.pdf

Title: Cooperative Control of Distributed Autonomous Systems with Applications to Wire

Author: Mark G. Richard

Subject: Radio Communications

Keywords:

Created: 7/8/09 9:24:09 AM

Modified: 7/8/09 2:49:39 PM

Application: DoD Reports http://dodreports.com

Advanced

PDF Producer: Logic PDF http://logicpdf.com

PDF Version: 1.5 (Acrobat 6.x)

Location: MacHD:Users:sailorchuck:Downloads:

File Size: 2.41 MB (2,526,649 Bytes)

Page Size: 8.50 x 11.00 in

Number of Pages: 68

Tagged PDF: Yes

Fast Web View: Yes

Figure 18. PDF Document Properties

c. *File Size*

When PDF files contain a large number of images, the file size can become very large. At the time of this writing, Google's search engine has a PDF size limit of 10MB (Google Webmaster Tools, 2011).

d. *Fast Web View*

Fast web view is a feature that allows page-at-a-time downloading from webservers. This allows clients faster access to pages of a PDF without the need to download the complete file first.

e. Font Size and Placement

Like HTML files, font size and placement is important for search engines. For example, if a PDF file is not tagged with a title, most search engines will look for the largest text near the top of the page and set this text as the title.

f. Reading Order

When PDF documents are structured in multiple columns, it can be difficult for a search engine to determine the reading order of the page. Setting the reading order of paragraphs can help search engines and users follow the reading order.

2. Website Extensions

It is possible to extend the presence of a website through syndication. Syndication allows content to be shared in channels other than the primary website. Common syndication methods include Really Simple Syndication (RSS) feeds and social media.

a. RSS Feeds

RSS Feeds are a way to syndicate content through Extensible Markup Language (XML) files. This is useful for frequently updated information such as news headlines. Search engines are able to monitor RSS feeds and update their index when new content is posted.

b. Social Media (Twitter, Facebook, Google+)

Social Media is still a relatively new and rapidly expanding part of the Internet. Social Media support can help people share knowledge where direct face-to-face interactions are not possible due to the distance between individuals. IT supports this transfer through informal means by enabling communication between two individuals or socialization. When socialization occurs, new knowledge can be created when someone gains insight. The social media company, Facebook, is only seven years old but at the time of this writing has 600 million+ active monthly users (Carlson, 2011).

3. Internal Search Engines

Commercial search engines cannot access all of an organization's databases. In the DoD, commercial search engines cannot be given access to any information that is controlled or items marked as classified. In order to access specialized databases and controlled information, internal search engines are required. Internal search engines can use technology that ranges from simple structured query language (SQL) keyword matching to a dedicated search appliance. The Google Search Appliance (Figure 19) works like Google's public search engine but can be customized and tuned to meet an organization's needs and remain on a private network.

Figure 19. Google Search Appliance (From Dell, 2011)

4. New Technology

Technology advances quickly and organizations need to have the ability to adapt to take advantage of it. This is particularly important in the DoD where military units need to maintain competitive advantages. One new and quickly advancing technology is the use of small mobile computing devices. These devices often have small screens. Without modification, existing webpages may not display well on these devices. It is important to constantly adapt and plan for these types of changes.

D. PROTOTYPE APPLICATION THAT CAN BE APPLIED GENERALLY

The construction of the demonstration application (dodreports.com) required approximately six months to create. The website was built using NPS and DTIC documents to evaluate how search plays a role in knowledge management systems. To make reports available to the large percentage of Internet searchers, the demonstration website was built to be discoverable, accessible, and extractable. Many modifications were made to the design in order to test the effects on search results. The construction of the application required gathering content, processing content, and creating the application to host the content.

1. Gathering Content

From the period of December 2010 to February 2011, 450,000 DTIC documents and 10,200 NPS documents were downloaded through a custom built search robot. The search robot was designed to look for files through links in the sitemap.xml, internal webpages, and through predictable URL structures. The DTIC documents collected consisted of HTML citation pages and PDF documents. The NPS documents consisted of HTML link pages and PDF documents.

2. Processing Content

To process the documents gathered, several custom scripts were created. The primary task of these scripts was to locate the title, author, publication date, and abstract of the documents collected. In the DTIC collection, it was relatively easy to collect this information from the HTML citation pages. In the NPS collection, this information was extracted directly from the PDF files using OCR technology on documents dated on or before September 2001. For documents after September 2001, the information was extracted text directly from the front cover/abstract page. The extraction of text from the NPS documents was not precise and errors were encountered in approximately 10% of the reports.

3. Use of Enterprise Architecture

An enterprise architecture strategy was used to plan the development of the demonstration website and its components. Key components of the enterprise architecture used include the web server, database server, and application server.

a. Databases

A database can describe everything from telephone book listings to complex programs such as a database management system (DBMS). This report uses the term database to describe the combination of structure and data. Structure includes both the procedures used to retrieve data and the rules used to protect data. Data are the physical text and variables stored in tables. A database provides many advantages. It can simplify programming and reduce costs, allow managers to redesign or reorganize quickly, and enforces integrity constraints that improve data quality. A database can reduce unwanted redundancy and reduce the likelihood of errors during updates through the process of normalization.

The demonstration website was built using a MySQL database with four tables. The entity-relationship (ER) used by the database is shown in Figure 20. Referential integrity was built into the demonstration using application server code.

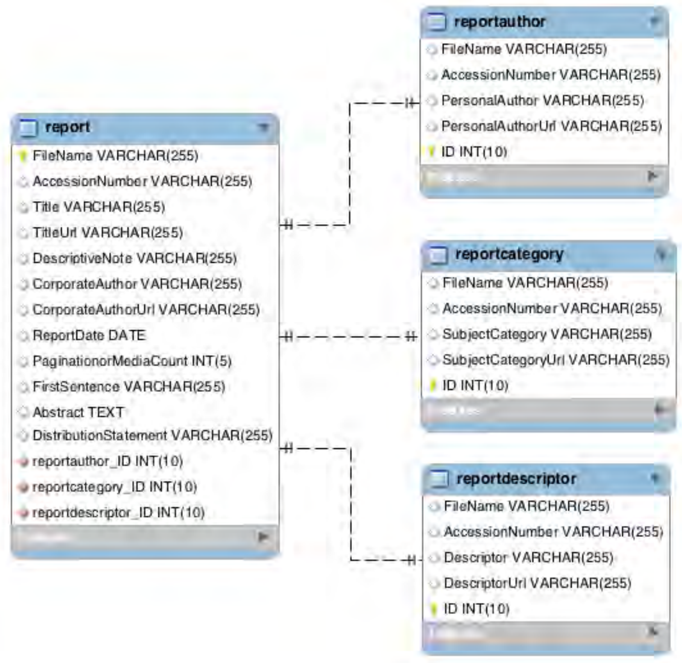

reportauthor
FileName VARCHAR(255)
AccessionNumber VARCHAR(255)
PersonalAuthor VARCHAR(255)
PersonalAuthorUrl VARCHAR(255)
ID INT(10)
report
FileName VARCHAR(255)
AccessionNumber VARCHAR(255)
Title VARCHAR(255)
TitleUrl VARCHAR(255)
DescriptiveNote VARCHAR(255)
CorporateAuthor VARCHAR(255)
CorporateAuthorUrl VARCHAR(255)
ReportDate DATE
PaginationorMediaCount INT(5)
FirstSentence VARCHAR(255)
Abstract TEXT
DistributionStatement VARCHAR(255)
reportauthor_ID INT(10)
reportcategory_ID INT(10)
reportdescriptor_ID INT(10)
reportcategory
FileName VARCHAR(255)
AccessionNumber VARCHAR(255)
SubjectCategory VARCHAR(255)
SubjectCategoryUrl VARCHAR(255)
ID INT(10)
reportdescriptor
FileName VARCHAR(255)
AccessionNumber VARCHAR(255)
Descriptor VARCHAR(255)
DescriptorUrl VARCHAR(255)
ID INT(10)

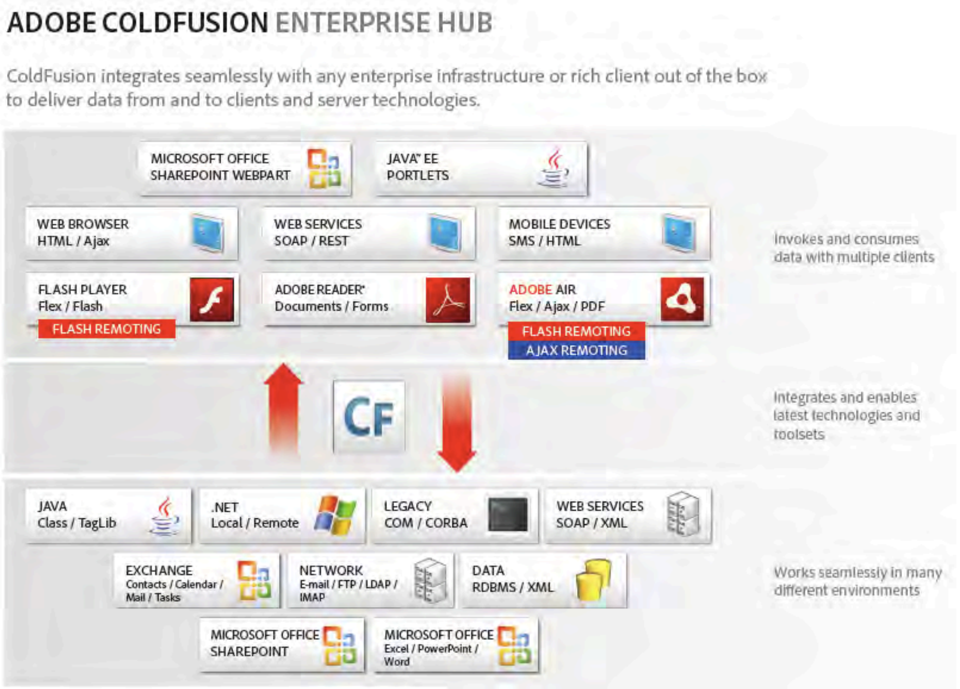
Figure 20. Entity Relationship Diagram

b. Web Server

A web server is a piece of software that uses the HyperText Transfer Protocol (HTTP) to transmit Hypertext Markup Language (HTML) files or other standard files from a server to a client computer's browser. A client's browser is able to process the HTML files and render text and images for display. Web server software is simple and can only process static files. Common web servers include Apache HTTP Server and Microsoft Internet Information Services (IIS).

c. Application Server

Application software allows users to interact with a database. Typically, users will access forms and reports to create, read, update and delete information from a database. Users typically interact with application software through desktop programs and web browsers. Application software can be standalone in nature and interact with an internal database, or it can be web/cloud based and interact with a shared database. The use of web or cloud based solutions allows greater sharing and collaboration. An application server allows an organization to transparently integrate different systems into a single interface, as shown in Figure 21. Application server software increases agility and allows for quicker changes to needs and demands. It can also increase usability by decreasing the number of systems users need to learn and interact with. Application servers serve logic to application programs through many protocols, including HTTP, Post Office Protocol (POP), Simple Mail Transfer Protocol (SMTP), File Transfer Protocol (FTP), etc. They are equipped to support more complex enterprise infrastructures when compared to standard web servers.


ADOBE COLDFUSION ENTERPRISE HUB

ColdFusion integrates seamlessly with any enterprise infrastructure or rich client out of the box to deliver data from and to clients and server technologies.

Invokes and consumes data with multiple clients

Integrates and enables latest technologies and toolsets

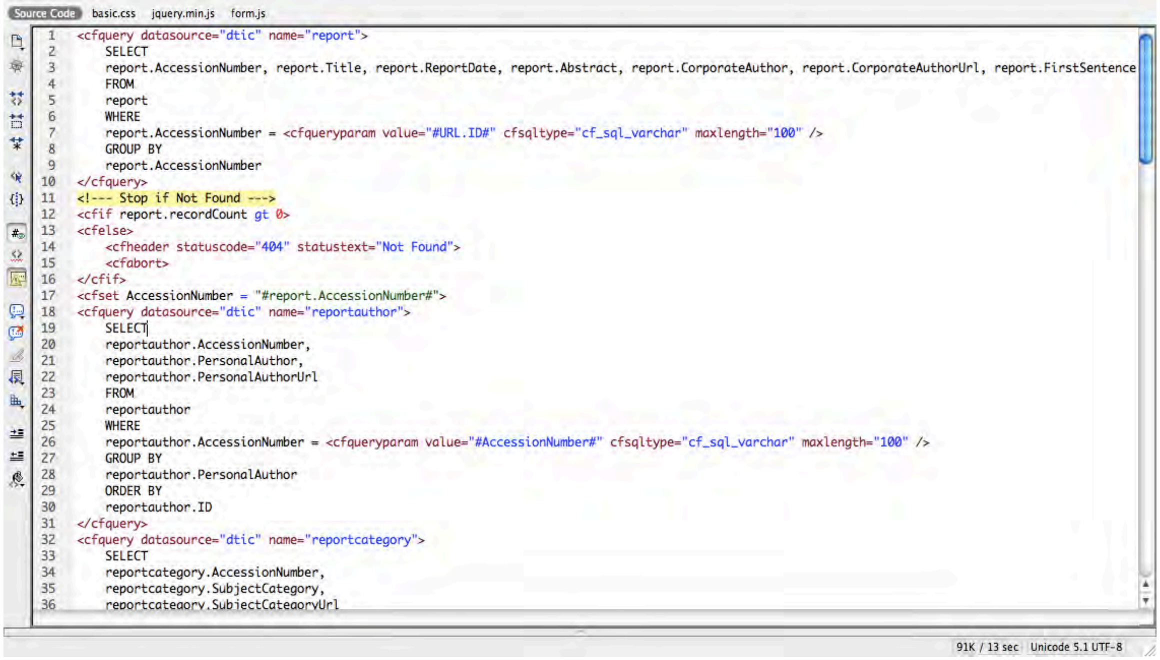
Works seamlessly in many different environments

MICROSOFT OFFICE SHAREPOINT WEBPART
JAVA EE PORTLETS
WEB BROWSER HTML / Ajax
WEB SERVICES SOAP / REST
MOBILE DEVICES SMS / HTML
FLASH PLAYER Flex / Flash
FLASH REMOTING
ADOBE READER Documents / Forms
ADOBE AIR Flex / Ajax / PDF
FLASH REMOTING
AJAX REMOTING
CF
JAVA Class / TagLib
.NET Local / Remote
LEGACY COM / CORBA
WEB SERVICES SOAP / XML
EXCHANGE Contacts / Calendar / Mail / Tasks
NETWORK E-mail / FTP / LDAP / IMAP
DATA RDBMS / XML
MICROSOFT OFFICE SHAREPOINT
MICROSOFT OFFICE Excel / PowerPoint / Word

Figure 21. Integration of Many Services with ColdFusion (From: Adobe Evangelism Kit, 2010)

Application servers can be used to present rich information in the form of HTML, flash, and images, Microsoft Documents and PDF documents. They also allow users to update databases, enabling the use of user-generated content. Application servers provide value to the enterprise because they are relatively transparent; they enforce a standard way of doing things and they tie together complex systems. In essence, it is the application server that can be used as the glue that holds together enterprise systems.

The demonstration website used application server code to connect the database to the webserver. In Figure 22, queries were used to connect to the MySQL database in order to extract report information.



```
<cfquery datasource="dtic" name="report">
    SELECT
    report.AccessionNumber, report.Title, report.ReportDate, report.Abstract, report.CorporateAuthor, report.CorporateAuthorUrl, report.FirstSentence
    FROM
    report
    WHERE
    report.AccessionNumber = <cfqueryparam value="#URL.ID#" cfsqltype="cf_sql_varchar" maxlength="100" />
    GROUP BY
    report.AccessionNumber
</cfquery>
<!--- Stop if Not Found --->
<cfif report.recordCount gt 0>
<cfelse>
    <cfheader statuscode="404" statustext="Not Found">
    <cfabort>
</cfif>
<cfset AccessionNumber = "#report.AccessionNumber#">
<cfquery datasource="dtic" name="reportauthor">
    SELECT
    reportauthor.AccessionNumber,
    reportauthor.PersonalAuthor,
    reportauthor.PersonalAuthorUrl
    FROM
    reportauthor
    WHERE
    reportauthor.AccessionNumber = <cfqueryparam value="#AccessionNumber#" cfsqltype="cf_sql_varchar" maxlength="100" />
    GROUP BY
    reportauthor.PersonalAuthor
    ORDER BY
    reportauthor.ID
</cfquery>
<cfquery datasource="dtic" name="reportcategory">
    SELECT
    reportcategory.AccessionNumber,
    reportcategory.SubjectCategory,
    reportcategory.SubjectCategoryUrl
```

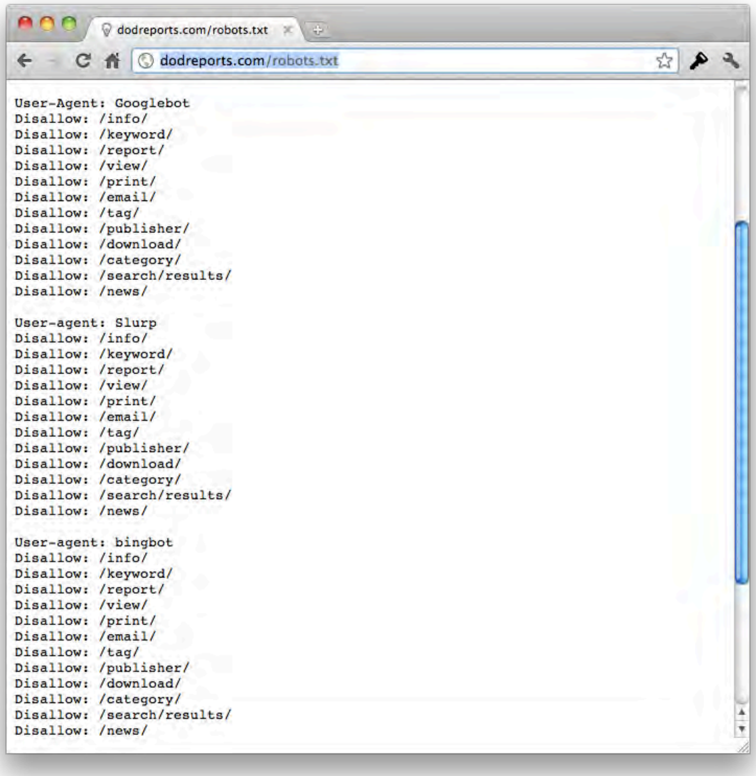
Figure 22. Application Server Code

4. Search Components

The demonstration website used search components identified in Table 10 (Search Components Checklist). The robots.txt shown in Figure 23. DoDReports Robots.txt was used to target specific search engines (Google, Yahoo and Bing) and prevent them from indexing directories that contained duplicate content and pages that are not useful to index. For example, following links on the /search/results directory would result in an infinite number of possible pages and links, making it is important to prevent robots from indexing this page. The sitemap.xml shown in Figure 24 contained more than 50,000 links and needed to be split up into five files. The sitemap.xml links to the five sitemap files. An attempt was made to extract all the components shown in Figure 18 for the PDF documents on the http://dodreports.com website. This proved to be impractical due to several factors, including the high central processing unit (CPU) demands of OCR software, an OCR character conversion error rate of 1% and an OCR formatting error rate of 10%. As a result, only the title and author were extracted.

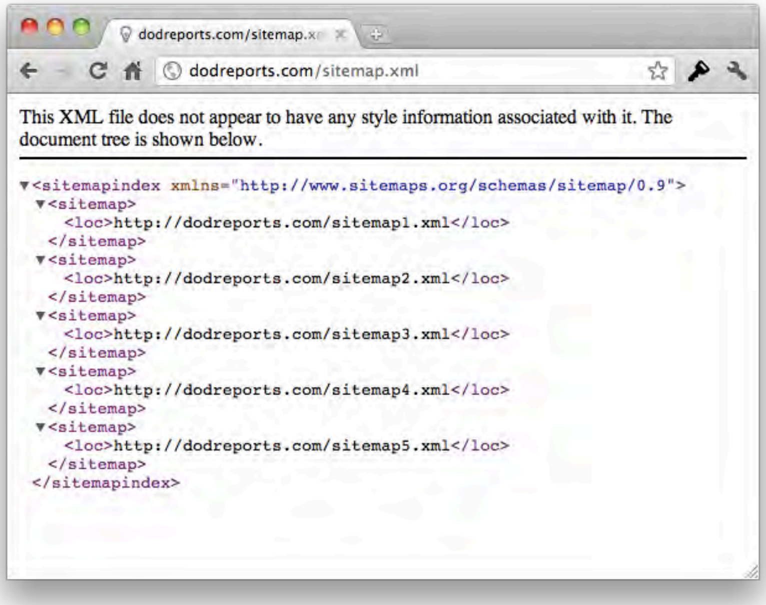
Table 10. Search Components Checklist

Item
Search Engine Friendly Website
Robots.txt
Sitemap.xml
Extractable Text and Links
Accessible URL Structure
Fast response / load times
Stable / Permanent Links
Server Redirects
Canonicalization
Avoiding Duplicate Content
Search Engine Friendly Webpages
Standard HTML (W3C Validation)
Page Title
Meta Description
Tags
Headings
Paragraphs
Alt Image Text
Use of Internal Links
Clear Site Navigation



```
User-Agent: Googlebot
Disallow: /info/
Disallow: /keyword/
Disallow: /report/
Disallow: /view/
Disallow: /print/
Disallow: /email/
Disallow: /tag/
Disallow: /publisher/
Disallow: /download/
Disallow: /category/
Disallow: /search/results/
Disallow: /news/

User-agent: Slurp
Disallow: /info/
Disallow: /keyword/
Disallow: /report/
Disallow: /view/
Disallow: /print/
Disallow: /email/
Disallow: /tag/
Disallow: /publisher/
Disallow: /download/
Disallow: /category/
Disallow: /search/results/
Disallow: /news/

User-agent: bingbot
Disallow: /info/
Disallow: /keyword/
Disallow: /report/
Disallow: /view/
Disallow: /print/
Disallow: /email/
Disallow: /tag/
Disallow: /publisher/
Disallow: /download/
Disallow: /category/
Disallow: /search/results/
Disallow: /news/
```

Figure 23. DoDReports Robots.txt



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<sitemap>
     <loc>http://dodreports.com/sitemap1.xml</loc>
   </sitemap>
  ▼<sitemap>
     <loc>http://dodreports.com/sitemap2.xml</loc>
   </sitemap>
  ▼<sitemap>
     <loc>http://dodreports.com/sitemap3.xml</loc>
   </sitemap>
  ▼<sitemap>
     <loc>http://dodreports.com/sitemap4.xml</loc>
   </sitemap>
  ▼<sitemap>
     <loc>http://dodreports.com/sitemap5.xml</loc>
   </sitemap>
 </sitemapindex>
```

Figure 24. DodReports Sitemap

Table 11. PDF Search Components Checklist

Item
Text Based PDF Files
Document Properties (Title)
File Size (under 10MB) – if possible
Fast Web View

a. Discovery

The first step external search engines use to find content is through discovery. The website was made discoverable by submitting the website's URL to Webmaster tools at Google, Bing and Yahoo. The website was designed to use a sitemap.xml file, and internal links to guide the search engines' bots. To ensure the bots could effectively crawl the website, a robots.txt file was used to block pages the robots should not crawl, and server redirects were used to identify the preferred URLs.

b. Accessibility

The URL structure was made accessible by using a simple and short structure that gave each report is own alphanumeric identifier. The URLs were structured so they only could resolve in one way. When pages moved, permanent redirects were used to automatically redirect users and search engines.

c. Extractability

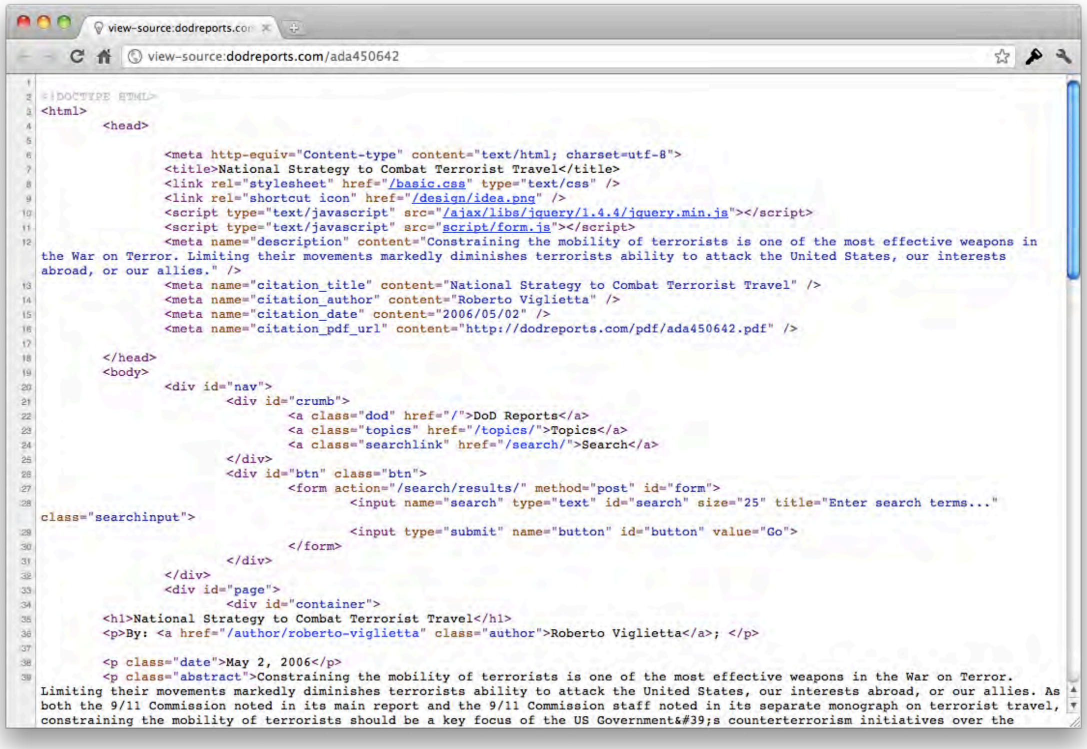
To make content extractable, every report was represented by a HTML page that was tagged with a page title, meta description, headings, publication date, author's name, and a paragraph containing the abstract. In a similar fashion, PDF files were made more extractable by tagging them with the title and author's name. The PDF files were also modified to enable fast web view.

d. Usability

Each webpage was built with usability in mind. Each page was styled with cascading style sheets (CSS) to make important elements (title, date, author) stand out. To make the website more useful, dynamic pages were created to display reports that shared a common author or common tags. An internal search engine was integrated to allow users to quickly find reports by title, abstract, author and publisher.

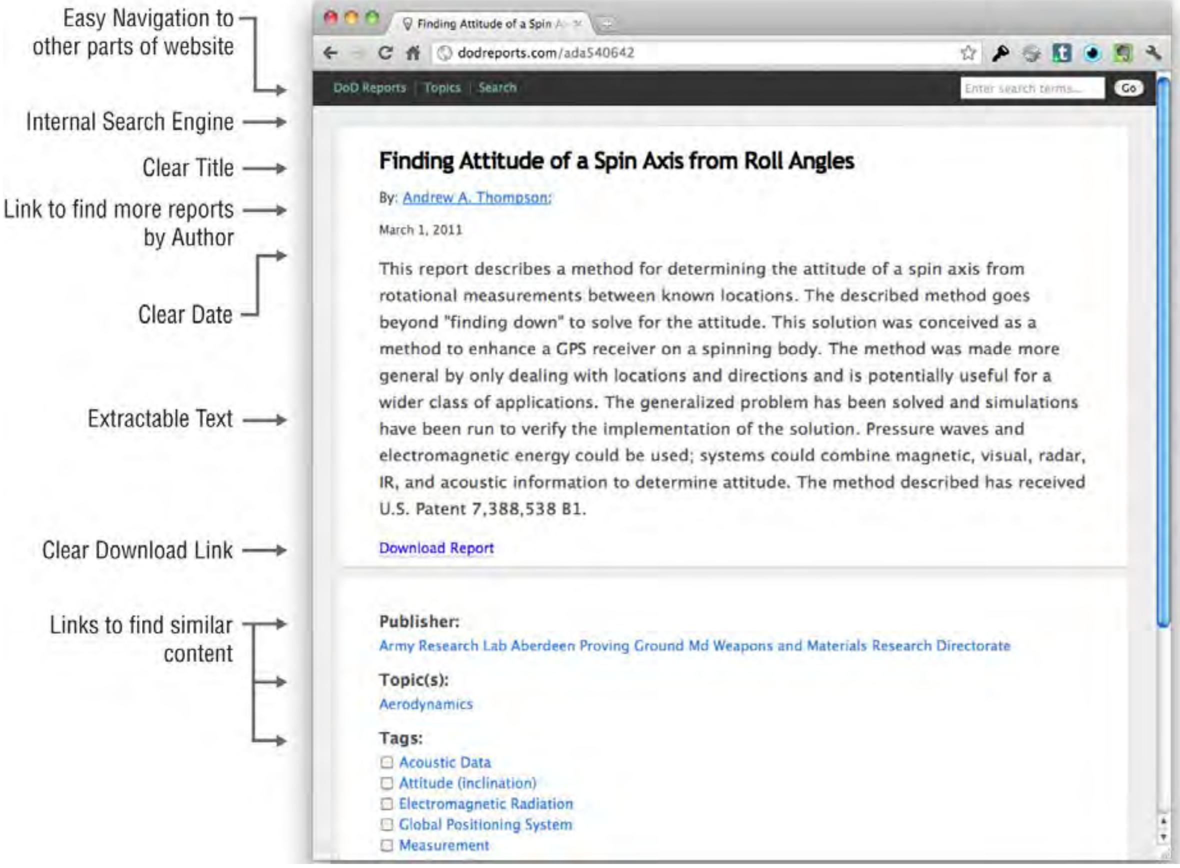
e. Design Components

The dodreports.com uses the following design components: clear website navigation, internal search engine, clear title, author's name with link, clear date, clear download link and links that could be used to find similar content. The design was created using HTML and CSS markup shown in Figure 25. The resulting display is shown in Figure 26.


view-source:dodreports.com
view-source:dodreports.com/ada450642

```

<!DOCTYPE HTML>
<html>
	<head>

		<meta http-equiv="Content-type" content="text/html; charset=utf-8">
		<title>National Strategy to Combat Terrorist Travel</title>
		<link rel="stylesheet" href="/basic.css" type="text/css" />
		<link rel="shortcut icon" href="/design/idea.png" />
		<script type="text/javascript" src="/ajax/libs/jquery/1.4.4/jquery.min.js"></script>
		<script type="text/javascript" src="script/form.js"></script>
		<meta name="description" content="Constraining the mobility of terrorists is one of the most effective weapons in
the War on Terror. Limiting their movements markedly diminishes terrorists ability to attack the United States, our interests
abroad, or our allies." />
		<meta name="citation_title" content="National Strategy to Combat Terrorist Travel" />
		<meta name="citation_author" content="Roberto Viglietta" />
		<meta name="citation_date" content="2006/05/02" />
		<meta name="citation_pdf_url" content="http://dodreports.com/pdf/ada450642.pdf" />

	</head>
	<body>
		<div id="nav">
			<div id="crumb">
				<a class="dod" href="/">DoD Reports</a>
				<a class="topics" href="/topics/">Topics</a>
				<a class="searchlink" href="/search/">Search</a>
			</div>
			<div id="btn" class="btn">
				<form action="/search/results/" method="post" id="form">
					<input name="search" type="text" id="search" size="25" title="Enter search terms..."
class="searchinput">
					<input type="submit" name="button" id="button" value="Go">
				</form>
			</div>
		</div>
		<div id="page">
			<div id="container">
	<h1>National Strategy to Combat Terrorist Travel</h1>
	<p>By: <a href="/author/roberto-viglietta" class="author">Roberto Viglietta</a>; </p>

	<p class="date">May 2, 2006</p>
	<p class="abstract">Constraining the mobility of terrorists is one of the most effective weapons in the War on Terror.
Limiting their movements markedly diminishes terrorists ability to attack the United States, our interests abroad, or our allies. As
both the 9/11 Commission noted in its main report and the 9/11 Commission staff noted in its separate monograph on terrorist travel,
constraining the mobility of terrorists should be a key focus of the US Government&#39;s counterterrorism initiatives over the
```

Figure 25. DoD Reports HTML Code


Easy Navigation to other parts of website
Internal Search Engine
Clear Title
Link to find more reports by Author
Clear Date
Extractable Text
Clear Download Link
Links to find similar content
Finding Attitude of a Spin A
dodreports.com/ada540642
DoD Reports | Topics | Search
Enter search terms...
Go

Finding Attitude of a Spin Axis from Roll Angles

By: Andrew A. Thompson;

March 1, 2011

This report describes a method for determining the attitude of a spin axis from rotational measurements between known locations. The described method goes beyond "finding down" to solve for the attitude. This solution was conceived as a method to enhance a GPS receiver on a spinning body. The method was made more general by only dealing with locations and directions and is potentially useful for a wider class of applications. The generalized problem has been solved and simulations have been run to verify the implementation of the solution. Pressure waves and electromagnetic energy could be used; systems could combine magnetic, visual, radar, IR, and acoustic information to determine attitude. The method described has received U.S. Patent 7,388,538 B1.

Download Report

Publisher:
Army Research Lab Aberdeen Proving Ground Md Weapons and Materials Research Directorate

Topic(s):
Aerodynamics

Tags:
☐ Acoustic Data
☐ Attitude (inclination)
☐ Electromagnetic Radiation
☐ Global Positioning System
☐ Measurement

Figure 26. Webpage Design Properties on DoD Reports

f. Side-by-Side Evaluation

In August 2011, a side-by-side evaluation was conducted between NPS, DTIC and the dodreports.com website. Using 7,217 reports that were common between all three websites, a random sample of 365 titles was evaluated with the following results:

- With a 95% Confidence Level, the percentage of NPS documents that can be found on the first page of Google search results is between 26.5% and 36.5%.
- With a 95% Confidence Level, the percentage of DTIC documents that can be found on the first page of Google search results is between 39.1% and 49.1%.
- With a 95% Confidence Level, the percentage of dodreports.com documents that can be found on the first page of Google search results is between 47.6% and 57.6%.

DTIC's website performed 12.6% better than the NPS digital archive. The dodreports.com website performed 21.1% better than the NPS digital archive and 8.5% better than DTIC's website. This is direct evidence that the items presented in this thesis have a high impact on search results.

Table 12. Google Search Statistical Results

Website	**Confidence Level**	**Mean**
NPS	95%	31.5%
DTIC	95%	44.1%
DoD Reports	95%	52.6%

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

A. RESEARCH RESULTS

The overarching goal of this thesis was to develop techniques for implementing search engine ready knowledge management systems. Twelve years ago, commercial search engines were primarily used by early adopters. Now, 92% of online adults use search engines to find information (Purcell, 2011). It is important for organizations to understand the importance of search and to implement a strategy to take advantage of commercial search engines.

1. Objectives Accomplished

The demonstration website accomplished the goal of creating a search engine ready knowledge management system. The primary research question has been answered with a useable system. The system satisfies the very specific requirements for commercial search engines that are currently not well implemented in the DoD.

Most of the effort in developing the application (70%–80%) was spent in testing potential designs, learning about the technologies used, and creating solutions to challenges faced. Having developed a unique process, the effort required to reproduce the work would be an estimated one tenth of the original. The output of the research resulted in the following:

- Identified 9 critical components that are necessary for search in websites (robots.txt, sitemap.xml, extractable content, fast response times, stable links, server redirects, canonicalization, and no duplicate content)
- Identified 9 critical components that are necessary for search in webpages (standard HTML, page title, meta description, tags, headings, paragraphs, alt image text, internal links, and clear site navigation)
- Identified 4 critical components that are necessary for search in PDFs (text based, title, file size, and fast web view)
- Created 1 MySQL database with 4 tables and 200,000 records

- Created 1 Web Crawler used to gather the data
- Created 7 ColdFusion Markup (CFM) pages (home, report, topic, author, publisher, tag, and search)
- Created 1 JavaScript file (used by the HTML forms)
- Created 1 CSS file (used by every HTML page)
- Tagged 200,000 PDF files

2. Performance

The primary measure of performance used to evaluate the developed application, involved testing whether reports could be found on the first page (top 10 results) of Google search results by searching for their exact title. During the first 6 months of research, performance gains of 8.5%–21.1% were recorded over the current DoD systems at a very low cost.

Further development efforts on this topic should include other commercial search engines like Yahoo and Bing and other search parameters like report headings and sentences. The ability to search for specific text, range of dated material, and content by particular data category (e.g., military aircraft organization) should also be explored.

B. RECOMMENDATIONS

1. Implementation

The current application is a production-ready system. Having proved that the demonstration concept can work and identifying the remaining challenges, the work required to develop the application, and test the solution would require approximately six months to twelve months. That work should be done in-house by NPS and DTIC. Failure to pursue the integration of search engines into KMS systems will result in duplication of effort and loss of knowledge.

2. Further Research

The current application can also be used as a baseline for implementing a knowledge management system in classified environments. The Google Search Appliance (GSA) is a rack-mounted server that provides indexing functionality on a private intranet. The GSA system provides an interface and results similar to the public Google Search Engine.

Exploration and testing of internal search engines is another area requiring further research. As the search field matures, more robust alternatives will become available and begin to challenge the need for expensive search engine appliances. A solution such as the open source Apache Solr search library may fulfill the same requirements as the off-the-shelf product. Investment in this technology will require comparative performance testing, security verifications, and thorough risk assessments.

Finally, keeping up with current search technologies is important. For example, during the time of this research the Google Search Engine started to track real-time search via Twitter; however, this feature was removed in favor of its own Google+ social networking site. Also during the time of this research, DTIC implemented an internal search engine using Google search technology. The new internal Google search at DTIC reflects modernization that will help users find relevant information and is a big improvement over the old search system. Keeping up with the ever-changing market of commercial search will require constant research on the emerging technologies.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly 25*(1), 107–136.

Bell, D. K., & Jackson, L. A. (2001). Knowledge management: understanding theory and developing strategy. *Competitiveness Review 11*(1), 1–11.

Bixler, C. H. (2005). Developing a foundation for a successful knowledge management system. In *Creating the discipline of knowledge management: The latest in university research*. Burlington, MA: Elsevier Butterworth-Heinemann.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems* (107–117). Stanford, CA: Stanford University.

Carlson, N. (2011, January 5). Goldman: Facebook has 600 million users. *Msnbc.com.* Retrieved January 6, 2011, from http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/

Davenport, T. H., De Long, D. W., & Beers, M. C. (1998). Successful knowledge management projects. *Sloan Management Review 39*(2), 43–57.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, Mass: Harvard Business School Press.

Dell simplifies search with Google. (n.d.). *Dell*. Retrieved May 30, 2011, from http://content.dell.com/us/en/corp/d/press-releases/2007–11–27–00-google.aspx.aspx

Dierickx, I., & Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management Science 35*(12), 1504–1511.

Draeger, M. (2009). *Use of probabilistic topic models for search* (Master's thesis). Naval Postgraduate School, Monterey, CA.

Drucker, P. F. (1993). *The new society: The anatomy of industrial order*. New Brunswick, U.S.A.: Transaction.

Farrell, B. S., & Hutton, J. P. (2011). Department of Defense strategy for operating in cyberspace. Washington DC: Department of Defense.

Fox, V. (2010). *Marketing in the age of Google: Your online strategy is your business strategy*. Hoboken, NJ: Wiley.

Google Webmaster Tools. *Google*. Retrieved January 30, 2011, from http://www.google.com/support/webmasters/?hl=en

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal 17, Special Issue: Knowledge and the Firm*, 109–122.

Halawi, L. A., McCarthy, R. V., & Aronson, J. E. (2006). Knowledge management and the competitive strategy of the firm. *The Learning Organization 13*(4), 384–397.

Hawkins, B. (2009). *Developing a modular framework for implementing a semantic search engine* (Master's thesis). Naval Postgraduate School, Monterey, CA.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science 3*(3), 383–397.

Levy, S. (2010). Exclusive: How Google's algorithm rules the web. *Wired.com*. Retrieved September 07, 2011, from http://www.wired.com/magazine/2010/02/ff_google_algorithm/all/1

Love, P., Irani, Z., & Fong, P. (2004). *Management of knowledge in project environments*. Burlington, MA: Elsevier Butterworth-Heinenmann.

Liebowitz, J. (1999). *The knowledge management handbook*. Boca Raton, FL: CRC Press.

Massey, A. P., Montoya-Weiss, M. M., & O'Driscoll, T. M. (2002). Knowledge management in pursuit of performance: Insights from NORTEL networks. *MIS Quarterly 26*(3), 269–289.

Naval Postgraduate School Strategic Plan. (2008, January 1). Retrieved February 30, 2011, from http://www.nps.edu/About/NPSStratPlan.html

NPS Website. (n.d.). Naval Postgraduate School. Retrieved May 15, 2011, from http://www.cnic.navy.mil/Monterey/InstallationGuide/NavalPostgraduateSchool/index.htm

Nissen, M. E. (2002a). An extended model of knowledge-flow dynamics. *Communications of the Association for Information Systems*, 251–266.

Nissen, M.E. (2005). Toward designing organizations around knowledge flows. In K. Desouza (Ed.), *New frontiers in knowledge management*. New York, NY: Palgrave McMillan.

Nissen, M. E. (2006). *Harnessing knowledge dynamics*. Hershey: IRM Press.

Nissen, M. E., Kamel, M. N., & Sengupta, K. C. (2000, January-March). Integrated analysis and design of knowledge systems and processes. *Information Resources Management Journal 13*(1), 24–43.

Pentland, B. T. (1995). Information systems and organizational learning: The social epistemology of organizational knowledge systems. *Accounting, Management and Information Technologies 5*(1), 1–21.

Pollock, N. (2002). *Knowledge management and information technology (Know-IT encyclopedia)*. Fort Belvoir, VA: Defense Acquisition University Press.

Porter, M. E., & Millar, V. E. (1985, July/August). How information gives you competitive advantage. *Harvard Business Review 63*(4), 149–160.

Purcell, K. (2011, August 9). Findings: Search and e-mail remain the top online activities. *Pew Research Center's Internet & American Life Project*. Retrieved August 15, 2011, from http://pewinternet.org/Reports/2011/Search-and-e-mail/Report.aspx

Ryan, R. P. (2009, June 1). DTIC: Your key to DoD scientific & technical information. *DTIC Online*. Retrieved September 6, 2011, from http://handle.dtic.mil/100.2/ADA510750

Saviotti, P. P. (1998). On the dynamics of appropriability, of tacit and of codified knowledge. *Research Policy 26*, 843–856.

Silvi, R., & Cuganesan, S. (2006). Investigating the management of knowledge for competitive advantage: A strategic cost management perspective. *Journal of Intellectual Capital 7*(3), 309–323.

Snider, K. F., & Nissen, M. E. (2003). Beyond the body of knowledge: A knowledge-flow approach to project management theory and practice. *Project Management Journal 34*(2), 4–12.

Snyman, R., & Kruger, C. J. (2004). The interdependency between strategic management and strategic knowledge management. *Journal of Knowledge Management 8*(1), 5–19.

Tirpak, T. M. (2005). Five steps to effective knowledge management. *Research Technology Management. 48*(3), 15.

Winter, S. G. (2000, October/November). The satisficing principle in capability learning. *Strategic Management Journal 21*(10/11), 981–996.

Yin, R. K. (2009). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publications.

Zach, M. (1999). Managing codified knowledge. *Sloan Management Review 40*(4), 45–59.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Dan C. Boger
   Naval Postgraduate School
   Monterey, California

4. Mark E. Nissen
   Naval Postgraduate School
   Monterey, California

5. Eleanor Uhlinger
   Naval Postgraduate School
   Monterey, California