Planning Report 10-1

Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program

Prepared by:
RTI International
for

National Institute of
Standards & Technology

December 2010

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

December 2010

Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program

Final Report

Prepared for

Gregory Tassey
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 1060
Gaithersburg, Maryland 20899-1060

Prepared by

Brent R. Rowe
Dallas W. Wood
Albert N. Link
Diglio A. Simoni
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0211875

RTI Project Number
0211875

Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program

Final Report

December 2010

Prepared for

Gregory Tassey
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 1060
Gaithersburg, Maryland 20899-1060

Prepared by

Brent R. Rowe
Dallas W. Wood
Albert N. Link
Diglio A. Simoni
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

RTI International is a trade name of Research Triangle Institute.

Contents

Figures

Number **Page**

Tables

Number **Page**

EXECUTIVE SUMMARY

Information retrieval (IR) is the science and practice of matching information seekers with the information they seek. Internet users depend on IR tools each time they use a web search engine, such as those built by Google, Microsoft, and Yahoo!. However, a much larger IR industry exists that has improved the process of finding information within a single computer ("desktop search") or set of computers ("enterprise search"), as well as within large databases, such as library catalogs ("database search"). Further, IR techniques have been used to identify key links within, for example, legal records, genomics data, and spam.

IR tools are ubiquitous today, but in the early 1990s IR as a field was relatively immature with limited ongoing research. Around that time, Donna Harman, a NIST IR researcher, was assisting the Defense Advanced Research Projects Agency (DARPA) with its TIPSTER IR research program; Harman saw an opportunity to radically improve IR research by developing a NIST program that would leverage TIPSTER's accomplishments to provide new data and standard evaluation methodologies to IR researchers and create an objective competition among IR system creators.

The National Institute of Standards and Technology (NIST) established the Text REtrieval Conference (TREC) in 1992 to solve two major problems in IR. First, IR researchers lacked large data sets, referred to as "document collections" or "document sets," with which to test IR systems and techniques. Second, IR researchers lacked methodologies to facilitate the standardized comparison of IR systems. A lack of standard evaluation methodologies stemming from a lack of coordination resulted in duplicative research and information asymmetries. TREC offered the possibility to push IR researchers to invest at the socially optimal level.

In 2008, NIST contracted with RTI International to perform a retrospective economic impact assessment of NIST's TREC Program activities from 1991 to 2009. The study's goals were to quantify direct economic impacts, identify the role that TREC has played in improving IR products and services for IR researchers and end users, and provide recommendations to NIST regarding the future of TREC and any implications for NIST's strategic planning.

ES.1 TREC's Technical Accomplishments

TREC aimed to improve IR system evaluation through five primary mechanisms.

- ***Creation of new, larger test collections***. Before 1991, publicly available test collections were small, consisting of at most a few thousand documents and thus had a questionable application to real-world situations, in which an IR system might have been required to search hundreds of thousands of documents. During this period, the largest test collection popularly used contained about 12,000 documents (Voorhees and Harman, 2005). By contrast, the test collections used at the first TREC in 1992 contained approximately 750,000 documents. This represents an increase of over 80-fold in 1 year.

- ***Development of standardized IR evaluation methods***. The TREC Program developed and distributed "test collections" that IR researchers could use to evaluate their IR systems. Test collections consisted of three components: (1) a document collection, (2) a set of information needs or queries, and (3) a set of judgments indicating which documents are relevant for different queries. By using the same test collections and the same testing methodologies developed by TREC, IR researchers could also use these resources to compare the performance of their IR systems in a more systematic and standardized fashion.

  TREC initially created new test collections and evaluation methodologies for routing and ad hoc tasks and later expanded to support such novel IR areas as video, e-discovery, and spam. TREC quickly became the foundational place for developing new IR evaluation methodologies, testing new IR algorithms, and training new IR researchers. Over the past 18 years, TREC test collections and methodologies have been the de facto standard with which IR researchers publish research results that are defensible, comparable, and reproducible.

- ***Organization of annual IR research workshops***. TREC used the resources it developed to facilitate annual competitions and host an annual workshop (at which the results of the competition were presented), but IR researchers who did not participate in the TREC competitions and/or attend the workshops could still use the test collections and read the methodology descriptions and papers that were published after each workshop. The TREC Program had created an objective process for IR techniques to be compared and the results discussed and disseminated.

- ***Distribution of research results***. In addition to creating an objective process for evaluating IR techniques, the TREC Program facilitated the dissemination of the evaluation results. TREC participants can read all TREC papers at the conference and thereafter; TREC papers are released to nonparticipants approximately 6 months later.

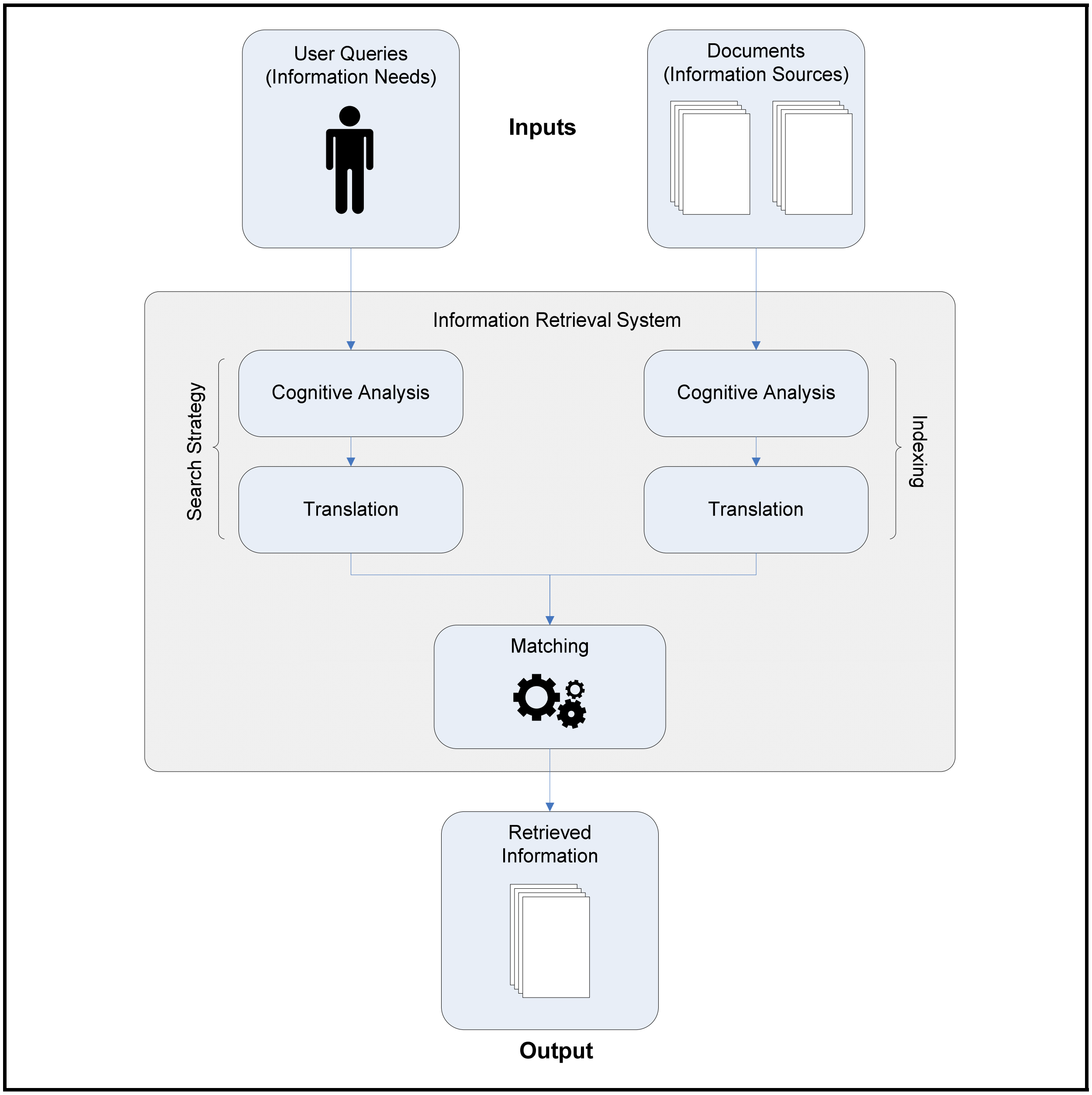
- ***Development of a model for other IR workshops***. Building off evaluation techniques and a format first designed by Cyril Cleverdon at the Cranfield College in the mid-1950s, TREC created the first of a growing number of programs and workshops that aim to facilitate rigorous, objective IR system evaluation.

ES.2 TREC's Impact on IR Research

If the goal of IR is to match people with the information they are seeking, then IR systems can be considered the methods used to achieve that goal. Figure ES-1 depicts a generalized illustration of how IR systems work. This diagram outlines the tasks an IR system must complete to retrieve the set of documents relevant to the needs of the information seeker; however, IR systems differ in the way and even the order in which these tasks are accomplished.

TREC has supported the improvement of IR systems primarily by increasing the document set sizes available to the IR research community by, providing a standard methodology, and by hosting annual IR research workshops (Table ES-1 provides an overview of significant events in the history of TREC). Evidence of the impact of TREC on IR system improvement has been presented throughout the academic literature. For example, Buckley, Singhal, and Mitra (1997) analyzed the performance of systems being evaluated in TREC's Ad Hoc and Routing tracks.

Figure ES-1. Tasks Performed by Typical IR Systems



Source: RTI.

They tested the IR system improvements made each year on the original test collection from 1992 and found that the majority of systems improved significantly between 1992 and 1997. With regard to their own research, they found that they were able to improve the Ad Hoc results of the SMART system by 20% to 25% annually, on average.

Table ES-1. Major TREC Historical Events

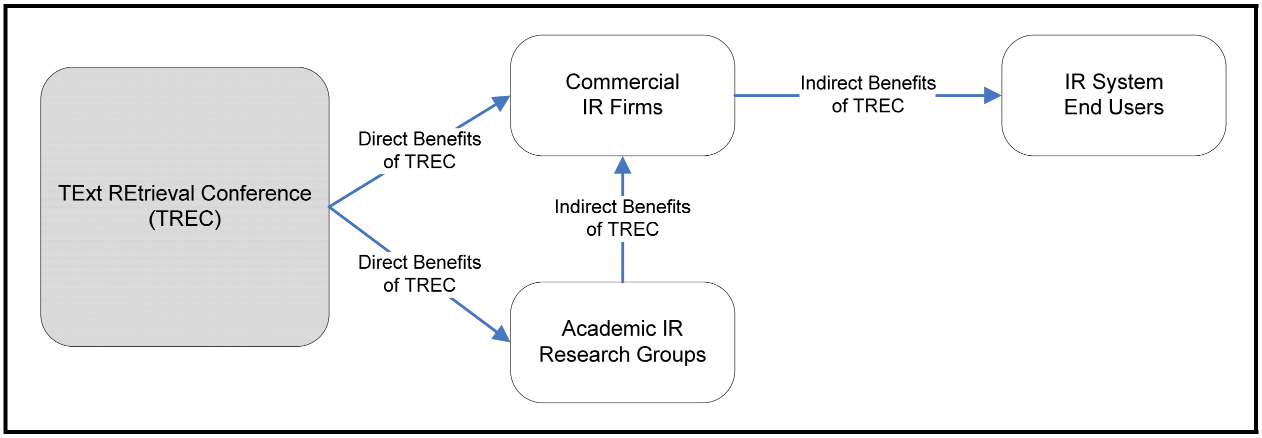
Year	Event Details
1990–1991	Charles Wayne (DARPA) asks Donna Harman (NIST) to help create a new, large test collection for the TIPSTER Program
1991	Donna Harman creates data set with DARPA funding and suggests releasing the data to the public and holding a workshop to compare researchers' IR techniques
1992	First TREC held in Rockville, MD
1993	TREC 2 provides first true baseline performance analysis
1994	TREC 3 expanded to include new tracks
1995	TREC 4 involves official track structure
2000	TREC 9 is first "all-track TREC"

Anecdotally, many individuals have also indicated that TREC drastically accelerated IR research. In 2008, Hal Varian, Chief Economist at Google, wrote about the impact of TREC, saying that TREC "revitalized research on information retrieval" (Varian, 2007). Varian talked further with RTI researchers during this study, praising NIST for both identifying a set of problems in the field of IR—data sets were too small and evaluation methodologies were not standardized—and for helping to solve these problems directly. However, Varian noted that TREC's influence has been felt less in the subfield of web search from the early 2000s forward, when TREC did not keep up with document set sizes being used by web search companies.

TREC's success in stimulating IR research has led to the creation of new IR evaluation programs, both in the United States and abroad, that have used TREC as a generic technology platform. These TREC-like projects, that either started as TREC tracks or were created to be similar to TREC, are now making their own contributions toward stimulating IR research (Robertson, 2008).

TREC has made significant contributions to the technology infrastructure supporting IR system development, the benefits of which flow directly or indirectly to a variety of stakeholder groups (see Figure ES-2). The direct beneficiaries are IR researchers in academic research groups and commercial firms; TREC's accomplishments improved both the efficiency and the effectiveness of their research and development (R&D) activities. R&D benefits that accrued to academic labs have also flowed indirectly to commercial firms through technology transfer and knowledge sharing. Improvement in the R&D of commercial IR firms led to improvements in the performance of IR systems commercialized into products and services. End users of these IR systems have also indirectly benefited from TREC through higher quality IR products and services.

Figure ES-2. Flow of TREC Benefits across Stakeholder Groups



Source: RTI.

TREC provided benefits at each stage of the IR system development process:

- Research results generated through TREC improve theoretical model development and adjustment.
- TREC's competitive workshops motivated improvements in system implementation by requesting that specific system goals be achieved and tested objectively.
- TREC's large test collections, evaluation paradigm, and workshops enabled more effective experimentation. TREC workshop participants received benefits greater than nonparticipants, particularly related to experimentation. Nonetheless, nonparticipants also benefited from the test collections, evaluation methodologies, and research results that were shared with the public.

NIST's involvement was crucial to turning the research conducted by DARPA into a workshop format that directly addressed the primary problems all IR researchers faced in the early 1990s. The importance of Harman's and NIST's accomplishments in creating TREC is maximized by the fact that no other organization or government agency (U.S. or foreign) was actively working to address these issues.

ES.3 Methodology for Estimating Net Economic Benefits of TREC

Economic costs and benefits were measured relative to a counterfactual scenario under which TREC was not created and IR R&D progressed in the absence of NIST, DARPA, and other public agency support of TREC. This counterfactual scenario established a framework for identifying, describing, and estimating the net benefits of TREC by making explicit the costs that would have been incurred in the absence of the Program and the benefits that would have been lost.

Benefits associated with TREC have accrued to *two* stakeholder groups—IR researchers (direct benefits) and end users (indirect benefits). By offering large test collections, standardized evaluation methods, and annual workshops and by disseminating new research, TREC reduced the cost of R&D conducted by IR researchers. In turn, these improvements have led to the development of higher quality IR systems that enable end users to satisfy their information needs in fewer search iterations.

By contrast, costs associated with TREC are incurred by *three* stakeholder groups—IR researchers (utilization costs), end users (utilization costs), and "investors" in TREC (creation/facilitation costs). TREC investment costs are defined as the costs incurred to create the test collections and evaluation methodologies for each track, facilitate the judging of TREC participant entries, host the annual workshop, and disseminate research results.

Data to inform this analysis were collected during semistructured interviews of IR researchers and experts and through a web survey fielded to IR researchers. The survey asked researchers in private, academic, nonprofit, and government organizations about how they used TREC resources, what value they place on the benefits of using those resources, and how they believe these resources affected the development of their IR systems.

A contingent valuation approach to soliciting estimates of the value of TREC was used to fully account for differences in estimation approaches. Contingent valuation is a survey technique that asks individuals directly to estimate how much value they receive from a particular resource. Although much more sophisticated techniques could have been used, such as discrete choice experiments, a variation of this simple open-ended approach was determined to be the most appropriate approach for this study because the magnitude of benefits and size of the sample were uncertain.

RTI received results from 404 respondents, of which 93 were based in the United States. Thirty percent of respondents (28 individuals) were located at U.S.-owned software or IR service companies and represent 58% of the total 2008 R&D expenditures by U.S. companies in IR. Approximately 66% of survey respondents (61 individuals) were employed by U.S. universities or academic research laboratories and represent 47% of total 2008 research expenditures of universities.

ES.4 TREC's Significance to IR Researchers

TREC's impact was most strongly felt by IR researchers—both TREC workshop participants and IR researchers who used TREC's test collections, evaluation methods, and research papers. Table ES-2 describes survey participants' use of TREC resources. The vast majority (over 95%) of respondents indicated that they had used TREC test collections, evaluation methods, and research papers at some point in time. However, the percentage who had attended a TREC workshop was somwhat lower—approximately 83% of respondents.

IR researchers also provided information on the use of TREC resources and the importance of TREC resources to those researchers who used them (Tables ES-2 and ES-3, respectively):

Table ES-2. Use of TREC Resources

	Number of Respondents Not Using Product or Service	Number of Respondents Using Product or Service	Blank	Total
Workshops	15	77	1	93
Test collection(s)	1	91	1	93
Evaluation method(s)[a]	3	89	1	93
Research papers	0	91	2	93

[a] Evaluation methods refer to the performance metrics (such as mean average precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

Table ES-3. Perceived Benefit of TREC Resources (as a Percentage of Users)

	Not Very Important for IR Research	Somewhat Important for IR Research	Very Important for IR Research
Workshops (N = 77)	6%	39%	55%
Test collection(s) (N = 91)	3%	12%	85%
Evaluation method(s)[a] (N = 89)	3%	15%	82%
Research papers (N = 91)	3%	31%	66%

[a] Evaluation methods refer to the performance metrics (such as mean average precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

- 55% of the researchers who attended TREC found the workshops to be very important to their research, while 39% found them to be somewhat important to their research;
- 85% and 82% of those who used the test collections and evaluation methods, respectively, indicated that they were very important to their research; and
- 66% of those who read the research paper rated them as very important, while 31% rated them as somewhat important.

Note that over 50% of those IR researchers who used any type of TREC resource found them to be very important to their research, and less than 6% found them to be not very important to their research.

Table ES-4 summarizes additional perceptions of the benefits of TREC. Of particular interest:

- 75% of survey respondents (70 individuals) have published papers using TREC test collection data,
- 47% (44 individuals) have used TREC test collections for tasks not studied at TREC (these tasks ranged from using TREC test collections to test algoritms to evaluating document download speeds),
- 71% (66 individuals) have used TREC evaluation methods with non-TREC test collections to evaluate the performance of IR systems,

Table ES-4. Use of TREC Resources

	Number	% of Total
Respondents publishing papers using TREC test collections	70	75%
Respondents using test collection data from TREC for tasks not studied at TREC	44	47%
Respondents using TREC evaluation methods to study the performance of an IR system using non-TREC test collections	66	71%
Respondents who have ever referenced a paper from TREC in a peer-reviewed journal article or a paper presented at conference	79	85%
Respondents who have referenced TREC papers in patent filing	13	14%

- 85% (79 individuals) have referenced a TREC paper in a peer-reviewed journal article they wrote or a paper they presented at a conference, and
- 14% (13 individuals) have referenced a TREC paper in a patent filing.

These responses suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study's economic benefits calculations.

Interviews also captured information about the improvement in human capital attributable to TREC. According to one researcher working for a web search services company, being able to hire individuals who have used TREC resources offers great value because they have hands-on experience with how IR systems work, making them more competent as employees.

Interviews and surveys indicate that TREC had a substantive impact on the interest of individuals to pursue graduate education in IR and the human capital that developed as a result of that education. Academic respondents were asked to evaluate how they percieved TREC's impact on the pursuit of and return on graduate education. We found that 67% of the 61 academic respondents believed that the number of individuals pursuing a doctoral or master's degree had increased as a result of TREC. Similarly, we found that 84% of academic respondents believed that participation in TREC workshops as a graduate student improved their employment prospects upon graduation.

ES.5 TREC's Impact on Web Search Results

Web search products, such as those develped by Google, Yahoo!, and Microsoft, have improved significantly over the last 10 years, and the results of this study suggest that TREC played a significant role in this improvement.

On average, IR researchers who responded to the survey estimated that end users of web search products would be able to fill an information need 215% faster in 2009 than in 1999 as a result of improvements in web search engine performance. In other words, information needs could be filled in approximately half the time with newer web search engines. Respondents, on average, also estimated that 32% of this improvement was enabled by TREC Program activities.

Under the counterfactual scenario that U.S. Internet users would have attempted to fill the same information needs using web search engines that did not experience any TREC-related improvement between 1999 and 2009, without TREC, U.S. Internet users would have spent up to 3.15 billion additional hours using web search engines between 1999 and 2009.

ES.6 Net Economic Benefits

Total extrapolated benefits were over $153 million for private, academic, and nonprofit organizations, and total unextrapolated benefits were $105 million. Inclusive of TREC Program and resource adoption costs, net extrapolated benefits were $123 million and net unextrapolated benefits were $75 million. Table ES-5 assembles the complete time series of quantified costs and benefits for the period between 1991 and 2009, and Table ES-6 provides performance measures.

After applying the Office of Management and Budget (OMB)-approved discount rate of 7%, the net present value (NPV) of extrapolated net benefits was $65 million and $37 million for unextrapolated net benefits. The benefit-to-cost (BCR) ratio, which is the ratio of the NPV of total benefits to that of costs, was estimated to be 5.07 for extrapolated benefits and 3.35 for unextrapolated benefits. In other words, for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers. The internal rate of return (IRR) was estimated to be over 250% for extrapolated benefits and over 130% for unextrapolated benefits.

ES.7 Summary Remarks

In 1990, IR research was largely being conducted by a handful of companies and universities whose techniques could not be compared easily. TREC transformed the IR landscape by providing large test collections at relatively low cost, developing and publishing robust evaluation methodologies and creating a competition through which researchers could objectively compare IR systems and discuss the results (successes and failures). TREC reduced the costs for IR research and opened the door to more robust IR system development.

TREC created an atmosphere where specific points of failure were discussed, which is uncommon even in academic circles. As a result, researchers in the private sector and in academia could see which IR techniques were most successful and integrate these findings into their products, thus benefiting these companies and their customers.

Table ES-5. Net Quantified Economic Benefits of TREC

Year	Total TREC Investment Costs (thousands $2009)	Total User Adoption Costs (thousands $2009)	Total Unextrapo-lated Benefits (thousands $2009)	Total Extrapolated Benefits (thousands $2009)	Net Unextrapo-lated Benefits (thousands $2009)	Net Extrapolated Benefits (thousands $2009)
1991	–$753	—	—	—	–$753	–$753
1992	–$713	–$19	$744	$1,177	$12	$445
1993	–$674	–$23	$3,060	$6,420	$2,363	$5,723
1994	–$1,522	–$25	$3,103	$6,403	$1,556	$4,857
1995	–$1,282	–$27	$3,231	$6,482	$1,922	$5,172
1996	–$2,129	–$29	$3,280	$5,894	$1,122	$3,736
1997	–$61	–$38	$4,307	$7,114	$4,208	$7,015
1998	–$1,739	–$42	$9,267	$14,037	$7,486	$12,255
1999	–$1,848	$51	$8,213	$11,878	$6,315	$9,980
2000	–$1,844	–$54	$8,182	$11,657	$6,285	$9,760
2001	–$1,544	–$68	$9,445	$12,938	$7,833	$11,326
2002	–$2,173	–$72	$6,778	$9,148	$4,533	$6,903
2003	–$1,880	–$73	$6,771	$8,972	$4,818	$7,020
2004	–$1,634	–$79	$6,461	$8,116	$4,748	$6,403
2005	–$2,143	–$94	$6,582	$8,659	$4,345	$6,423
2006	–$1,788	–$87	$6,484	$8,671	$4,609	$6,796
2007	–$1,668	–$78	$6,404	$8,473	$4,658	$6,727
2008	–$1,982	–$46	$6,387	$8,477	$4,359	$6,450
2009	–$1,671	–$66	$6,387	$8,477	$4,649	$6,740
Total	–$29,046	–$970	$105,084	$152,994	$75,068	$122,978

Note: All dollar values were inflation-adjusted to 2009 dollars using the Consumer Price Index, U.S. City Average for all Items.

Table ES-6. Performance Measures

Measure	Value
Unextrapolated Performance Measures	
Total quantified benefits (Discounted, Base Year = 1991)	$53,267,846
Total quantified costs (Discounted, Base Year = 1991)	−$15,916,193
Net present value of net benefits (NPV) (Base Year = 1991)	$37,351,653
Benefit-to-cost ratio (BCR)	3.35
Internal rate of return (IRR)	128%
Extrapolated Performance Measures	
Total quantified benefits (Discounted, Base Year = 1991)	$80,655,082
Total quantified costs (Discounted, Base Year = 1991)	−$15,916,193
Net present value of net benefits (NPV) (Base Year = 1991)	$64,738,889
Benefit-to-cost ratio (BCR)	5.07
Internal rate of return (IRR)	250%

Note: All dollar values were inflation-adjusted to 2009 dollars using the Consumer Price Index, U.S. City Average for all Items. Benefits and costs were discounted using the 7% real social discount rate recommended by OMB.

In addition to the retrospective findings of this study, the future of TREC was investigated. Many industry and academic experts who were interviewed noted that TREC's value will continue to be worth the investment moving forward if NIST focuses on solving new problems—identifying new subfields of IR that need new data and evaluation methodologies. Further, several trends in survey responses are noteworthy: 37 survey respondents indicated that TREC should expand into new tracks, 20 said TREC should develop new evaluation methods, and 17 said TREC should develop new data sets. Common suggestions were the following:

- Focus on more user behavior data (e.g., social data, Twitter, geographically based) to improve on the Interactive track.
- Continue to look at multimedia search techniques (e.g., pictures, video).
- Expand into more focused search areas (e.g., chemistry, drug design, evidence-based medicine).

More broadly, several respondents suggested that TREC should work with industry to increase their participation in the TREC workshops, as well as to solicit data that they might allow the TREC audience to use, thus increasing the usefulness of TREC results. One respondent suggested that more time should be spent discussing the improvements in search techniques, instead of spending so much time talking about the methodologies used to compare system results in the TREC competition. Another respondent suggested that TREC should try to partner with a leading journal to expand the dissemination of TREC results farther and/or collocate with another conference to increase participation.

TREC has proved to be a success, and our research suggests that TREC will continue to be highly beneficial moving forward as long as TREC continues to focus on its key benefits—providing new data sets, developing methodologies to approach new IR topics, and providing a venue for IR researchers in all sectors to compare their IR techniques objectively and to share both successes and failures in a safe environment.

1. INTRODUCTION

Information retrieval (IR) is the science and practice of matching information seekers with the information they are seeking. Internet users use IR-based tools each time they use a web search engine, such as those built by Google, Microsoft, and Yahoo!. However, a much larger IR industry exists that developed products and services for businesses and government alike over the past 40 years. IR techniques have been used to improve the process of finding information not only on the web, but also within a single computer ("desktop search") or set of computers ("enterprise search"), as well as within very large databases, such as libraries ("database search"). Further, IR techniques have been used to identify key links within, for example, legal records, genomics data, and spam. Improvements in IR techniques result in improved efficiency of business operations as well as an increase in the level of general information awareness by government, businesses, and consumers.

In the early 1990s, IR as a field was relatively immature with limited ongoing research. One of the primary barriers facing IR researchers during this period was a lack of large data sets, referred to as "document collections" or "document sets," with which academic researchers and companies could test novel IR techniques. Before 1991, document collections were small, consisting of only a few hundred documents, and thus had a questionable application to real-world situations, in which an IR system might have been required to search several thousand documents. The second major barrier to IR was a lack of methodologies to facilitate the standardized comparison of IR systems.

The National Institute of Standards and Technology (NIST) established the Text REtrieval Conference (TREC) in 1992 to solve these two major problems. NIST first began by working with the Department of Defense's Defense Advanced Research Project Agency (DARPA), which had funded NIST IR researcher Donna Harman to create a new, very large test collection for its use. Harman convinced DARPA to make this new test collection available to a wider audience, and NIST set out to establish a competition that would allow IR researchers to test their systems using this new collection.

In order to facilitate this competition, the TREC Program developed and distributed "test collections" that IR researchers could use to evaluate their IR systems. Test collections consist of three components: (1) a document collection (also known as a "document set" or "data set"), (2) a set of information needs or queries, and (3) a set of judgments indicating which documents are relevant for different queries. Test collections are important for IR research because researchers can use them to test the IR systems they develop and determine how well they perform at identifying documents in a particular collection that are relevant to particular queries.

By using the same test collections and the same testing methodologies developed by TREC, IR researchers could also use these resources to compare the performance of their IR systems in a more systematic and standardized fashion. TREC used the resources it developed to facilitate annual competitions and host an annual workshop (at which the results of the competition were presented), but IR researchers who did not participate in the TREC competitions and or attend the workshops could still use the test collections and read the methodology descriptions and papers that were published after each workshop. The TREC Program had created an objective process for IR techniques to be compared and the results discussed and disseminated.

As a result of TREC, the size of document sets available to IR researchers to test their IR systems increased from an average of approximately 12,000 documents in 1991 to approximately 750,000 documents in 1992 with the first TREC document set, representing an increase of over 80-fold in 1 year. In subsequent years, TREC created new test collections and evaluation methodologies for such novel IR areas as video, e-discovery, and spam. TREC quickly became the foundational place for developing new IR evaluation methodologies, testing new IR algorithms, and training new IR researchers. Further, over the past 18 years, TREC test collections and methodologies have been the de facto standard with which IR researchers publish research results that are defensible, comparable, and reproducible.

In 2008, NIST contracted with RTI International to perform a retrospective economic impact assessment of NIST's TREC Program activities from 1991 to 2009. The study's goal was to quantify direct economic impacts, identify the role that TREC has played in improving IR products and services for IR researchers and end users, and provide recommendations to NIST regarding the future of TREC. This report is the study's final deliverable.

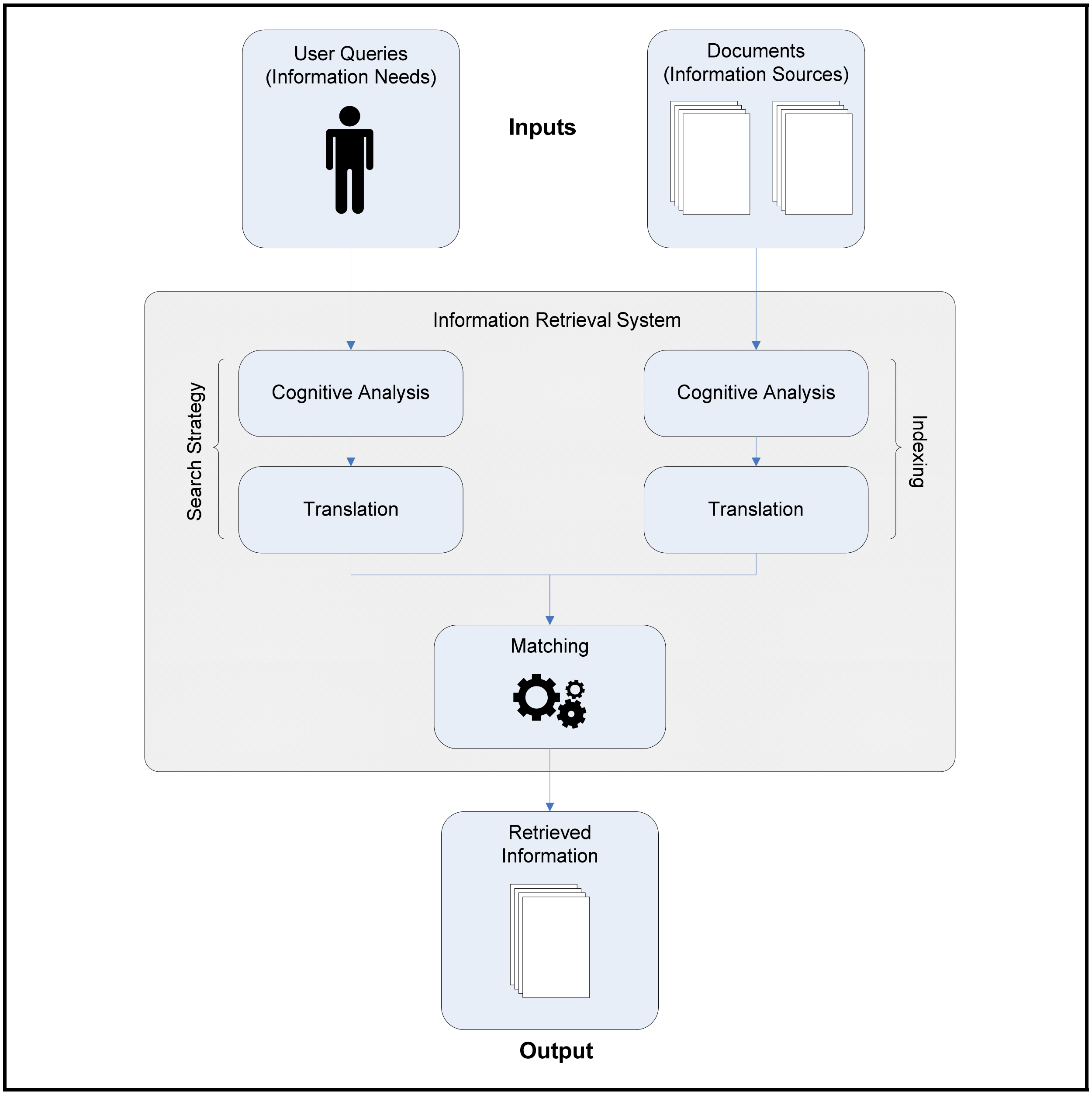
This introductory section offers a discussion of key IR concepts and the rationale for NIST's involvement in IR, both of which are necessary to fully contextualize TREC's scientific accomplishments and economic impacts.

1.1 How Do IR Systems Work?

If the goal of IR is to match people with the information they are seeking, then IR systems can be considered the methods used to achieve that goal. Figure 1-1 depicts a generalized illustration of how IR systems work. Although the order in which these tasks are completed and the method used will differ depending on the IR system in question, this simplified diagram shows the tasks that must be completed to retrieve the set of documents relevant to the needs of the information seeker.

An IR system requires two basic inputs: a collection of source documents and a query. First, the sources containing information that users would like to find may be books, websites, scientific journal articles, or any number of different media. Regardless of the information's medium, these information sources are commonly referred to as "documents" in the field of IR.

Figure 1-1. Tasks Performed by Typical IR Systems



Source: RTI.

These documents are organized into an index, which serves as a guide for locating documents that best meet a user's information need. The **indexing** process comprises two steps:

- Step 1. Cognitive analysis: the determination of what information the documents contain and why users may find it useful.
- Step 2. Translation: the translation of the results of the cognitive analysis into a vocabulary or "index language" that can be searched according to the information desired (Lancaster, 1979).

Once indexing is complete, the raw documents are organized and stored where they can be retrieved when needed. The index itself is likewise organized and stored so that it can be used to identify and locate documents efficiently inside the collection based on their content.

The second input into an IR system is the user's information need, commonly referred to as a query. When a user has a query (e.g., "articles on Benjamin Franklin's light bulb"), she inputs it into the IR system, and the IR system uses a **search strategy** that comprises two steps that are similar to the indexing process:

- Step 1. Cognitive analysis: the determination of what the user's information needs are so those needs can be met.
- Step 2. Translation: the translation of the information need into the "index language" so that documents can be located to meet that need (Lancaster, 1979).

After both the information sources and the user's information need have been analyzed and translated into the same format, the IR system matches indexed documents with the criteria laid out by the search strategy. Although the techniques used to complete these steps can vary significantly, the end goal of the process is to retrieve a set of documents that meet the user's information needs.

1.2 NIST's Involvement in IR: Market Barriers

NIST's involvement in IR through the creation of the TREC Program provided a solution to a market failure that had been observed. Both the inability of IR researchers to appropriate (or capture) the total benefits of their investments and the lack of coordination among IR researchers were preventing IR research from progressing at a socially optimal level.

Underinvestment in technology development can occur when conditions exist that prevent firms from fully realizing or appropriating the benefits created by their investments, causing firms to view prospective investments as having expected rates of return below the firm's minimum acceptable rate of return (hurdle rate). The resulting level of private investment leads to a suboptimal social level of use of the technology. Although firms may recognize that there are spillover benefits to other markets or consumers, they are likely to ignore or heavily discount these benefits. Infratechnology research to support the development of shared data and methodologies, such as the efforts of TREC, are all paradigmatic examples of cases where private returns to investment can be less than both social returns and private hurdle rates. As a result, those activities are frequently supported by government activities; without them, investment will be suboptimal from a social perspective.

The development of TREC offered to solve the problem of suboptimal investment that resulted from appropriability issues and lack of coordination. TREC facilitated the creation and distribution of new, larger, and more varied document sets that few private, public, or academic organizations would create on their own because they could not appropriate the full value of their investment. Further, TREC developed common evaluation methodologies to analyze IR system performance and compare systems; this helped solve the problem of a coordination failure that existed as a result of asymmetries in incentives and information between market participants. Firms acting in their self-interest often invest in standards or technologies that are not optimal for the industry as a whole, or in competing implementation procedures developed independently that may not interoperate. It has been shown that coordination activities can lower the cost of development and increase the quality of the technologies. Prior to TREC, a handful of researchers had the resources to conduct robust research but without common methodologies, their claimed results could not be verified. TREC lowered the barrier to entry and provided a means to compare IR system results objectively.

The TREC Program also created positive externalities for IR research. The more organizations that participated in TREC, the more everyone benefited. TREC's workshops provided a way in which the IR research community could benefit from each others' successes and failures. And beyond TREC, the test collection materials (document sets, queries, and relevance judgments) were used for additional research that was often shared with the research community through other forums (e.g., ACM SIG-IR). Because common data sets and evaluation methodologies were used, information asymmetries were eliminated.

1.3 Study Overview and Report Organization

The purpose of this study was to analyze the net benefits of the NIST TREC Program. This study involved conducting background research on technical and economic contributions of TREC, including a set of scoping interviews with IR experts; developing a robust methodology for estimating the economic costs and benefits of TREC and additional qualitative metrics; and fielding a survey instrument that was completed by over 350 IR researchers. This report, summarizing the findings, is organized as follows:

- Section 2 presents a historical overview of the IR field, including the creation of the TREC Program.
- Section 3 reviews TREC's technical contributions.
- Section 4 discusses the stakeholder groups affected by TREC.
- Section 5 presents the methodology for conceptualizing economic impacts and quantifying economic benefits.
- Section 6 presents the analytical results from economic modeling.
- Section 7 concludes with remarks about the future of the TREC Program and broader implications for NIST's strategic planning activities.

2. TREC IN CONTEXT: IR RESEARCH FROM 1945 TO 2009

Innovation in IR systems between 1945 and 1970 was followed by 20 years of little progress. A lack of standardized test collections, including larger document sets, became a significant barrier to technological advancement. Around 1990, as this problem became increasingly apparent, DARPA and NIST created TREC as a means for providing resources to foster IR research. This section provides historical context (see Figure 2-1) and explains the motivation for creating TREC by describing how the lack of both standardized evaluation methodologies and large test collections impeded the transfer of technology from the laboratory to the marketplace.

2.1 Origins of IR Research

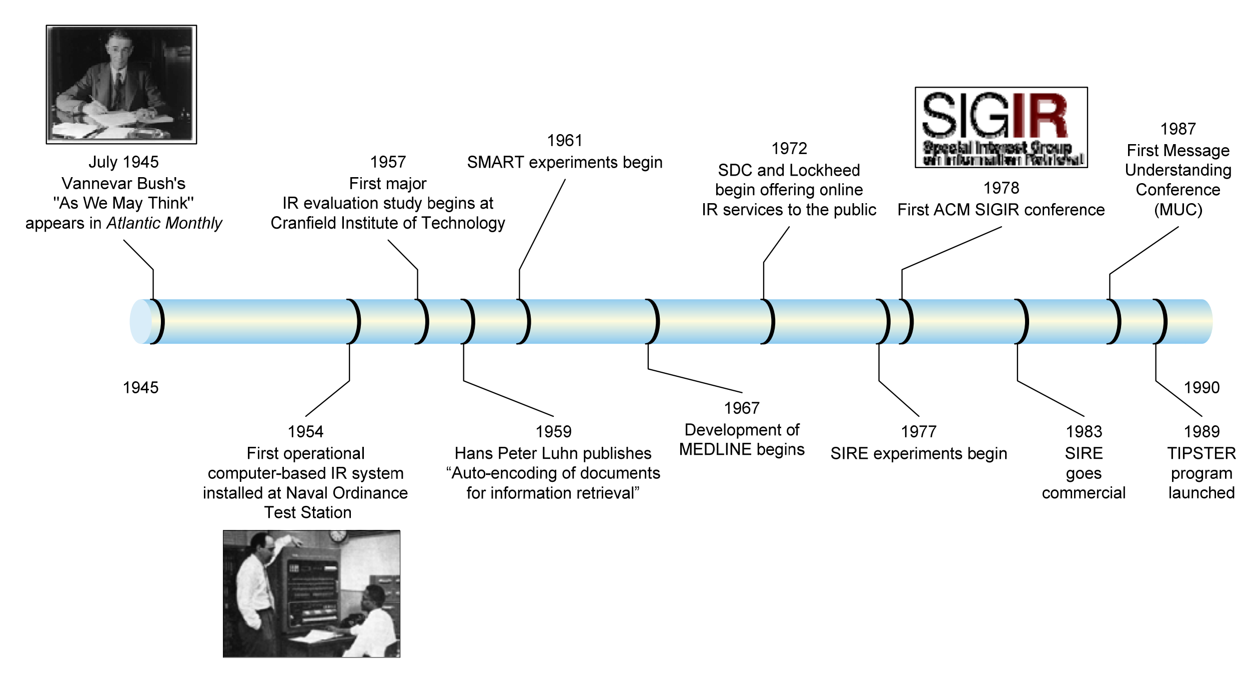
Prior to the 1950s, all of the tasks of an IR system had to be completed manually without the aid of computers. Users wanting to identify documents, books, or articles that addressed certain topics consulted printed indexes. These indexes provided bibliographic data (such as citations) that would point to the locations of documents on a particular topic. The contents of documents listed in these indexes were typically described using short descriptions called subject headings (e.g., a book or article on France in the Middle Ages may fall under the subject heading "France, History, Middle Ages") (Meadow, Boyce, and Kraft, 2000).

This process is still familiar to individuals who have used a library card catalog (potentially a shrinking number of people as more libraries move their records online). A user with a need for information on a particular subject or for a work by a particular author can go to a library and search a card catalog for books or periodicals to satisfy those needs. This catalog provides relevant index listings and locations for documents in the library. Typically card catalogs are organized by author, title, and subject.

In the case of scientific literature, professional societies and other organizations published printed indices to serve as guides to recent research for scientific investigators. These indices were created by professional indexers who analyzed the content of documents to assign them relevant subheadings. New indices of the scientific literature were published routinely and made available to libraries and laboratories. Accumulations of these indices were also printed regularly to help researchers search the literature across multiple years more easily (Meadow, Boyce, and Kraft, 2000).

Although historical methods such as card catalogs and printed indexes of scientific literature provided individuals with a structured way to find information, they were often difficult and time consuming to use. Finding relevant documents for a request could also be complicated if an information seeker's needs could not be translated easily into existing subject headings. As a result, the assistance of a research librarian or information specialist was often required to complete extensive literature searches.

Figure 2-1. Major Developments in IR Research: 1945–1990



The challenges associated with manual IR methods accumulated at an accelerated pace during and after World War II when the number of scientific articles being published increased significantly. In his much celebrated 1945 *Atlantic Monthly* article "As We May Think," Dr. Vannevar Bush, Director of the U.S. Office of Scientific Research and Development during World War II, argued that new IR systems had to be developed to meet this growing challenge of having too much data and to facilitate the search of large document collections. Bush complained that existing index systems were too "artificial," and he called for the application of computers to improve retrieval methods (Bush, 1945).

After World War II, a great deal of research was aimed at addressing Bush's concerns about existing index systems. Among the earliest of the new postwar indexing systems was the Uniterm system created by Dr. Mortimer Taube. This system indexed documents using single key words (called Uniterms) instead of complete subject headings, as had been the case in previous indexing systems (Meadow, Boyce, and Kraft, 2000).

To conceptualize how the Uniterm system operated, consider the following example. Suppose that a user wanted to locate material on Medieval French history using a printed index organized by subject heading. To find this material, she would consult the index and look for a subject heading like "France, History, Middle Ages." However, if the indexer who created the index did not create this or a similar subject heading, or if she applied different terminology, then the user would have a much harder time finding the material she wanted. For example, if the only similar subheading available were "French, History," then the user would likely have to scan many irrelevant documents before finding one specifically on Medieval French history. The inflexibility of these types of systems was an inherent problem for retrieving information efficiently (Meadow, Boyce, and Kraft, 2000).

Taube realized that "France," "History," and "Middle Ages" were separate concepts being combined and, as such, that a more flexible system would allow individuals to search using a combination of these concepts as needed, rather than try to predict their information needs. Under the Uniterm system, a book on French history in the Middle Ages would be indexed on three separate cards: the card listing all documents on "France," the card listing all documents on "History," and the card listing all documents associated with "Middle Ages." An individual wanting to find only books on French history during the Middle Ages would obtain all three of these cards and identify the documents that are common to all three (Meadow, Boyce, and Kraft, 2000).

IR researchers of the early 1950s realized that matching documents using index systems like Uniterm was a process that could be described in mathematical terms that computers could understand using the algebra of sets developed by George Boole (Meadow, Boyce, and Kraft, 2000). This realization implied that computers could be used to perform the relatively time-consuming task of matching documents to key word queries quickly.

The first demonstration and implementation of such a computer-based literature search system was at the Naval Ordinance Test Station (NOTS) in 1954. Library staff at NOTS worked with Taube to install an "in-house" computerized search service based on Taube's Uniterm indexing system (using an IBM 701 computer) that offered NOTS researchers access to 15,000 bibliographic records indexed by Uniterms (Bourne, 1999).

2.2 Cranfield's Evaluation of IR Systems

The proliferation of new indexing systems in the early to mid-1950s naturally led to the question of which system was most useful for IR. In 1957, Dr. Cyril Cleverdon of the Cranfield College of Aeronautics (now Cranfield University) in the United Kingdom led a study that attempted to compare the performance of four index systems.[1]

Cleverdon studied how well indexing systems performed IR using a collection of 100 source documents chosen from the field of aeronautics. Three human indexers were chosen for each system to index each source document 5 times (spending 2, 4, 8, 12, and 16 minutes per document). This yielded a total of 6,000 indexed items (100 documents X 3 indexers X 4 index systems X 5 times).

[1] The four indexing systems were (1) the Uniterm system, (2) the Universal Decimal Classification (a hierarchical library classification), (3) an alphabetical subject catalogue (subject headings expressed as phrases), and (4) a faceted classification scheme (allowing the construction of complex categories by combining elements from different facets) (Robertson, 2008).

Next, technical experts from outside the study were asked to analyze each of the test documents and to formulate queries to which the given document would be relevant. A total of 400 queries were formulated.

Indexers then attempted to identify the documents that were relevant to these 400 queries from the 6,000 indexed items. This was done 3 times to see if more relevant documents were retrieved as the indexers became more experienced (Chowdhury, 2004). The overall performance of each system during these tests was measured using two metrics:[2]

1. **Recall:** the fraction of all documents in a collection relevant for a particular query that are actually retrieved by an IR system.

$$\text{Recall} = \frac{\left|\{\text{Relevant documents}\} \cap \{\text{Retrieved documents}\}\right|}{\left|\{\text{Relevant documents}\}\right|}$$

2. **Precision:** the fraction of documents retrieved by an IR system that are actually relevant for a particular query.

$$\text{Precision} = \frac{\left|\{\text{Relevant documents}\} \cap \{\text{Retrieved documents}\}\right|}{\left|\{\text{Retrieved documents}\}\right|}$$

A perfect IR system would retrieve all of the documents that are relevant for an individual's query (resulting in a recall score equal to 1) and no documents that were irrelevant (resulting in a precision score equal to 1).

In reality, actual IR systems typically face a trade-off between these two measures of performance because if an IR system simply retrieves more documents, it is likely to increase recall (by retrieving more documents, one would more likely retrieve more that are relevant). But retrieving more documents is likely to reduce precision: by simply retrieving more documents, one is also increasing the chances of retrieving irrelevant documents.

Surprisingly, the results of Cleverdon's study indicated that there was little difference in the performance of the four indexing systems. For example, the recall ratios for these systems ranged from 74% to 82%. The vast majority of retrieval errors that did occur were determined to be caused by human mistakes in indexing and searching. It was revealed that fewer of these errors occurred as the indexers were given more time to index the documents and as the indexers gained more experience in conducting the searches (the success rate in the third round of searches was 3% to 4% higher than in the second round) (Chowdhury, 2004). These results were important for subsequent IR research because they clearly illustrated which factors affected the performance of IR systems and which did not (Meadow, Boyce, and Kraft, 2000).

[2]The recall and precision performance equations use set notation. The symbol ∩ is used to represent the intersection of two sets, in this case, the intersection of two sets of documents (the set of relevant documents and the set of retrieved documents).

In addition, Cleverdon's study was important because it was the first major IR evaluation study conducted, and it developed a methodology that could be applied successfully to the evaluation of any IR system. The components of this methodology can be summarized as

- a collection of documents to be searched (often called a "test collection," "document set," "document collection," or "corpus"),
- a series of questions or "queries" answered by the documents in this collection,
- an IR system to match the information needs with the information in the document collection, and
- performance measures based on relevance judgments (Meadow, Boyce, and Kraft, 2000).

This basic methodology would serve as the foundation for many subsequent evaluation studies. However, the test collections used in these studies were relatively small compared to the expected demand of an applied IR system. This limitation created problems for researchers in later years as they attempted to transfer the retrieval techniques they developed into commercial applications.

2.3 Further Developments in Using Computers to Automate IR Tasks

IR research during the mid-1950s largely focused on using computers to automate the task of matching words used in a search statement with key words listed in an index file. It was widely assumed that the time-consuming task of assigning key words to information items within a document collection was a job that only trained human indexers could complete.

Hans Peter Luhn of IBM argued against this conventional wisdom in a series of influential papers published between 1957 and 1959.[3] He suggested that statistical techniques could be applied so that computers could handle the "intellectual work" of analyzing the content of written text. Specifically, Luhn proposed automatic indexing methods based on how frequently a key word occurred inside the text and where it was located relative to other words (Salton, 1987).

Luhn's theories were expanded by subsequent researchers seeking to automate the indexing process. One of the most prolific of these later researchers was Dr. Gerard Salton of Harvard University and, later, Cornell University. In 1961, Salton began a long-running theoretical and experimental program to explore and evaluate various indexing and retrieval techniques. The computer programs that Salton and his colleagues created to facilitate this research were collectively known as SMART.[4] These programs were used to systematically

[3]This is according to Salton (1987) though no list of these "influential papers" was in the paper.
[4]The SMART acronym took on several meanings over the course of Salton's experiments. These meanings included the System for the Mechanical Analysis and Retrieval of Text, System for the Manipulation and Retrieval of Texts, and Salton's Magical Automatic Retriever of Text (Bourne and Hahn, 2003).

- accept search queries posed in natural English;
- perform fully automated indexing of documents (rather than indexing documents by selected index terms as Luhn suggested, SMART preserved all terms in a document);
- match analyzed search statements and the contents of documents; and
- retrieve documents most similar to the queries presented, ranked in terms of their likely relevance (Chowdhury, 2004).

The SMART system was later described by IR researcher Harold Barko in 1985 as "one of the most automatic retrieval systems being used experimentally" during this period (Bourne and Hahn, 2003).

To evaluate the performance of the IR techniques he was developing, Salton used a methodology similar to that developed by Cleverdon in the late 1950s. However, like most other IR evaluation studies conducted during this time, the test collections Salton used were very small, meaning that they were measured in tens or hundreds of documents (Robertson, 2008). As a result, some IR researchers were skeptical of how the techniques he developed would perform on large document collections (Bourne and Hahn, 2003).

2.4 Introduction of Online Information Services

Throughout the 1950s and early 1960s, several research institutions and government agencies began developing computerized "in-house" IR systems like the one implemented by Taube at NOTS in 1954. However, a significant disadvantage of these systems was that none of the computers running them were devoted exclusively to IR; they were used instead for completing multiple unrelated calculations and tasks. As a result, queries were processed along with other tasks through "batch processing." *Batch processing* is where tasks are stored in a queue while the computer system is busy processing other tasks. Although scheduling algorithms differ, eventually all the tasks in the queue are processed one after the other.

The following problems were associated with processing queries through batch processing:

- **Time delays:** Because searches were not processed as they were received and because computers' processing times were generally very slow, a user would typically have to wait hours or even days to obtain the results.
- **"One-chance" searching:** A user had to think of all possible search approaches in advance so that he or she could construct a search strategy that, when matched with the database, was likely to retrieve all the relevant literature. These systems were not interactive, and a user could not alter strategies easily based on search results (Lancaster, 1979).

By the mid-1960s, mainframe computers had enough speed and memory both to carry out their routine research tasks and to perform IR requests simultaneously. Researchers were now able to perform interactive searches in minutes (versus searching and waiting hours or days) through terminals that were connected directly to the mainframe. In addition, developments in network technology enabled these terminals to be connected to a mainframe computer over long distances through telephone lines, thus making an individual IR system available to users across the country (Meadow, Boyce, and Kraft, 2000).

The first major IR system of this type (called an “online” system) was the Medical Literature Analysis and Retrieval System Online (MEDLINE) (Chowdhury, 2004). In 1967, the National Library of Medicine (NLM) contracted with Systems Development Corporation (SDC), which had developed a “general purpose” IR system called On-line Retrieval of Bibliographic Text, to install a system that would allow medical institutions across the country to search NLM’s electronic database of medical journal citations.

In 1970, a preliminary system, called AIM-TWX, was launched. Its name was derived from the fact that the system provided medical institutions access to the Abridged Index Medicus (AIM) database over the Teletypewriter Exchange Network (TWX). The AIM-TWX system was expanded between 1971 and 1973 to include a larger collection of citations and renamed MEDLINE.

NLM initially only offered access to this database to noncommercial medical institutions. Soon, however, there was growing demand from commercial organizations (such as the Pharmaceutical Manufacturers Association) for access to a system that provided similar services. This swell of demand convinced SDC that a profitable business model could be based on providing online IR services to the public.

Under this model, SDC would contract with entities creating large databases and sell the service of searching those databases to customers. When SDC launched its service business in 1972, it provided customers access to MEDLINE’s medical journal database, the Education Resources Information Center (ERIC) database (which contained citations for journal and nonjournal education literature), and the Chemical Abstracts Condensates (CHEMCON) database (which contained citations for chemistry-related journals from around the world) (Bourne and Hahn, 2003).

At approximately the same time, Lockheed Martin began offering online services using an IR system that it developed called DIALOG. In 1971, DIALOG indexed the ERIC database. In 1972, DIALOG’s database offerings included the ERIC database and the National Technical Information Service (NTIS) database of government-sponsored research.

The type of information that these online information services provided to users was largely determined by the hardware capability of the online systems themselves. Limited disk space on mainframe computers mandated that only citations of articles (and not the full text themselves) could be stored. In addition, early terminals were paper teletypewriters that would print out the results of the search as they came in. This also made the reporting of bibliographic data such as citations more economically feasible than the reporting of full-text articles (Schatz, 1997).

As time passed and computers became more powerful, the scale of documents that could be indexed and searched became greater. In the early 1970s, mainframe computers could store full-text documents, rather than just citations. The first large-scale demonstration of the practicality of using IR systems to search and retrieve full-text documents was provided by Mead Data Central when, in 1973, it launched the Lexis system to retrieve full-text U.S. court records for legal professionals. In 1979, Mead introduced a full-text news and business-information service called Nexis (Schatz, 1997). Table 2-1 provides an overview of the three pioneering online IR service companies discussed above.

2.5 Lack of Large Test Collections and Barriers to Commercialization

The number of companies offering IR services continued to grow throughout the 1970s. By 1975, as many as 300 public access databases were available from a range of different vendors (Chowdhury, 2004). However, the search methods these companies used typically did not employ the statistical techniques that Salton and other IR researchers developed for automating indexing and other tasks. This was largely because the companies were skeptical that statistical techniques would not be able to work on large document collections, and large test collections were typically not publically available to prove otherwise.[5]

In 1975, Dr. Karen Spark Jones at the Computer Laboratory at Cambridge University and Keith van Rijsbergen of the University of Glasgow proposed the creation of a larger, "ideal" test collection to address this need. She reasoned that one of the major difficulties in building a large test collection is that it becomes increasingly hard to determine how many of the documents in a particular collection are relevant for a particular query, which is essential to calculating the recall performance measure. This is because, at some point, a document collection becomes so large that it is not feasible for subject matter experts to analyze each of the documents to determine which are relevant. Spark Jones proposed a solution to this problem called pooling.

The pooling process requires using a variety of IR systems to search the large document collection and provide a ranked list of documents relevant to a particular search. The first 100 ranked items from each system are then "pooled" and analyzed for relevance to the initial query. After analysis, the portion of these 100 documents that are relevant is used to indicate what documents in the collection are relevant to a particular query. Because of a lack of funding, Spark Jones's project for creating an ideal test collection never came to fruition, but TREC would later use the pooling method she developed when creating its document collections (Robertson, 2008).

[5] Based on interviews with IR researchers active during this time, it is clear that a select few researchers had access to larger, proprietary data sets that could be used for IR system evaluation. For example, at the University of Massachusetts, Dr. Bruce Croft and others had access to a relatively large Westlaw data set. However, because such data sets were not publicly available, there could be no verification of results.

Table 2-1. Pioneering Online IR Services Companies

Company Name	Year Services First Publicly Offered	IR System(s)	Initial Service Offerings
Systems Development Corporation (SDC)	1972	ORBIT	Bibliographic search of MEDLINE’s medical journal database; ERIC, the educational literature database; and the CHEMCON, the chemistry journal database
Lockheed Martin	1972	DIALOG	Bibliographic search of ERIC, the educational literature database, and NTIS, the government-sponsored research database
Mead Data Central	1973	Lexis	Full-text search of U.S. court records and legal documents
	1979	Nexis	Full-text search of news articles and popular periodicals

Despite the reluctance of commercial online IR service providers to adopt statistical techniques developed in academic labs, IR research continued. In 1978, the Association for Computing Machinery (ACM) Society’s began an annual series Special Interest Group for Information Retrieval (SIGIR) conferences. According to IR researchers, the SIGIR conference provided the best mechanism for sharing IR research results of search techniques and evaluation methods throughout the late-1970s and 1980s.

In 1977, a major project began at Syracuse University to design an online bibliographic retrieval system that used techniques pioneered by Salton’s SMART experiments. The project was called the Syracuse Information Retrieval Experiment (SIRE). The SIRE IR system was similar to SMART in that it also aimed to use the full text of a document for indexing and statistical methods to offer users ranked results of their search. After the system had been developed over several years, a version was commercialized in 1983 by KNM, Inc. partly in response to individuals who claimed that IR research was not generating practical applications for commercial systems (McGill et al., 1976; Fox and Koll, 1988).

Even though the SIRE system was successfully commercialized, the lack of large test collections still posed a barrier to conducting and commercializing IR research. In the late 1980s, Salton lobbied the senior management of DAILOG[6] to adopt some of the retrieval techniques he had developed as part of his SMART system. However, DIALOG was reluctant to try Salton’s methods because they had not been evaluated using large data collections. As a result, they were unsure whether the benefits of improved retrieval resulting from adopting Salton’s methods would outweigh the costs of installing them (Bourne and Hahn, 2003).

[6]Between 1981 and 1988, DIALOG was owned and operated by a separate subsidiary of Lockheed Martin called DIOLOG Information Services. Subsequently, DIALOG was sold several times; most recently, Dialog was purchased by ProQuest in 2008.

2.6 Government Involvement in IR System Evaluation

During the late 1980s, government research and investment into evaluating IR systems began to expand. In particular, two major government efforts began that would be pivotal in the later creation of TREC: the Citator System and Message Understanding Conferences.

In the late 1980s, Donna Harman of NIST created the NIST Citator System, a new IR system based on statistical techniques developed by IR researchers such as Salton and others. Harman and her colleagues at NIST used this system to search a test collection of three relatively large databases, representing over a gigabyte of text. This was done to illustrate the effectiveness of statistical IR techniques on relatively large test collections (Harman and Candela, 1990). Harman's personal experience with this system would prove influential in later years because she had demonstrated the feasibility of working with large data collections in evaluation experiments.

The second major government IR initiative at this time was the creation of the Message Understanding Conferences (MUCs), which were initiated in 1988 by the Naval Command, Control and Ocean Surveillance Center (NOSC) Research, Development, Test and Evaluation Division (NRaD) with funding support from DARPA and were designed to assess and foster research on the automated analysis of military messages containing textual information (Grisham and Sundheim, 1996). The distinguishing characteristic of these MUCs was the evaluations in which participants engaged. For each conference, participating groups were given an initial set of sample messages and instructions on what type of information was to be extracted from those messages. These groups would then develop systems to process these messages to extract the particular relevant information. Shortly before the conference, participants were given a new set of test messages to process without making any changes to the system. At the conference, each participating group would submit the results of this second run, which were then evaluated against a manually prepared answer key to determine how well they performed (Grisham and Sundheim, 1996).

Although results from early MUCs were encouraging, the MUCs highlighted the need for rigorous IR evaluation standards and metrics as well as a method for handling and analyzing large volumes of text. These needs would be addressed by TREC.

2.7 Creation of TREC: NIST's Entry into IR Research and Evaluation

In 1992, the TREC Program, a collaboration between NIST and DARPA, was created to address these problems. The overall goal of TREC was to support and encourage research within the IR community by providing the infrastructure necessary for evaluating IR methodologies using large data sets and to improve the transfer of IR technologies from research labs to commercial products. Table 2-2 provides an overview of significant events in the history of TREC.

Table 2-2. Major TREC Historical Events

Year	Event Details
1990–1991	Charles Wayne (DARPA) asks Donna Harman (NIST) to help create a new, large test collection for the TIPSTER Program
1991	Donna Harman creates data set with DARPA funding and suggests releasing the data to the public and holding a workshop to compare researchers' IR techniques
1992	First TREC held in Rockville, MD
1993	TREC 2 provides first true baseline performance analysis
1994	TREC 3 expanded to include new tracks
1995	TREC 4 involves official track structure
2000	TREC 9 is first "all-track TREC"

TREC was created as an extension of research conducted during the early 1990s as part of the TIPSTER Program. DARPA initiated the TIPSTER Program in 1989 as a way to improve IR technologies through several activities. The primary goal of this program was to advance the state of the art in text-processing technologies through the cooperation of researchers in government, industry, and academia. Specifically, TIPSTER focused on developing IR systems (referred to as document detection) as well as technologies for information extraction and information summarization.

During the founding of TIPSTER, it was realized that a method for evaluating the performance of the IR systems would be required. TIPSTER Program director Charles Wayne asked Donna Harman of NIST to lead this effort.[7] Over the next year, Harman worked with her colleagues at NIST to develop the test collection and evaluation methods that would be used as part of the TIPSTER Program.[8] Completed in 1991, the test collection Harman created included approximately 750,000 documents (2 gigabytes of data). This represented the largest known test collection in existence at the time. Upon delivery, Harman proposed that the new test collection be made available to the larger IR research community and that DARPA fund a workshop using the new test collection to solicit a competition among various IR systems.[9] This workshop would come to be known as TREC.

Based on a background interview with Charles Wayne, it is clear that the involvement of NIST was crucial for turning the research conducted for TIPSTER into a workshop like TREC that directly addressed the primary problems all IR researchers faced in the early 1990s. Mr. Wayne

[7] Mr. Wayne indicated in an interview for this study that Harman was the best person to lead this task both because she had served on the committee that worked to create the TIPSTER Program and because she was an accomplished IR researcher who had studied with Gerard Salton at Cornell University and had already built an IR system and a robust test collection at NIST.

[8] DARPA funneled the money through NIST to pay for Harman's and other NIST staff members' time spent working on this project.

[9] TIPSTER had already funded the international MUCs, organized by NRaD, as described in Section 2.6.

indicated that if TIPSTER had worked with another agency or a private company to develop its evaluation methodology, it is unlikely that they would have had the vision Donna Harman and NIST had for the broader implications of this research. The importance of Harman's and NIST's accomplishments in creating TREC is also supported by the fact that no evidence exists indicating that another organization or government agency (U.S. or foreign) was actively working to address these issues.

In 1992, NIST and the Department of Defense (DoD) cosponsored the first TREC workshop. Approximately 25 different organizations participated in the evaluation competition and attended the first workshop. Because IR systems have historically focused on searching text documents, the first TREC workshops were dedicated to creating common evaluation methods for these types of systems. This was accomplished in two "core" tracks—the Ad Hoc track and the Routing track. *Tracks* are evaluation exercises, each with specific data sets, queries, and evaluation methods. The Ad Hoc track focused on evaluating the ability of IR systems to analyze static text with different queries (this is relevant for retrospective retrieval tasks such as literature searches). The Routing track, by contrast, involved IR systems interpreting changing or "streaming" text with static queries (relevant for tasks involving the selective dissemination of information, such as analysts who wish to monitor news feeds on certain subjects) (Chowdhury, 2004).

2.8 Expansion and Development of TREC: Adding New Tracks

Encouraged by the success of its initial core tracks, TREC sought to extend the common evaluation paradigm to new research areas. This was done, starting in 1994, by adding new tracks to TREC's annual workshops. Some of the major research areas that TREC became involved with include the following:

- **Web Search:** In 1992, when the first TREC was held, the World Wide Web consisted of 130 websites. By 1996, the number had exploded to 600,000 (Battelle, 2005). Today, popular web search engines claim to index 20 billion web pages or more (Robertson, 2008). As a result, the test collections initially developed by TREC were considered small relative to the task of searching the entire Internet. In response, TREC initiated several tracks for the purpose of generating increasingly larger collections and developing evaluation methodologies to accompany them. These tracks include the Very Large Corpus track, the Web track, and the Terabyte track (Robertson, 2008).
- **Content Beyond Text:** The declining cost of computer storage and increasing processor speeds have enabled multimedia documents such as video and audio to be stored economically in electronic formats (Schatz, 1997). TREC launched Video and Speech tracks to encourage research in content-based retrieval of video and audio documents.
- **Information Extraction:** Traditional IR techniques have focused on providing users with documents that may meet their information needs. However, in recent years, the growth of potentially relevant documentation available for any given query has initiated much research in developing systems that actually extract the information users desire from the relevant documents rather than just retrieving the documents themselves (Hersh,

2003). The Q&A track was launched in 1999 to create evaluation methods for these types of systems.

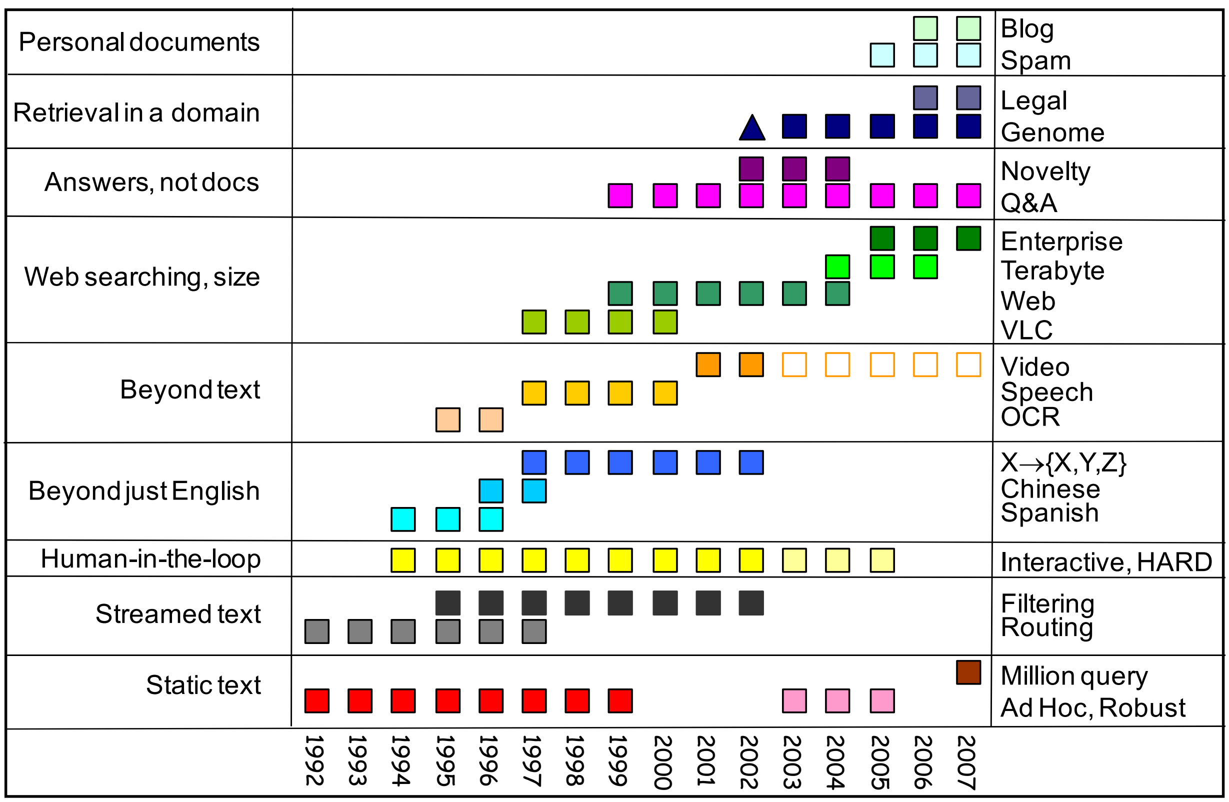
- **Domain Search:** In 2002, TREC introduced the Genomics track. This was the first track devoted to retrieval in a specific domain. Its primary goal was to see whether exploiting domain-specific information could improve retrieval effectiveness (Voorhees and Harman, 2005). In this vein of research, TREC launched the Legal track in 2006.

By the year 2000, TREC ended its two core tracks (Ad Hoc and Routing) and has since continued to expand into other areas of research, such as those listed above. A more complete timeline of all TREC tracks and the research areas they engaged is provided in Figure 2-2.

2.9 TREC Processes and Timeline

As TREC expanded into new research areas, the full TREC planning and execution timeline came into focus. As shown in Table 2-3, the planning and implementation of each TREC evaluation exercise and workshop take an entire year.

Figure 2-2. TREC Tracks by Research Area, Title, and Year



Note: The box colors indicate individual tracks. For example, red boxes identify the Ad Hoc track, which was held from 1992 to 1999, and pink boxes identify the Robust track, which was held from 2003 to 2005. The orange boxes identify the Video track, which was spun off into its own conference series (TRECVid) in 2003, represented by empty boxes from 2003 to 2007.

Table 2-3. TREC Annual Timeline/Steps

Month	Activity	Responsible Party(ies)
November	New tracks proposed	IR researchers and NIST
December	Tracks determined and Call for Participation released	NIST Program Committee
February	Organizations submit intent to participate	IR researchers
November–May	(1) Data set found/created (2) Questions and comparison methodology determined	(1) Each track "leader" (2) NIST and track leader
March	Data sets released to participants	NIST, organizations hosting data sets
May/June	Questions released to participants	NIST and track leaders
August	Submissions due to NIST	TREC participants
October	Relevance judgments released to participants	NIST and track leaders
September/October	Effectiveness results finalized and released	Usually NIST, sometimes with help from track leader
November	TREC workshop held	NIST

Each year new ideas for tracks are proposed at the TREC workshop, and the following year's TREC agenda is set in December. In most cases, a university or other organization proposes new tracks. The TREC Program Committee then determines what tracks will be sponsored (i.e., which NIST will help organize and for which there will be a session at the TREC workshop).

Between November and March, NIST and the organizations that have proposed new tracks work to create new data sets, if needed;[10] write appropriate queries; and develop the evaluation methodologies needed. In some cases, this entire process is very simple; however, in others it can be quite complex. NIST has largely focused on developing the evaluation methodology for each track, whereas the data sets were usually created by outside organizations. Table 2-4 provides a summary of major data sets and their creators.

Around March of each year, the data sets are released to participants, but the queries are usually released in late spring. In most cases, the output generated by a certain system is due back to NIST or the relevant evaluating organization for the track within 6 to 8 weeks. Participants have very little time to make changes to their systems; instead, they must prepare ahead of time and then generate results quickly.

Between early summer and early fall, NIST or the organizing institutions analyze the results of the organizations that participate in each track. This typically involves both automated analysis as well as manual analysis by relevance assessors. For many of the tracks, NIST hires former

[10]Data sets are reused if appropriate.

Table 2-4. Major Document Set by Creator

Document Set	Creator	Tracks Using Document Set
TREC Disks 1–5	Linguistic Data Consortium (LDC), DARPA, and NIST	Ad Hoc Track (1992–1999), Routing Track (1992–1997), Question Answering Track (1999–2001)
AQUAINT	LDC and NIST	Question Answering Track (2002–2006)
Reuters vol1 and vol2	Reuters	Filtering Track (2001–2002)
.gov and .gov2	Australian National University	Web Track (1999–2004), Terabyte Track (2004–2006), Million Query Track (2007–2008)
Blog06 and Blog08	University of Glasgow	Blog Track (2006–2009)
Spam Test Collection	University of Waterloo	Spam Track (2005–2007)
ClueWeb09	Carnegie Mellon University	Web Track (2009)
W3C Corpus	Nick Craswell (Microsoft)	Enterprise Track (2005–2007)

intelligence analysts to review the system output and provide judgment results. In some cases, however, where subject matter expertise is required, alternate relevance assessors are found; for example, in the genomics track, Bill Hersh successfully sought a National Science Foundation (NSF) grant to fund the subject matter experts he hired to assess the results of IR systems' analysis of genomics data.

In late September or early October, results are disseminated to participating organizations, and in November, all participating organizations meet in Gaithersburg, MD, on NIST's campus to discuss the positive and negative results. Since the mid-1990s, these workshops have usually lasted about 2 days. During this time, almost all organizations that participated in the exercises are in attendance, and a large percentage of them present their results.

The next section provides a summary of TREC's main technical accomplishments, the impact of which this study set out to quantify.

3. TREC TECHNICAL ACCOMPLISHMENTS

TREC has revolutionized IR system evaluation through five primary mechanisms:

1. Creation of new, larger test collections.
2. Development of standardized IR evaluation methods.
3. Organization of annual IR research workshops.
4. Distribution of research results.
5. Model for other IR workshops.

By 1997, evidence existed suggesting that TREC's accomplishments had stimulated significant improvement in IR systems. Buckley, Singhal, and Mitra (1997) analyzed the performance of systems being evaluated in TREC's Ad Hoc and Routing tracks. They tested the IR system improvements made each year on the original test collection from 1992 and found that the majority of systems improved significantly between 1992 and 1997. With regard to their own research, they found that they were able to improve the Ad Hoc results of the SMART system by 20% to 25% annually, on average.

Anecdotally, many experts have also indicated that TREC drastically accelerated IR research. In 2008, Hal Varian, Chief Economist at Google, wrote about the impact of TREC, saying that TREC "revitalized research on information retrieval" (Varian, 2007). Varian talked further with RTI researchers during this study, praising NIST both for identifying a key set of problems in the field of IR—data sets were too small and evaluation methodologies were not standardized—and helping to solve these problems directly. However, Varian noted that TREC's influence has been felt less in the subfield of web search from the early 2000s forward, when TREC did not keep up with document set sizes being used by web search companies (Varian, 2010).

Prior to TREC, objective analysis of IR system performance improvement and comparison among systems were not possible. Anecdotally, many TREC participants have indicated that TREC drastically accelerated their ability to improve their IR systems' performance. TREC's success in stimulating IR research has led to the creation of new IR evaluation projects that have used TREC as a role model. These TREC-like projects, that either started as TREC tracks or were created to be similar to TREC, are now making their own contributions to stimulating IR research (Robertson, 2008).

In this section, each of TREC's main accomplishments is discussed. First, the creation of a large test collection is described. Second, the common evaluation paradigm that has benefited IR research is discussed. Next, the benefits of an IR research forum are described, and the types of organizations that have participated in TREC over the past 16 years are identified. The section concludes with a discussion of other IR evaluation projects that are based on the TREC approach.

3.1 Creation of Larger Test Collections

As previously stated, large test collections for evaluating IR systems did not exist prior to TREC. The largest test collection popularly used during this period contained about 12,000 documents. By contrast, test collections developed by NIST were about 80 times larger (Voorhees and Harman, 2005).

The test collections that were used in the initial TREC ad hoc and routing tracks were compiled by NIST with the assistance of the Linguistic Data Consortium (LDC) at the University of Pennsylvania.[11] The document sets for these test collections were stored on three CD-ROMs and were composed of government documents, newspaper articles, and other materials. A description of the size and types of data included in these test collections is provided in Table 3-1. As with many future TREC data sets, because some of this material was obtained from private sources, a fee had to be paid (in this case by DARPA) to gain the intellectual property rights to make the data sets publically available (for a fee).

Over time, the type of data included in TREC test collections expanded to accommodate the growing number of tracks TREC pursued. For example, video data sets were difficult to

Table 3-1. Size of TREC Data

TREC Disk	Collection	Size (MB)	Number of Documents
1	*Wall Street Journal*, 1987–1989	267	98,732
1	Associated Press, 1989	254	84,678
1	Computer Select, Ziff-Davis	242	75,180
1	*Federal Register*, 1989	260	25,960
1	Abstracts of U.S. Department of Energy (DOE)	184	226,087
2	*Wall Street Journal*, 1990–1992	242	74,520
2	Associated Press, 1988	237	79,919
2	Computer Select, Ziff-Davis	175	56,920
2	*Federal Register*, 1988	209	19,860
3	*San Jose Mercury News*, 1991	287	90,257
3	Associated Press, 1990	237	78,321
3	Computer Select, Ziff-Davis	345	161,021
3	U.S. Patents, 1993	243	6,711

Source: Grossman, D., and O. Frieder. 2004. Information Retrieval, Algorithms and Heuristics. Dordrecht, The Netherlands. Springer Publishing.

[11] The LDC is an open consortium of universities, companies and government research laboratories that supports language-related education, research, and technology development by creating and sharing linguistic resources: data, tools, and standards. The University of Pennsylvania is the LDC's host institution. Of note, LDC was founded in the same year as TREC and was originally funded by DARPA; in subsequent years, LDC received ongoing support from NSF.

distribute because of the incredibly large file sizes, and the spam data set was difficult to create because of the strict privacy laws that protect e-mail (e.g., even a university cannot allow researchers to use e-mail from its network for research). For the video track, hard drives were mailed to participants; for the spam track, e-mails that were released to the public during the legal investigation concerning the Enron corporation were used. Most test collections created for TREC provided participants either with more data than was easily available to test IR system performance or with unique data aimed at a particular application that was difficult to create, either based on the subject matter or intellectual property rights regarding data.

3.2 Development of Standardized IR Evaluation Methods

In addition to creating large test collections, TREC has also developed standardized IR evaluation methods that enable IR researchers to evaluate the performance of their systems in a shared fashion. TREC's methodologies and test collections have become the de facto tools for researchers interested in publishing because of the comparability and reproducibility.

The evaluation methods used at TREC are based on the experimentation approach used by Cleverdon and his colleagues, as described in Section 2 (Voorhees and Harman, 2005). Specifically, the TREC evaluation method requires researchers to use their IR systems to retrieve documents that match a listing of queries that NIST provides. NIST typically releases the queries in the spring before the TREC conference. Later, typically in October, NIST releases the relevance judgments that are used to evaluate and score the results of TREC participants. Nonparticipants can also obtain the queries and relevance judgments via TREC's website in February after each annual workshop; data sets used for each track are also made available, often later in the year.

Researchers wanting to evaluate their IR systems can, first, complete a TREC retrieval task using the queries and data sets provided; second, compare their results with the relevance judgments; and finally, compute a variety of metrics to evaluate the performance of the retrieval system. The specific metric that is used to evaluate system performance will typically depend on the evaluation methods used in a particular track.

In many TREC tracks, the evaluation methodology is generally based on an analysis of Mean Average Precision (MAP). The document MAP score is calculated as follows. First, for a given query, the average of all the precision values at each recall point in a document ranked list is first calculated. Second, the mean of all the query average precision scores is determined. The resulting MAP value provides a single measure of quality across recall levels. According to a book by several leading IR researchers, MAP has been shown to have particularly good discrimination and stability as compared to alternate evaluation methods (Manning, Raghavan, and Schütze, 2008).[12]

[12]A detailed description for how MAP scores are calculated can be found in Manning et al. (2008).

Although the simple MAP score can be used for many tracks, the methodology might have to be completely reinvented for some specific exercises. The Question and Answering (Q&A) track, for example, focuses on finding a single answer to a question as high in the ranked output as possible. Thus, the evaluation metric used is mean reciprocal rank (MRR). The performance metric in the interactive track has varied depending on the specific user task but is usually a measure reflective of what the user has been asked to do, such as find one or many answers to a given question.

Furthermore, for certain tasks, determining the MAP components—recall and precision—can be difficult. For example, specific subject matter experts are needed to help design the methodology for analyzing the relevance of system results for some tracks. In other contexts outside TREC, MAP may not be an appropriate metric. For example, web search companies do not deal with searching a document set with a finite size, making measures like recall impossible to calculate. For some tracks, the organization that proposed the tracks rather than NIST manages the methodology development.

3.3 Organization of Annual IR Research Workshops

As described in Section 2, TREC organizes annual IR workshops that provide IR researchers with a competitive forum to openly compare the results of the IR systems they develop. Although these researchers are instructed to avoid making claims of "being the best at TREC" (particularly for product marketing purposes), relative success at TREC does carry a certain amount of weight in the IR community. These IR research workshops also encourage the rapid spread of ideas through the IR community by helping demonstrate what works and what does not work (Robertson, 2008). As a result, these workshops provide an excellent opportunity for training PhD students on the state of the art in IR research. Based on background interviews with TREC participants, it is unlikely that a private organization would have been able to arrange this type of workshop. This is because one of the attractive features of the TREC workshop is that it is organized by NIST, an unbiased third party.

More than 200 organizations have participated in one or more TREC workshops. This level of participation indicates that researchers have benefited from this experience. Table 3-2 summarizes the breakdown of these organizations by type, as well as lists some of the main participants over the years. In general, universities in the United States and abroad have been the primary participants, representing over 50% of organizations. However, many nonprofit research organizations, U.S. and foreign government entities, and corporations have also participated over the years. Table 3-3 provides more detailed information on how many organizations participated in each track.

Table 3-2. TREC Participants 1992 to 2008

Percentage of Participating Organizations	Organization Type	Frequent Participants/Contributors
53%	Universities/colleges	Carnegie Melon, Cornell, Queens College (CUNY), UMass, Univ. of Glasgow, Univ. of Waterloo, Univ. of Illinois (Chicago)
33%	Corporations	Microsoft, IBM, Thomson Reuters
8%	Foreign government agencies/labs	Commonwealth Science and Industrial Research Organization (CSIRO)
4%	Nonprofit research	The MITRE Corporation, Environment Research Institute of Michigan
2%	U.S. government agencies/national labs	National Security Agency

Source: Voorhees, E. 2008. Personal communication with RTI International.

3.4 Dissemination of Research Results

After the annual workshops have concluded, TREC requests that workshop participants prepare papers discussing the results of their research. These papers are published on TREC's website where they are made freely available 3 months after the TREC workshop each year (so that TREC participants receive an extra benefit for participating). In addition, papers written by NIST authors on the results of the TREC workshop are also published on TREC's website. By disseminating research results in this fashion, TREC enables IR researchers that did not participate in a workshop to still benefit from them.

One way to quantitatively measure the importance of TREC's research papers is to estimate the number of times they have been cited by other published papers. Table 3-4 provides citations to five highly cited papers. It is important to note that this list includes only papers that included a NIST author. Many researchers also use TREC data when analyzing their systems to present results at other conferences, such as the ACM SIGIR conference, or to publish in IR journals, such as *Information Retrieval*.

3.5 Model for New IR Evaluation Workshops

TREC also created a model for how other IR workshops should be developed. The IR workshops that have been inspired by TREC or directly spun off from a track at TREC include the following:

- *Cross Language Evaluation Forum (CLEF)—Europe*: CLEF was launched in 2000 and began as a track inside TREC. CLEF, which is held annually in a European country, was initially focused on developing an infrastructure for evaluating cross-language IR systems—systems that retrieve information written in a language different from the language of the user's query. However, its focus has gradually expanded to include other "tracks." This includes an ad hoc track and a web-searching track. CLEF also

Table 3-3. Number of Organizations Participating in TREC by Track

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Ad hoc	18	24	26	23	28	31	42	41									
Routing	16	25	25	15	16	21											
Interactive			3	11	2	9	8	7	6	6	6						
Spanish			4	10	7												
Confusion				4	5												
Database merging				3	3												
Filtering				4	7	10	12	14	15	19	21						
Chinese					9	12											
NLP					4	2											
Speech						13	10	10	3								
Cross-language						13	9	13	16	10	9						
High precision						5	4										
Very large corpus							7	6									
Query							2	5	6								
Question answering								20	28	36	34	33	28	33	31	28	
Web								17	23	30	23	27	18				
Video										12	19						
Novelty											13	14	14				
Genome												29	33	41	30	25	
HARD												14	16	16			
Robust												16	14	17			
Terabyte													17	19	21		
Enterprise														23	25	20	16
Spam														13	9	12	
Legal															6	14	15
Blog															16	24	25
Million Q																11	7
Feedback																	15
Total participants	25	31	33	36	38	51	56	66	69	87	93	93	103	117	107	95	56

Source: Voorhees, E. 2008. Personal communication with RTI International.

Table 3-4. Papers Frequently Cited in the Literature

Publication Information	Number of Citations
Buckley, Chris and Ellen Voorhees. 2000. "Evaluating Evaluation Measure Stability." *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 33-40. Athens, Greece.	101
Buckley, Chris and Ellen Voorhees. 2004. "Retrieval Evaluation with Incomplete Information." *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 25-32. Sheffield, UK.	83
Voorhees, Ellen. 2001. "Evaluation by Highly Relevant Documents." *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 74-82. New Orleans, LA.	50
Voorhees, Ellen. 2000. "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness." *Information Processing and Management* 36:697-716.	61
Voorhees, Ellen and Dawn Tice. 2000. Building a Question-Answering Test Collection. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 200-207.	47

Note: Citation counts obtained through ACM Portal, April 19, 2010.

includes an image retrieval track, ImageCLEF, which itself includes a medical image retrieval task (Hersh, 2003).

- *TRECVID—United States*: TRECVID began as a TREC video track, held from 2001 to 2002, that was devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, the track became an independent NIST-organized workshop, held in the days prior to TREC on NIST's campus, that continues annually.
- *NII Test Collection for IR Systems (NCTIR) Project—Japan*: NCTIR, held annually in Tokyo since 1999, focuses on East Asian languages (predominantly Japanese and Chinese) and include a wide range of IR tasks such as question answering, web search, and text summarization (Hersh, 2003).
- *INitiative for the Evaluation of XML Retrieval (INEX)—International*: INEX organizes an international, coordinated effort to promote the creation and utilization of common evaluation procedures for evaluating IR systems focused on retrieving Extensible Markup Language (XML) documents. INEX was organized annually between 2002 and 2008.
- *Forum for Information Retrieval Evaluation (FIRE)—India*: Started in 2008, FIRE is a relatively new cross-language IR evaluation conference for Indian languages with the aim of encouraging research in Indian-language information-access technologies by providing reusable large-scale test collections for IR experiments (FIRE, 2009).

4. AFFECTED STAKEHOLDERS AND APPLICATIONS

TREC has made significant contributions to the technology infrastructure supporting IR system development, the benefits of which flow directly or indirectly to a variety of stakeholder groups (see Figure 4-1). The direct beneficiaries are IR researchers in academic research groups and commercial firms: TREC's accomplishments improved both the efficiency and the effectiveness of their research and development (R&D) activities. R&D benefits that accrued to academic labs have also flowed indirectly to commercial firms through technology transfer and knowledge sharing, and improvements in the R&D of commercial IR firms have led to improvements in the performance of IR systems commercialized into products and services. End users of these IR systems have indirectly benefited from TREC through higher quality IR products and services. This section describes the overall characteristics of the stakeholder groups that have benefited from TREC.

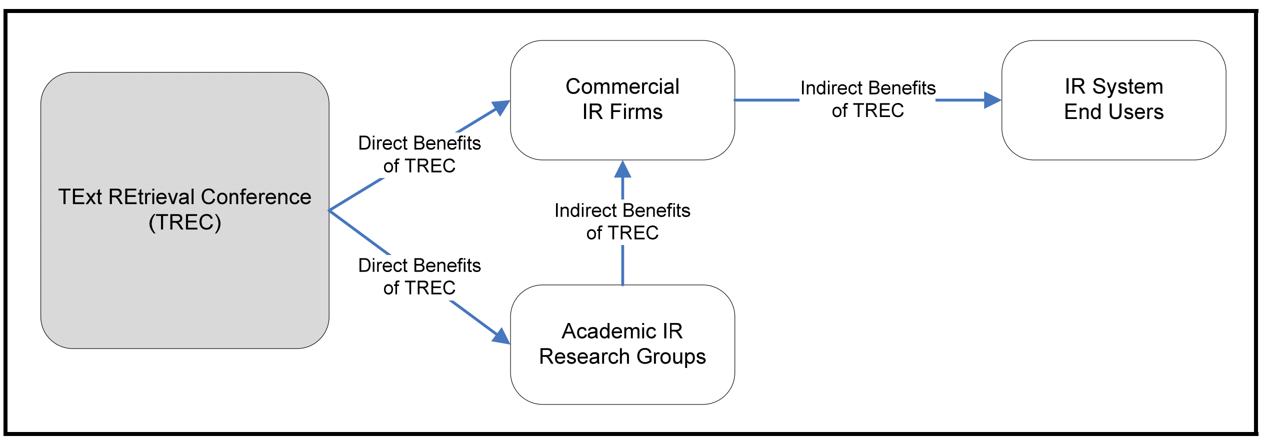
4.1 Developers of IR Systems

In Section 1, IR systems were defined as software systems that conduct three basic tasks:

- indexing source documents
- processing user queries
- identifying source documents that match user queries

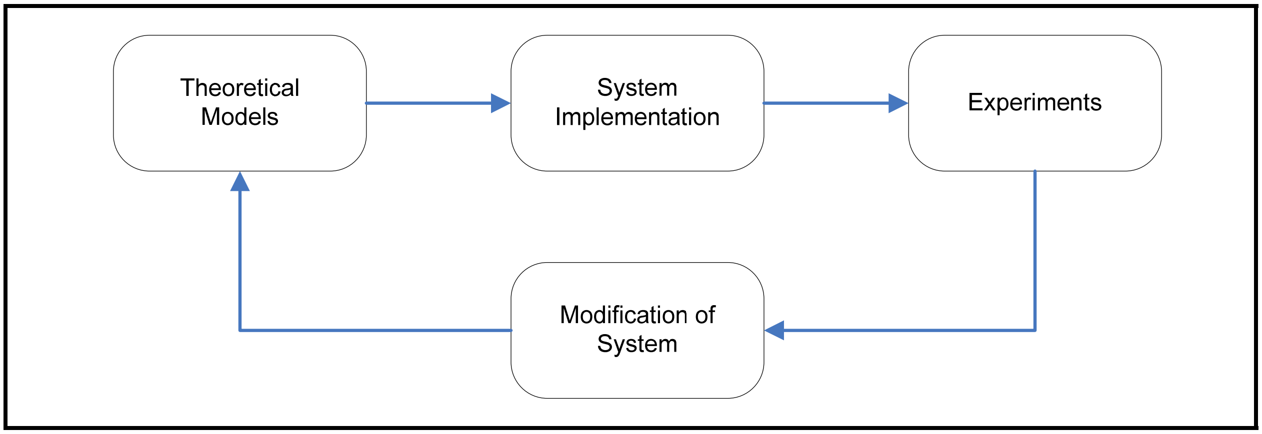
The job of IR researchers is to formulate theoretical models for how each of these tasks should be completed and to integrate those models into a workable system. Once a system has been implemented, IR researchers conduct experiments to evaluate its performance. The results of these experiments help researchers identify ways to modify and improve their IR system (Milic-Frayling, 1999). Figure 4-2 illustrates this development process.

Figure 4-1. Flow of TREC Benefits across Stakeholder Groups



Source: RTI.

Figure 4-2. The IR System Development Process



Source: Modified from Milic-Frayling (1999).

TREC benefited each stage of the development process:

- Research results generated through TREC improve theoretical model development and adjustment.
- TREC's competitive workshops motivated improvements in system implementation by requesting that specific system goals be achieved and tested objectively.
- TREC's large test collections, evaluation paradigm, and workshops enabled more effective experimentation. TREC workshop participants received benefits greater than nonparticipants, particularly related to experimentation. Nonetheless, nonparticipants also benefited from the test collections, evaluation methodologies, and research results that were shared with the public.

Although all types of IR researchers benefited from TREC, differences among groups affected the level of and way in which benefits were received. Because R&D conducted by academic research groups is different from R&D conducted by commercial firms, the impacts of TREC accrue differently to researchers in each group. The following two sections describe each group of IR researchers and their R&D activities in greater detail.

4.1.1 Academic IR Research Groups

The primary focus of academic IR researchers is not to simply create an improved IR system, but instead to better understand the IR systems themselves. Academic research groups generate knowledge—basic, applied, and infratechnologies (e.g., data sets)—that is later made available through conference presentations, reports, and peer-reviewed journal articles. As a result, academic research groups add to the knowledge base that all IR researchers draw on when creating their own systems.

Academic IR researchers are primarily funded by government research grants.[13] In 2007, total computer science R&D expenditures of U.S. universities participating in TREC were approximately $780 million. Although IR research occurs in many university departments, this study's research and interviews suggest that most IR research occurs in computer science departments. Table 4-1 provides a list of the major universities conducting IR research listed in order of their computer science department research funding in fiscal year 2008.

4.1.2 Commercial IR Firms

The primary focus of IR researchers located in commercial firms is to develop IR systems that will either be sold as part of a software application or will be used to offer IR services (collectively referred to as "search products and services").

Table 4-1. Top 20 Universities Pursuing IR Research by Computer Science Department Funding

University	FY2008 Funding ($thousands)
Carnegie Mellon University	99,279
Johns Hopkins University	76,125
University of California, San Diego	55,791
Georgia Institute of Technology, all campuses	52,455
Massachusetts Institute of Technology	45,762
Pennsylvania State University, all campuses	41,862
University of Texas, Austin	37,688
University of Maryland, College Park	33,720
Ohio State University, all campuses	29,255
Cornell University, all campuses	23,304
Oregon Health & Science University	20,850
Stanford University	19,959
University of Massachusetts, Amherst	17,660
University of Utah	16,273
University of California, Los Angeles	16,005
University of Minnesota, all campuses	15,951
University of Wisconsin, Madison	14,827
University of Illinois, Chicago	13,067
University of California, Santa Barbara	11,821
University of North Carolina, Chapel Hill	11,422

Source: NSF, 2010.

[13] Although as the number of IR products and services increased in the mid-1990s, funding from private companies increased.

The search products and services offered by commercial IR firms have changed significantly in recent decades. When IR systems first became commercialized in the 1960s and 1970s, commercial IR firms were primarily involved in enabling a relatively small number of consumers to remotely access and search databases of textual documents (such as abstracts of scientific or medical journals). However, as time passed and the storage of electronic information became more cost-effective, the creation of electronic documents increased. Today, countless documents are stored on individuals' personal computers, companies' internal networks, external databases, and websites. Commercial IR firms have responded to this wealth of data by developing products and services that find documents stored in these locations. The majority of search products and services can therefore be grouped into four primary application categories:[14]

- **Desktop search—**used to locate documents stored on personal computers (e.g., e-mails, videos, to-do lists)
- **Enterprise search**—used to locate documents stored on internal networks (e.g., e-mails, intranet pages, electronic documents, presentations)
- **Database search**—used to locate documents stored on external databases (e.g., scientific journal articles, newspaper articles)
- **Web search—**used to locate information on the World Wide Web (e.g., web pages, videos, music)

A listing of major firms providing these types of products and services is provided in Table 4-2. As this table indicates, many of the major firms providing these products have been direct participants in TREC workshops. Google and Yahoo!, for example, currently employ IR researchers (in executive and other management and research positions) who worked with TREC data while working on their graduate education.

Table 4-2. Example Companies with Search Products, by Type

Search Product Type	Example Companies
Desktop search	Microsoft,[a] Apple,[a] Google, Yahoo!
Enterprise search	Microsoft,[a] Google, Autonomy, Endeca
Database search	News Corp, Reed Elsevier Group,[b] Thomson Reuters Corporation[b]
Web search	Microsoft,[a] Time Warner (AOL), Google, Yahoo!

[a]TREC participants.

[b]Companies that have acquired companies that participated in TREC. Specifically, Reed Elsevier acquired Lexis-Nexis in 1994 and Thomson acquired West Publishing in 1996.

[14]This categorization of search products and services is not necessarily exhaustive nor is it the only way that these products and services can be categorized. For example, Moulton (2008) and White (2007) both use different schemes of product categorization. Therefore, the categorization scheme used here should be understood as only one of many ways for better understanding the markets in which IR system researchers operate.

The sale of these products and services generates a significant amount of revenue for commercial IR firms. Table 4-3 provides revenue data collected by the U.S. Census Bureau for the Information Sector through its Annual Service Sector survey and reported using the 2002 North American Industrial Classification System (NAICS). Commercial IR firms represent only a portion of the total information sector and are typically located in either the software publishing industry (NAICS 5112) or the Internet service and data processing industry (NAICS 518). In 2007, these industries generated $144 billion and $104 billion of revenue, respectively, although it should be noted that these data include firms that are not involved in developing IR systems.

Because IR products are valued by consumers, commercial IR firms employ a significant number of IR researchers to continually improve the performance of their search products and services. Data on the R&D employment of firms occupying different portions of the information sector are provided in Table 4-4. As these data show, firms located in the industries that contain commercial IR firms (software publishing and Internet service and data processing providers) devoted large portions of their total employment to R&D. In 2005, firms in these two industries devoted 26% and 16% of their respective domestic employees to R&D activities. These proportions are significantly higher than the 9% average for the information sector as a whole (NSF, 2007).

In order to provide a better description of the scope of the market for search products and services, the following discussion is devoted to describing each of the four types of search products and services in more detail.

Desktop Search

Desktop search products are used to locate documents stored on users' PCs. This includes web browser histories, e-mail archives, text documents, sound files, images, and video. Currently, major computer operating systems like Microsoft's Windows and Apple's Mac OS include desktop search features. Although desktop search programs have been standard parts of popular operating systems like these for decades, the recent surge in the development of desktop search products did not begin until 2004, when companies such as Google and Yahoo! began offering alternative desktop search products that could be downloaded for free from their websites (Wildstrom, 2005).[15]

[15]Because desktop search products are typically sold as independent software applications or as part of a software package, firms producing these products would likely be included with software publishers under NAICS 5112. According to the data reported in Table 4-2, this industry generated $144 billion in sales in 2007. Unfortunately, estimates of the portion of this aggregated total that is represented by the desktop search market are not readily available. Collecting data on this market is further complicated by the fact that desktop search features are included in most major operating systems and are also freely available for download online. As a result, sales of desktop search software products may not be the most appropriate measure of market activity and measures of the use of desktop search products would be more useful. However, these data were also unavailable.

Table 4-3. Revenue by Industry: 2004–2007

Industry Description	NAICS	2004 ($million)	2005 ($million)	2006 ($million)	2007 ($million)
Information sector total	51	955,083	1,003,262	1,057,430	1,114,883
Publishing industries (except Internet)	511	256,301	269,715	282,880	297,709
Software publishers	5112	112,261	121,334	132,257	143,704
Internet service providers, web search portals, and data processing services	518	82,491	88,598	99,546	104,050
Internet service providers and web search portals	5181	25,161	25,969	28,749	31,168
Internet service providers	518111	20,201	18,977	19,092	19,086
Web search portals	518112	4,960	6,992	9,657	12,082
Data processing, hosting, and related services	5182	57,330	62,629	70,797	72,882

Source: U.S. Department of Commerce, Bureau of the Census. 2008. Service Annual Survey. Available at <http://www.census.gov/svsd/www/services/sas/sas_data/historical/sas-07.pdf>.

Table 4-4. Labor Resources Devoted to R&D in Information Industries: 2005

Industry Description	NAICS	R&D Scientists and Engineers (thousands of FTEs)	Total Domestic Employment (thousands of FTEs)	% Total
Information	51	134.2	1,493	9%
Publishing, including software	511	98.7	381	26%
Telecommunications	517	10.2	714	1%
Internet service and data processing providers	518	23.8	149	16%
Other information	Other 51	1.5	249	1%

Source: National Science Foundation (NSF). 2007. "Expenditures for U.S. Industrial R&D Continue to Increase in 2005; R&D Performance Geographically Concentrated." Available at <http://www.nsf.gov/statistics/infbrief/nsf07335/>. As obtained on April 28, 2009.

Enterprise Search

Enterprise search is similar to desktop search in that it deals with searching documents that are stored by the consumer of the product. However, the documents in question are stored not on an individual's personal computer, but instead on the internal networks of large companies or other enterprises. Enterprise search has been of growing importance since the early 1990s when firms began to more fully incorporate the use of computers and internal networks in their everyday operations.

According to estimates from the Gartner research company, as much as 85% of all data within an average company is contained in unstructured formats such as e-mail or documents (Levack, 2002). As a result, businesses are increasingly using enterprise search products and services so that they can better access the information they need to operate. This information may be required for decision making and planning, for complying with government regulatory requirements, or for complying with information requests during the discovery phase of civil litigation (often referred to as e-discovery). Enterprise search is typically performed using software packages purchased by the enterprise. However, service companies are also available for performing different enterprise search tasks. For example, companies such as H5 perform e-discovery services for Fortune 500 companies.

Enterprise search software vendors would likely be included with other software publishing firms in NAICS 5112. However, more detailed information can also be obtained for this market through private research companies. For example, according to Gartner Inc., the worldwide market for enterprise search software has grown rapidly in recent years. Between 2006 and 2008, worldwide enterprise search software revenue grew from $717.2 million to $989.7 million (an average annual increase of 13%). Gartner projects that enterprise search software revenue will surpass $1.2 billion in 2010 (Gartner, 2008).

Database Search

Commercial firms first began offering database search products and services in the 1960s and 1970s. Unlike desktop and enterprise search, database search typically deals with locating information contained on databases that are stored outside a person's personal computer or an enterprise's external network. These databases can be stored on CD-ROMs or on external servers that individuals can access remotely through communication networks like the Internet (referred to as online IR services). However, because databases stored on CD-ROM are difficult to update and have relatively limited storage space, the vast majority of database searches are involved in providing online IR services. Authors such as Chowdhury (2004) identify LexisNexis (owned by the Reed Elsevier Group), Westlaw (owned by Thomson Reuters), DIALOG (owned by ProQuest), and Dow Jones (owned by News Corp.) as popular online database search service providers.

Online database search companies have typically been included in NAICS 518111 along with other types of online service providers. According to data published by the Department of Commerce, revenue for this industry was approximately $20 billion in 2007 (Table 4-2).[16]

Web Search

Web search products and services were first developed in the early 1990s to search the growing number of web pages and file directories that made up the World Wide Web. The information contained on these pages included hypertext documents, images, video, and many other file types.

Companies offering web search services are classified under NAICS 518112. According to data reported in Table 4-3, firms in this industry received $12 billion in 2007. This is almost double

[16]Unfortunately, this total includes many companies that do not provide search services of online databases. Private estimates could not be found for this market.

the revenue this industry recorded in 2004. However, unlike online database search companies web search companies have typically offered their services for free and instead derive revenue from the sale of ad space. Therefore, these data do not necessarily provide an adequate sense of the extent to which end users are using web search services.

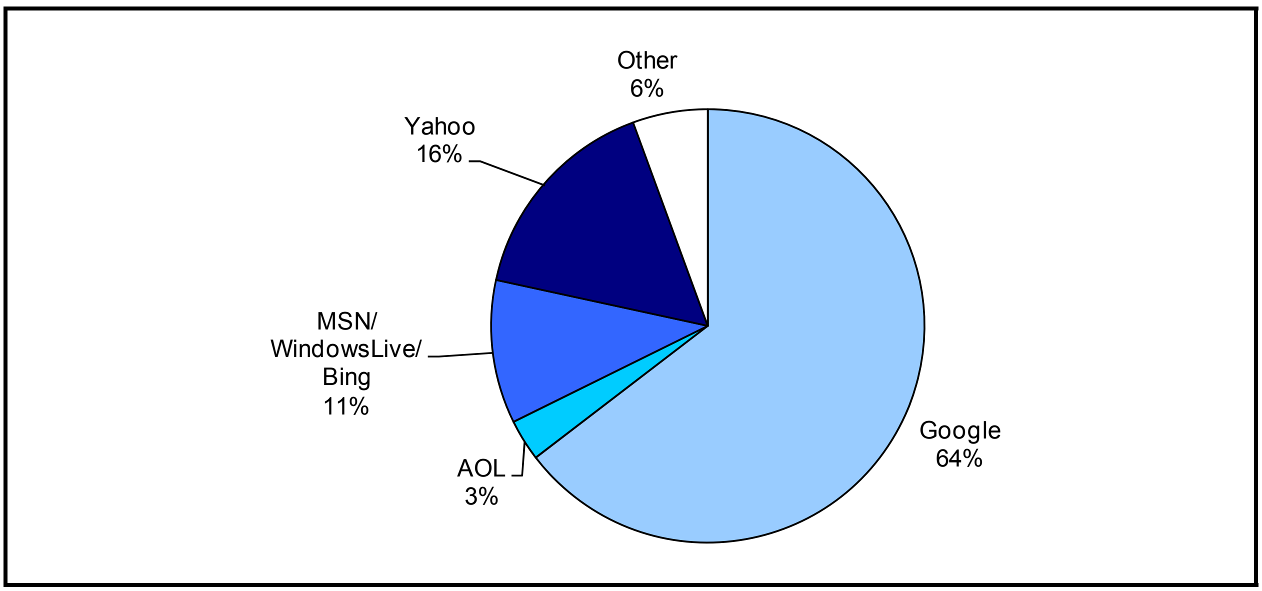
According to the Online Publishers Association (OPA), there are over 150 million unique visitors to search engine websites each month (OPA, 2010). As a result of these millions of search engine users, billions of searches are conducted each month. According to data from the Neilson ranking system, over 10 billion searches were conducted in August 2009 (Search Engine Watch, 2009). Figure 4-3 reports the share of total searches by engine. Over 60% of these searches were made using the Google web search engine.

4.2 End Users of IR Systems: Individuals and Organizations

End users of IR systems form the customer base of commercial IR firms. To understand end users, their information needs must also be understood. At a conceptual level, Wilson (1981) explains that "information needs" arise out of an individual's efforts to satisfy three basic human needs:

- physiological needs (e.g., food, water, shelter),
- emotional needs (e.g., need for attainment, affirmation), and
- cognitive needs (e.g., need to plan, to learn a skill).

Figure 4-3. Portion of Total Web Searches Accounted for by Popular Web Search Engines, 2008



Source: Johnson, N. 2009. "Nielsen Online December 2008 Search Engine Share Rankings." Search Engine Watch. Available at <http://searchenginewatch.com/3632382>. As obtained on May 19, 2009.

In attempting to fill these interrelated basic human needs, individuals may recognize missing information that needs to be found. For example, the physiological need for food may give rise to a cognitive need (such as the need to get groceries), which may require information to complete (such as knowledge of areas that have grocery stores).

However, this is not to suggest that the simple presence of an information need triggers information seeking. As Wilson (1981) points out, factors such as the cost of acquiring information may create barriers that discourage information seeking and result in leaving a particular information need unfilled. In cases where an individual must search for one or a small number of documents that contain the information he needs, the barrier is the cost of the time and energy it would require to search the collection of documents.[17] Thus, end users value IR systems because they reduce the cost of obtaining the information that helps fill these three needs.

The type of information being sought by an end user will depend on the specific context in which an individual is attempting to meet a basic human need. A single individual may have different information needs in the context of his job than in the context of his personal life. For example, in a work context, a manager working inside a large company may need information from an e-mail exchange he had with colleagues or financial data for the company to formulate future business plans. In this case, using desktop or enterprise search products or services may be beneficial. In these types of situations, search products and services and other outputs of the information services industry can be considered intermediate goods in the production of other goods and services.

To illustrate how workers and companies in other industries use the output of businesses in the information services industry (NAICS 51), Table 4-5 shows the total amount of intermediate goods consumed by each industry and the portion of these purchases that are accounted for by information industry output in 2007. As expected, the information industry is the biggest consumer of its own goods, followed by professional and business services; government; and finance, insurance, real estate, rental, and leasing. Overall, 6% of the $12 trillion of total intermediate goods purchases are accounted for by the information sector.

In addition to information needs in a work context, individuals may also require information in nonwork contexts. For example, an individual may require information on candidates running for office to make an informed voting decision. Using a web search engine may be particularly useful for identifying documents stored on websites that would contain this type of information. According to the Bureau of Economic Analysis, approximately $46 billion of output from the information industry is purchased directly by private citizens (BEA, 2009). This includes money spent by individuals on newspapers and subscription fees to online media, among other purchases.

[17]This is why Wilson (1981) suggests that “information seeking behavior” might be a more apt term than “information need.”

Table 4-5. Purchases of Information Service Industry Output as an Intermediate Good by Other Industries (2007)

Industrial Sector	NAICS 51: Information Sector Output Purchased as Intermediate Goods ($millions)	Total Purchases of Intermediate Goods ($millions)	% of Total
Agriculture, forestry, fishing, and hunting	1,119	200,103	1%
Mining	1,239	203,785	1%
Utilities	372	173,197	0%
Construction	14,370	717,535	2%
Manufacturing	43,365	3,428,693	1%
Wholesale trade	20,776	368,583	6%
Retail trade	21,884	408,726	5%
Transportation and warehousing	10,772	374,379	3%
Information	288,673	782,645	37%
Finance, insurance, real estate, rental, and leasing	50,929	1,668,822	3%
Professional and business services	118,648	1,067,145	11%
Educational services, health care, and social assistance	47,053	667,657	7%
Arts, entertainment, recreation, accommodation, and food services	25,516	433,752	6%
Other services, except government	22,814	352,318	6%
Government	75,769	1,154,024	7%
Total	**743,299**	**12,001,363**	**6%**

Source: Bureau of Economic Analysis (BEA). 2009. 2007 Input-Output (I-O) Accounts. Available at <http://www.bea.gov/industry/index.htm#annual>. As obtained on May 20, 2009.

5. ECONOMIC ANALYSIS METHODOLOGY

By offering large test collections, standardized evaluation methods, annual workshops, and new research, TREC has reduced the cost and improved the quality of R&D conducted by IR researchers. In turn, these improvements have led to the development of higher quality IR systems that enable end users to satisfy their information needs in fewer search iterations.

An appropriate measure of the net economic benefits generated by TREC can be defined as the net economic benefits that IR researchers and end users have received from TREC minus the costs of creating and operating TREC:

net economic benefits = ∑ IR researchers' net benefits +

∑ end users' net benefits − ∑ TREC investment costs

where the IR researchers' net benefits are defined as the change in R&D cost (efficiency) and the change in the value of R&D output (quality) that resulted from TREC's accomplishments minus the costs of utilizing test collections, workshops, and published research:

IR researchers' net benefits = ∑ Δ R&D cost +

∑ Δ product quality − ∑ utilization costs

End users' net benefits are defined as the reduction in user labor costs that result from using improved IR systems minus the change in the cost of using these systems:

end users' net benefits = ∑ Δ Search labor costs −

∑ Δ IR system purchase/utilization costs

TREC investment costs are defined as the costs incurred to create the test collections and evaluation methodologies for each track, facilitate the judging of TREC participant entries, host the annual workshop, and disseminate research results:

TREC investment costs = ∑ Test collection / methodology creation costs +

∑ Competition / workshop costs + ∑ Research dissemination costs

As shown in these equations, benefits have accrued to *two* stakeholder groups—IR researchers (direct benefits) and end users[18] (indirect benefits)—while costs are incurred by *three* stakeholder groups—IR researchers (utilization costs), end users (utilization costs), and "investors" in TREC (creation/facilitation costs).

[18]Of note, economic impact metrics were not developed for end users because of the difficulty in accurately estimating the improvement in IR products and services that occurred during the time that TREC has existed, the impact of TREC on this improvement, and the level of utilization and usage of the relevant IR products and services. However, qualitative impact metrics for end users are described in Section 5.1.2 and presented in Section 6.

This section describes

- how economic costs and benefits were conceptualized (Section 5.1),
- what data collection methods were used to estimate costs and benefits (Section 5.2),
- what extrapolation methods were used to calculate national net benefit estimates (Section 5.3), and
- how analytical results were summarized (Section 5.4).[19]

5.1 Taxonomy of Economic Costs and Benefits

Economic costs and benefits were measured relative to a counterfactual scenario under which TREC was not created and IR R&D progressed in the absence of NIST, DARPA, and other public agency support. Our research and interviews suggest that if TREC had not been created, another U.S. or foreign government agency might have eventually created a TREC-like program (as discussed in Section 2, efforts had already been pursued in the UK in the 1970s). However, in the early 1990s nothing similar to TREC was being discussed publically. Thus, the counterfactual scenario specifies that no TREC-like program would have existed in the absence of NIST.

The counterfactual scenario established a framework for identifying, describing, and estimating the net benefits of TREC by making explicit the costs that would have been incurred in the absence of the Program and the benefits that would have been lost. A summary of TREC's accomplishments and consequences and the costs and benefits they generate by stakeholder group is provided in Table 5-1. The remainder of this section provides a more detailed discussion of the specific costs and benefits of TREC's accomplishments, as compared to the counterfactual scenario, by stakeholder group.

5.1.1 IR Researchers: Cost and Benefit Categories

In order to benefit from TREC resources, IR researchers must first incur utilization costs with the use of each TREC resource. However, the type(s) of utilization costs differs by the resource type. To use the test collections, in some cases IR researchers could access the data for free by simply downloading the data from a website, or they paid the data host a fee to receive a copy of the data. IR researchers might have had to pay a registration fee and travel expenses, incur

[19] Although the discussion presented in this section focuses on efforts to monetize the costs and benefits of TREC, not all of the economic impacts could be monetized. Interviews with IR researchers and an Internet survey were used to collect information on the various ways that TREC has affected R&D which could not be easily monetized. For example, researchers were asked about the following perceived benefits of different types of TREC resources (e.g., workshops, test collections); the perceived benefits of different TREC tracks; the ways in which they have used TREC resources—for example, whether they have used TREC data to address research questions beyond those pursued officially at TREC workshops or whether they have cited TREC resources in patent filings; and whether TREC has affected human capital development by influencing the perceived return on graduate education in IR. The qualitative information that was collected helps better explain the breadth of TREC's impact. The results are presented in Section 6.

Table 5-1. Example Benefit and Cost Metrics, by Stakeholder Group

Technical Accomplishment/ Consequence of TREC	Benefit to Stakeholder	TREC Investment Costs (Government and Nongovernment)	Stakeholder Utilization Cost
IR Researchers			
Larger test collections	• Lower R&D costs from being able to avoid labor hours spent developing alternative test collections. • Improved R&D productivity from having better test collections than would have been available otherwise.	• Time spent compiling documents for inclusion in test collection, making relevance judgments, etc. • Fees paid for obtaining rights to use copyrighted material in test collection (when applicable).	• Fees related to purchasing test collection (when applicable).
Standardized IR evaluation methods	• Lower R&D costs from being able to avoid labor spent pursuing less fruitful research paths because standardized evaluation methods allow researchers to more easily look at the results of others to identify promising IR techniques.	• Time spent developing standardized evaluation methods.	• Time spent learning and understanding evaluation methods.
Annual IR research workshops	• Lower R&D costs from being able to avoid labor spent duplicating efforts or pursuing less fruitful IR techniques because workshops help researchers identify the most promising IR techniques. • Lower human capital development costs (e.g., training facilitated by TREC participation).	• Time spent organizing and administering IR research workshop.	• Time spent preparing for IR research workshop. • Fees related to workshop attendance and travel costs.
Novel research results (e.g., papers)	• Lower R&D costs from being able to avoid labor spent duplicating efforts or pursuing less fruitful IR techniques because research papers help researchers identify the most promising IR techniques. • Lower human capital development costs (e.g., training facilitated by TREC participation).	• Time spent writing papers summarizing the results of TREC workshops as well as related research.	• Time spent by workshop participants writing papers on the results of their research. • Time spent by others obtaining, reading, and understanding research results.
End Users of IR Systems			
Improved IR system performance	• Lower search-related labor costs as a result of improved IR systems that facilitate improved searching.	• None	• Price of purchasing search product or service (if applicable).

labor costs to attend TREC (2 to 3 days), and test their IR system using the TREC document set and queries for each track in which they participated.

Once IR researchers have acquired and used TREC resources, they can begin receiving benefits associated with their use. Since 1992, TREC has directly benefited researchers of IR systems through four primary accomplishments:

- *Creation of new, larger test collections*: Providing IR researchers with large test collections has enabled them to conduct higher quality evaluations of IR systems at a lower cost.
- *Development of standardized IR evaluation methods*: Providing IR researchers with IR evaluation methodologies has enabled them to use higher quality evaluation methods at a lower cost.
- *Organization of annual IR research workshops*: Offering a venue for IR researchers to compete against one another in narrowly defined forums and receive feedback on the relative performance of their IR systems helped researchers identify which IR techniques worked and which did not and also facilitated information exchange.
- *Dissemination of research results (e.g., papers)*: Working with workshop participants as well as NIST personnel to publish papers detailing competition results and findings via the TREC website.

Each of these benefit categories is described in detail below.

Creation of New, Larger Test Collections. By taking the lead in developing large test collections and making them available to IR researchers (sometimes for a fee), TREC is responsible for generating two significant research benefits that accrued to IR researchers:

- lower R&D costs because of avoided labor costs that would have been incurred to develop alternative test collections and
- improved quality of products coming out of R&D because of the use of better test collections than would have been available otherwise.

By developing large test collections and offering them to IR researchers at greatly reduced cost, TREC helped researchers avoid the cost of developing alternative collections. It was learned during background interviews with industry representatives that IR researchers routinely use several test collections when evaluating the performance of their IR systems. However, as evidenced by the history of IR evaluation before TREC and confirmed in background interviews, it is clear that the test collections that existed in the early 1990s were too small to be considered true alternatives to those created by TREC and its partners. Therefore, it is unlikely that the test collections identified or developed by individual research groups would have been of the same size or quality of the test collections created by TREC.

Thus, the second benefit generated by TREC's large test collections is that they made IR research more productive: the same amount of R&D inputs, such as labor, could generate higher quality R&D outputs, such as a new product or process. Without large test collections, IR researchers could not conduct experiments to accurately see how they would perform in realistic settings where they would be required to search very large amounts of data. As a result, there was often great difficulty in transferring technologies into application.[20]

Development of Standardized IR Evaluation Methods. TREC developed and validated performance metrics and evaluation processes that all members of the IR community could share easily. Standardized evaluation methods enabled IR researchers to evaluate IR systems in a common way so that the performance of one IR system could be compared to the performance of another. As a result, researchers better understood which IR techniques worked best for a given task. This effectively increased the productivity of IR R&D because it helped researchers avoid spending resources pursuing techniques that may not work well for a given task. Evaluation methods were developed for each track building on a basic framework established when TREC was first created. Each track had its own tailored evaluation process that any IR researcher could adopt. Although TREC participants have had access to the queries, relevance judgments, and other information required for evaluation while at the workshops, nonparticipants have also been able to access these materials via TREC's website several months after each workshop.

Organization of Annual IR Research Workshops. The benefits of the workshops organized by TREC primarily accrue to the researchers who attended them. Attending these workshops required that researchers pay for test collections with a fee, devote time to preparing IR systems for evaluation, travel to the workshop location, and pay certain fees for attendance (approximately $375 in 2008). In exchange for incurring these costs, IR researchers can expect to benefit from the annual workshops in two main ways:

- lower R&D costs because the workshops help researchers identify the most promising IR techniques and, therefore, help them reduce labor hours that would have been spent duplicating efforts or pursuing less fruitful IR techniques and
- lower human capital development costs (e.g., training facilitated by TREC participation).

TREC workshops have benefited IR researchers who have attended by providing a forum in which to evaluate the relative performance of the IR systems and to exchange information on research successes and failures. During background interviews, one representative of a relatively large IR service firm indicated that attending TREC workshops and reading the TREC research papers had been useful because he was able to see how other researchers addressed particular problems and learn which techniques were more or less appropriate for specific tasks. As a result, he indicated that he has avoided several lines of research already proven to be unpromising.

[20]Researchers use multiple test collections to give them a better indication of how their systems will perform using several data sets and to help them avoid "overtraining" their systems to a single test collection. As a result, if researchers did not use TREC test collections, they would have had to spend additional labor hours either identifying an alternative or creating their own test collections.

Participating in TREC workshops also provided educational benefits that were not available elsewhere. In particular, TREC workshops enabled newcomers to get up to speed quickly on the state of research in a particular IR field. This study hypothesizes that this educational atmosphere may have helped PhD students reduce the amount of time they would have to spend surveying the literature and developing applied IR research skills that would make them more attractive to potential employers. TREC essentially provided real-world experience to students that was very valuable to IR companies that were hiring.

Dissemination of Research Results. Although the benefits of attending TREC workshops accrue primarily to individuals who participate, participants often devote time to writing research papers that detail the lessons they learned during the workshop. As a result, individuals who did not attend a workshop can still benefit by taking the time to obtain, read, and understand papers that were written for a particular workshop. The benefits that IR researchers can expect to receive from incurring these costs include

- lower R&D costs because the research papers help researchers identify the most promising IR techniques and, therefore, help them reduce labor hours that would have been spent duplicating efforts or pursuing less fruitful IR techniques and
- lower human capital development costs (e.g., training facilitated by TREC participation).

5.1.2 End Users of IR Systems: Cost and Benefit Categories

As previously discussed, end users indirectly benefited from TREC by having access to improved IR systems. However, in order to receive benefits, end users of IR systems may have to incur costs to gain access to those systems, which may have changed as a result of TREC. For example, in some cases (e.g., web search) IR products were given out freely; in other cases, companies likely charged higher prices for IR products and services that could speed up an organization's ability to search within their own databases (enterprise search) or search other databases (database search). In these cases, the change in overall costs related to utilizing IR products and services affected by TREC could be significant.

Despite these utilization costs, the improved performance of the IR system could have resulted in enough improvement to result in a net benefit for end users. Without TREC, end users of IR systems would likely have used lower quality IR products and services over the last 18 years. TREC's impact on IR researchers resulted in an improvement in the quality of the IR products and services that they offer to customers. The resulting benefit that accrued to end users of IR systems was the differential improvement (the improvement that resulted explicitly from TREC) in the performance of IR systems, which yielded higher quality results in fewer searches. This means that business and consumer end users of IR products and services that were affected by TREC expended less time to acquire the same amount of information as before.

5.1.3 TREC Investors: Cost Categories

A number of investment costs incurred by NIST and other organizations would have been avoided if TREC had not been created. As described in Section 2, TREC was created as an extension of research started in the late 1980s as part of the TIPSTER Program. Therefore, some of the costs associated with the first TREC in 1992 can actually be tracked to DARPA and the TIPSTER Program in 1991, when DARPA funded Donna Harman at NIST to create the document set that would eventually be used for the first TREC workshop.

Since the inception of TREC, funding has been provided from a variety of sources to operate the Program. For example, funding has continued to flow intermittently to TREC from DARPA. Similarly, funding has also been provided by other U.S. government organizations (specifically, intelligence agencies) whose activities could be improved by TREC's accomplishments.

In addition to government funding used to operate the TREC Program, costs have been incurred by academic and nonprofit organizations in producing resources for TREC. For example, from 2003 to 2007, the Oregon Health & Science University (OHSU) led efforts to create and maintain test collections for the TREC Genomics track; OHSU was partially funded to conduct this work by a grant from the NSF. Similarly, various portions of test collections (such as document sets, queries, or relevance judgments) used in TREC tracks were created by organizations other than NIST. In many cases, these costs would also have been avoided if TREC had not been initiated. The various public and private investment costs described above are quantified in Section 6.

5.2 Measuring Economic Costs and Benefits

The previous section categorized and described the economic costs and benefits that were associated with various stakeholders. Data informing the measurement of these costs and benefits were collected through three primary phases:[21]

- Phase I consisted of scoping interviews over a wide range of topics with a variety of IR researchers from NIST, private companies, and universities. The purpose of the scoping interviews was to explore TREC and its impacts to inform hypotheses about TREC's economic benefits.
- Phase II was the distribution of an Internet survey (see Appendix A) to the IR research community. The survey included questions regarding how researchers personally used TREC resources in their research, what value their organization would place on these resources, and how they believe these resources affected the development of IR systems.[22] To maximize distribution of our survey announcements, we sent survey

[21] The formal data collection survey in this analysis was approved by Office of Management and Budget (OMB). This study respected the sensitive nature of the information needed to quantify economic benefits. Respondents were promised confidentiality and that only aggregated data would be presented in the report.

[22] Draft survey instruments were pilot tested with members of the IR research community to help determine the most suitable way to ask individuals about the costs and benefits of TREC. In-person and telephone interviews were conducted with survey respondents and other industry representatives to obtain information on these topics in an open-ended and conversational manner that cannot be obtained through an Internet survey.

participation solicitation requests through the TREC e-mail listserve and partnered with the ACM SIGIR, which sent information through the SIGIR-IRList.

- Phase III involved a series of follow-up interviews with respondents to the Internet survey. These interviews were conducted to verify our assessment and synthesis of data from Phase II.

The following sections describe how data were collected during this study and how these data were used to estimate costs and benefits for IR researchers and end users.

5.2.1 IR Researchers: Cost and Benefit Estimation

As described above, many IR researchers certainly received value from using TREC resources, but estimating the magnitude of these benefits was difficult because TREC resources are not bought and sold like market goods. In previous studies, the benefits of NIST technology programs accruing to research organizations were typically estimated using a cost-savings approach. A survey, fielded to relevant researchers, was designed to quantify how much researchers benefited, in terms of labor/material/time savings, from a particular innovation that resulted from the technology program; benefits were estimated relative to a counterfactual scenario in which the innovation was not present. These labor and material savings were then monetized using secondary data and extrapolated to the national level.

Although such an approach was explored as part of this study, it was determined to be infeasible for this project. Specifically, background interviews revealed that IR researchers had a difficult time estimating what types of costs they would incur if TREC did not exist. In many cases, IR researchers (particularly those located in private companies) indicated that the research they have pursued would not have been possible without TREC resources, and if those resources were not available, they and their organizations would likely have chosen different research paths. In some cases, according to interviews, whole companies may not have existed or been created had it not been for TREC.[23]

In response, it was determined that a contingent valuation approach would be the best approach for measuring the economic value generated by TREC. Contingent valuation is a survey technique that is well established in environmental economic studies and is used to value nonmarket resources such as clean air or water quality (Champ et al., 2003). In their simplest form, contingent valuation questions ask individuals directly to estimate how much value they place on a particular resource. Although much more sophisticated techniques could be used, a variation of this simple open-ended approach was determined to be the most appropriate approach for this study because it was unclear how many individuals would likely participate in the survey, and the magnitude of benefits themselves were uncertain.

[23] Around 1994, Eric Brewer, an assistant professor at Berkeley, contacted Hal Varian, then an economist at Berkeley, to ask him for references to researchers knowledgeable in IR techniques. Varian pointed Brewer to Ray Larson, an IR researcher at Berkeley. Larson pointed Brewer to the work of another Berkeley IR researcher, who had just retired, named William Cooper who had published a paper at TREC describing a novel IR formula. Brewer ended up using this formula as the basis for the Inktomi search engine (Varian, 2010).

One difficulty with using this contingent valuation approach is that TREC resources are not consumption goods but are intermediate inputs into the production of IR research. As discussed in Section 4, this research is pursued by groups of IR researchers collaborating inside private companies and universities. As a result, this survey focused on estimating the value that the entire IR research organization would place on TREC resources.

Box 1 reproduces the questions used to elicit estimates from individual survey participants regarding how much value TREC resources contributed to their research. Based on their previous responses, respondents saw the number of years that researchers in their organization have used TREC resources where X appears in Box 1. Respondents' statements about the size of their organization's IR R&D labor force were used to compute the valuations presented to respondents, where $Y appears in Box 1, according to the following formula:[24]

$$\$Y = (M + ((N * R * V) * P)) * X$$

where

M = membership fee to join LDC (approximately $20,000 for academics and $200,000 for private companies).

N = number of full-time equivalents (FTEs) working in the respondent's organization that are involved in IR system R&D.

R = the percentage of FTEs whose activities the respondent is familiar with and can respond for.

V = the value of a single IR FTE as computed by the weekly salary of a production worker in the software industry ($1,406.98 per week in 2008 [BLS, 2010]) times 52 weeks in a year, times 2 to account for benefits and overhead.

P = Percentage of the value of FTEs that respondents are being asked to consider, which varies as the game progresses. Respondents are first asked to consider a valuation based on 50% of the value of FTEs. If the respondent indicates that TREC did not contribute this amount of value to their organization, they are then asked to consider a valuation based on 25% of their FTEs. Conversely, if the respondent agrees with the initial valuation, they are then shown a higher valuation based on 70% of the value of FTEs. The exact percentages used are 1,000%, 200%, 100%, 70%, 50%, 25%, 10%, 1%, and 0%.

X = the number of years they have been using TREC resources.

[24] Implicit in this approach is the assumption that the value an organization would receive from TREC depended positively on the size of its IR R&D efforts.

Box 1

Estimating the Benefits of TREC Resources Using Contingent Valuation

During the contingent valuation portion of the survey, respondents are asked the following question(s):

> "Over the past 17 years, TREC has provided a number of resources (test collections, evaluation methods, workshops, and research papers) to help foster IR research. Without the use of TREC resources, your organization might have
>
> - spent **more time creating new data sets** (or using smaller or less robust data sets),
> - spent **more time evaluating IR systems** and been less confident in evaluation results,
> - spent **more time going down faulty paths,** and/or
> - created **IR systems of lower quality.**
>
> Earlier in this survey you indicated that researchers in your organization have used TREC resources for X years. Suppose that you were asked by a project manager or your organization's financial accountant to justify the use of TREC resources by specifying the value they contributed to your research.
>
> Do you believe these resources contributed at least $Y of value to your research?"

If respondents answer "yes," then the question is represented with a higher value and the process is repeated until the respondent answers "no." Alternatively, if the respondent answers "no," then the question is repeated with a lower value until the respondent answers "yes." This means of eliciting valuations from respondents is known as a bidding game (Freeman, 2003).

Conceptually, the lowest value a respondent would see (where P = 0%) would be the cost of joining an organization like LDC for X number of years. This lower bound was based on background interviews with several industry experts who believed that this was a reasonable minimum amount that most IR researchers would pay for TREC resources. Alternatively, the highest value that a respondent would see would be 1,000% of the value of FTEs their organization employs in IR R&D for X number of years plus the cost of joining an organization like LDC for X number of years.

If respondents reached either the upper or lower bound without identifying the value that represents the maximum value that TREC has contributed to their organization, they were then asked in an open-ended format to identify this value.[25]

After data were collected from individual researchers inside IR research organizations, valuation estimates were developed for the organizations employing them using data provided by the individual respondents. In cases where only one individual responded for a particular organization, it was assumed that the valuation estimates that individual provided expressed the total value their organization received from TREC during the time they used TREC resources. In cases where multiple individuals responded for a single organization, the responses of the individual who indicated he/she occupied a more senior position and/or could respond for a greater portion of IR researchers at their organization were used. If it was difficult to determine whose responses were most representative for an organization, an average of the responses provided was used. Once benefits were estimated for individual organizations, they were extrapolated to national-level benefits estimates using secondary information. Section 5.3 describes the extrapolation strategy in detail.

The advantages of contingent valuation for quantifying the benefits of TREC are as follows:

1. *Full value can be assessed efficiently and effectively*. When determining how much they would value each TREC resource, each respondent was asked to consider and account for all the ways that TREC benefited his/her research. As a result, this approach measured the full value that IR researchers place on TREC resources. This approach also implies that asking questions in this manner can be more efficient than asking respondents to consider individual benefits of TREC resources because it is less likely that important benefits will be left out.
2. *Monetization is simple and defensible*. Because the respondents' answers were already reported in terms of dollars, there was no need for secondary data to monetize the benefits metrics estimated by IR researchers. This reduced the data collection and analysis time and also reduced uncertainty in final estimates that may have resulted from making assumptions on which secondary data are most appropriate for particular monetization tasks.
3. *IR researchers were most comfortable with this approach*. Early pilot tests revealed that these types of questions were the easiest for IR researchers to answer, primarily because they allowed researchers to use their own methods for mentally estimating the value of TREC resources. As a result, respondents were not required to consider what direction their research would have taken in the absence of TREC if they were unable to

[25]Although bidding games like the one used in this survey, are one of the oldest procedures used by economists for eliciting contingent valuations from respondents, it is important to recognize that this approach has limitations that cannot be controlled for. There is evidence that different starting points can affect the outcomes of the game. This is known as "starting point bias." As a result, many contingent valuation surveys prefer to use a format where multiple respondents are shown only one value, and statistical analysis is used to infer their willingness to pay. However, these approaches typically require larger sample sizes that could not be guaranteed in this study. Additionally, it was unclear what range of values should even be considered if a more discrete format was pursued. As a result, the simpler bidding game approach was decided upon, with the starting point calculated based on company data provided by participants.

do so. Instead, they could estimate the benefits they received by thinking about the costs they saved, for example.

These advantages outweigh the approach's limitations. In particular, some survey respondents and interview participants may not have been able to accurately gauge how much they would have been willing to pay for TREC accomplishments, and some survey respondents were unwilling to provide estimates. These limitations are shared by all surveys to some degree, because survey results are only useful to the extent that respondents are willing and able to provide truthful and accurate responses to the questions being asked.

Follow-up interviews were conducted with several respondents to mitigate these limitations. During these interviews, respondents were asked about the time period during which they accrued benefits from TREC (e.g., did benefits accrue evenly throughout a researchers' use of TREC resources or did they mostly accrue in the first year[s] of use?) and the methods they used to mentally estimate the value of TREC resources.

Estimates of the utilization costs for IR research organizations were calculated based on nonlabor costs, including registration fees and travel costs for TREC workshop participants. Fees paid for document sets (e.g., to NIST and LDC) were not included because the costs for creating the document sets were already included in our analysis. Labor costs were not included for two reasons. First, it is likely that some survey participants implicitly provided net benefits in their valuation estimates. Thus, the benefit estimates might actually be higher if no utilization costs were included. Second, there was no clear way to estimate what percentage of TREC participation labor should be classified as "TREC utilization costs" versus time that would have been spent conducting IR research with or without TREC resources. Utilization costs related to new technologies or standard processes typically involve both time spent learning and time spent using a new technology or standard process. However, the time spent generally does not have a separate benefit. In the case of TREC, this time was spent on research, as it would have been without TREC. It was simply spent on a more effective and efficient form of research. Thus, only nonlabor utilization costs were calculated for this study.

Utilization costs were calculated using secondary data on the registration fees and travel costs paid by TREC workshop participants. To illustrate how this was accomplished, Table 5-2 provides an overview of the typical cost of participating in a single TREC track—specifically the 2009 Web track. Similar costs were computed for each track and multiplied by the U.S. number of respondents. Registration and travel costs were only multiplied by total U.S. participation in each TREC workshop, rather than per track.

Table 5-2. Costs Associated with Participating in 2009 Web Track

	Cost ($2009)
TREC registration fee	$395
Travel cost	$1,362
Total cost per participant	**$1,757**

Note: Per-participant travel costs were calculated using data from the Government Reporting Office (GAO) and the Bureau of Labor Statistics. Specifically, data on domestic per diem rates were obtained from the GAO ($300 per day), and it was assumed that participants spent 1 full day at TREC and 2 half days (with only 75% of per diem being spent). This yields a per diem estimate of $750. In addition, average domestic flight ticket prices were obtained from the Bureau of Travel Statistics. Specifically, in 2009, the average 1-way air fare was $306. This yields a round trip estimate of $612. In total, travel costs were estimated to be $1,362.

5.2.2 End Users of IR Systems: Cost and Benefit Estimation

End users' benefits were evaluated using qualitative data. In theory, end users saved time conducting information searches based on an improvement in the quality of search results. These benefits are difficult to quantify because they accrued to a multitude of businesses in many industries, as well as to all Internet users who used web search engines.

To estimate one set of benefits end users received, overall improvement in web search products between 1999 and 2009 was estimated. IR researchers who completed the online survey provided assessments of the reduction in time spent searching for a specific item during this 10-year period as well as the percentage of this improvement attributable to TREC. The specifics of how this question was asked are described in Box 2. Information was not collected directly from end users because end users likely would have a difficult time separating out improvements in their computer hardware, Internet speed, and web search engine.

Data from OPA's Internet Activity Index were used to develop estimates of the time spent by U.S. Internet users searching online between 1999 and 2009. Assuming linear growth in the improvement in web searching during this time period, time savings estimates for home users were estimated.

It should be noted that the resulting estimates are only an approximation of the time savings that resulted from improvements in web searches. In reality, if web search engine performance had not improved, people would have likely just spent less time using web search engines rather than spending more time using inferior systems.

Such analysis results should be viewed cautiously for a number of reasons and, thus, are not included in the quantitative estimates of the economic benefits attributable to TREC. Percentage improvements in performance do not easily translate into technical metrics that can be monetized easily. For example, the percentage improvement in an IR system may not directly equal the percentage of labor hours avoided conducting additional searches using less efficient IR systems. It is not clear if respondents were able to provide accurate estimates of the

Box 2

Estimating the Impact of TREC on Web Search Engines

"Do you believe that the time it takes to fill an information need using web search engines has decreased since 1999, holding improvements in hardware and nonsearch software constant? For example, if you used a web search engine from 2009 to search the Internet of 1999 using a computer made in 1999, do you think you could fill an information need in less time than if you had used a search engine from 1999?
<Yes/No>

> If yes, what percentage change (faster speed) do you think you would have observed, on average, using today's search engines on a 1999 computer looking at the 1999 web? (NOTE: Although you might find this question difficult to calculate, we are looking for your best estimate based on your expert knowledge of the field of IR.)
>
> > Select One
> > 0%, 1-4%, 5-9%, 10-14%, 15-19%, 20-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, 90-99%, 100-199%, 200-299%, 300-399%, 400-499%, 500-99%, 1000+%
>
> What portion of this time savings would you attribute to improvements in web search quality enabled by TREC's workshops and resources?
>
> > Select One
> > 0, 1-4%, 5-9%, 10-14%, 15-19%, 20-29%, 30-39%, 40-49% , 50-59%, 60-69%, 70-79%, 80-89%, 90-99%"

percentage improvement in search as a function of time or accurate estimates of what portion of this improvement is due to TREC. It is very difficult to attribute incremental improvements in system performance to any one factor, and significant uncertainty surrounds these figures. Further, to estimate the net economic benefits to end users, one would need to have accurate information on the difference in the price that end users paid to use IR systems as it compared with what they would have paid for IR search products if TREC had not existed. However, the actual and estimated prices vary significantly depending on the types of search products or services the end user needs. In this study, web search engine improvement was studied because web search engines are widely used and because price is not a factor in analyzing usage impacts given that most web search engines are free to use.

5.3 Extrapolating Survey Results to National Net Benefit Estimates

For private-sector respondents, the study extrapolated respondents' benefits valuations using an estimate of total industry IR-related R&D expenditures. To illustrate this extrapolation strategy, consider the following simple example. Suppose that each year firms involved in developing IR systems spend $100 billion on IR-related R&D and that completed surveys were received from firms that represent 80% of this total and their responses indicated that TREC resources contributed an average of $0.01 of value per dollar of total IR-related R&D expenditures ($800 million). To estimate the benefits received by companies not responding to the survey, the same benefit per dollar of R&D ratio was applied to the remaining $20 billion. In this example, it can be estimated that nonrespondents received approximately $200 million in benefits from using TREC resources ($0.01 x $200 billion). By contrast, benefits estimates for academic respondents were extrapolated using a similar method, except rather than using IR-related R&D, publicly available total R&D estimates associated with computer science departments were used.[26]

The first challenge to using this type of extrapolation strategy is properly identifying the companies and universities that will comprise the extrapolation base. To create this universe, the major organizations conducting IR research were identified, which included both private firms and academic institutions. Two primary sources to identify these organizations were used:

1. NIST provided a list of academic institutions and public companies that participated in TREC workshops between 1992 and 2009.
2. Publicly available information was used on the proceedings of the Association of Computing Machinery (ACM) SIGIR to identify organizations presenting IR research at SIGIR's annual conference.

These sources were supplemented by looking at information on attendance at other IR-related conferences (such as the 2009 Enterprise Search Summit) and by looking at industry reports for several submarkets related to IR (such as Gartner's 2008 report on the market for information access technologies and software). A list of all organizations included in our extrapolation base is provided in Appendix B.

After the universe was compiled, data were collected on the R&D expenditures for the organizations included. These data were collected from different sources and in different ways based on whether the organization was an academic institution or a private firm:

1. For academic institutions, data on R&D expenditures in computer sciences were obtained from the NSF Survey of Research and Development Expenditures at Universities and Colleges from 1992 to 2008.[27] R&D expenditures in computer sciences were used because academics conducting computer science research are the most likely to use TREC resources.
2. For private firms, several steps were performed to estimate total IR-related R&D. First, two methods were used to obtain estimates of *total* R&D expenditures. Data were

[26]Sales and employment data are traditionally used to extrapolate benefits; however, TREC most directly impacts the R&D activities of adopting companies and academic organizations. Therefore, R&D is the most appropriate metric for extrapolation.

[27]At the time of this study, R&D expenditure data from NSF were only available up to 2008. Therefore, for the purposes of extrapolation, it was assumed that R&D expenditures in 2009 were the same as the previous year.

collected on public companies that reported R&D expenditures through the S&P North American COMPUSTAT database provided through the Wharton Business School.[28] Then, R&D expenditures were estimated for public companies that did not report R&D expenditures by multiplying the sales of these companies (reported in the COMPUSTAT database) by the average R&D to the sales ratio for companies that did report R&D expenditures. This is also the method used to calculate R&D expenditures for private firms (however, sales data were instead collected for these companies from MANTA.com).[29]

After estimates of total R&D expenditures for each company were obtained, IR-related R&D was estimated using data collected from survey respondents. Specifically, as part of the survey, respondents were asked to estimate the total number of IR R&D workers currently in their company. These responses were combined with estimates of total R&D employment for each company that provided this information; R&D employment was estimated using publicly available data on total employment, sales, and total R&D—it was assumed that R&D employment was the same proportion to total employment that R&D expenditures were to sales. Specifically, the ratio of IR R&D workers to total R&D workers were estimated and assumed that IR R&D spending was proportionally the same to total R&D expenditures. This same ratio to total R&D was applied for all years from 1992 to 2009. For companies that did not respond to the survey or did not provide estimates of total IR workers employed in their organization, the ratios estimated were applied to similar companies that did respond to the survey.

After sample benefits were extrapolated using the data collected, the national benefits that IR researchers received from TREC were compared with national utilization costs and the costs of administering the TREC Program itself, which were obtained directly from NIST and other supporting organizations. Figure 5-1 provides a conceptual overview of how these costs and benefits accrued for a single track (the Ad Hoc track). This figure shows that the TREC Program costs grew and then dissipated during the time period when NIST was operating the track, as did users' utilization costs. However, the benefits associated with using the test collections, evaluation methods, and research generated by this track likely continued for some time afterward. It should be stressed that this figure only reflects hypothetical costs and benefits for a single track and that this study did not attempt to break down costs and benefits at the track level.

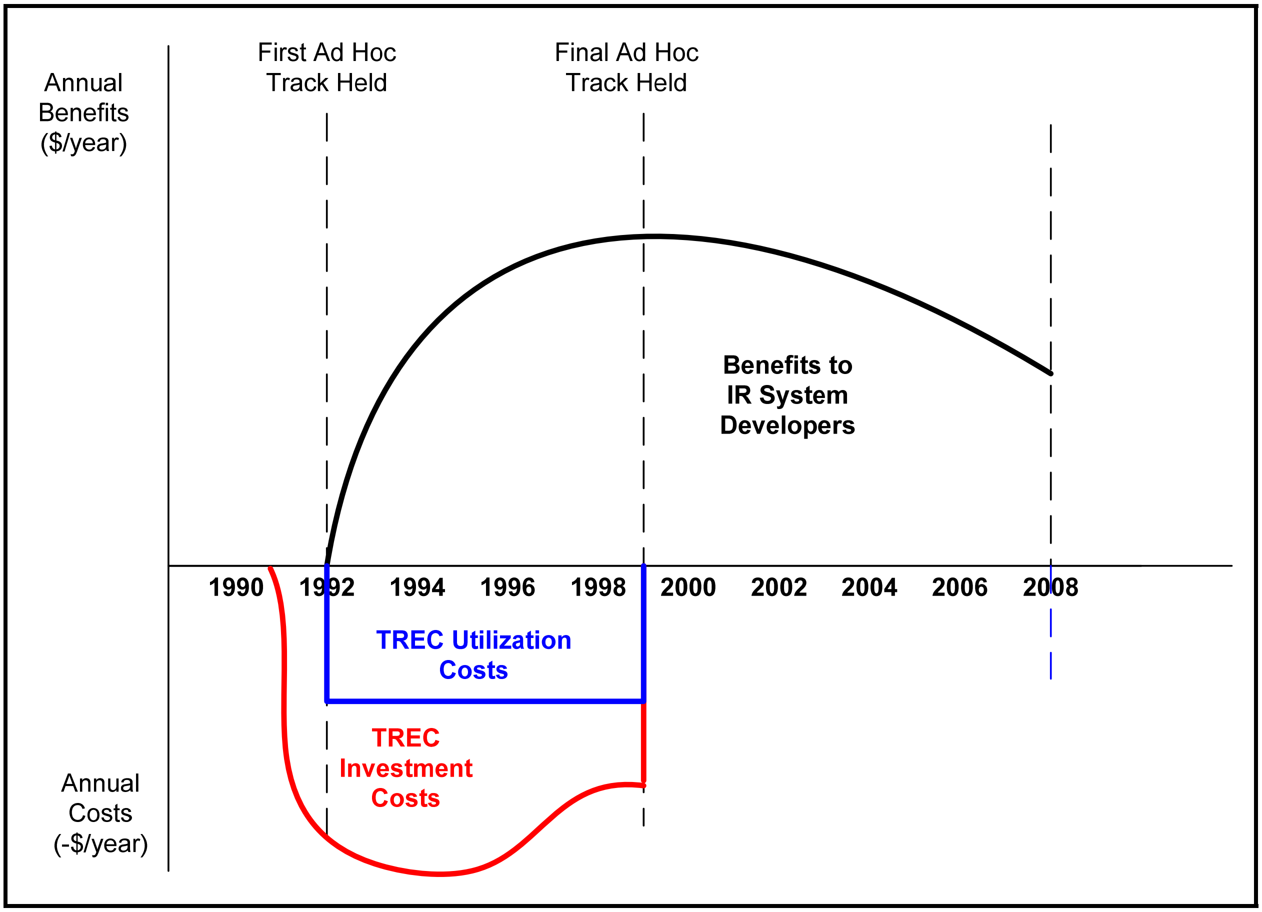
5.4 Summarizing Measures of Economic Return

After the benefits of TREC were estimated relative to the counterfactual, three benchmark measures were generated—benefit-to-cost ratio (BCR), net present value (NPV), and internal rate of return (IRR). These metrics are defined below; the resulting figures for the TREC analysis are presented in Section 6.

[28] At the time of this study, R&D expenditure data from COMPUSTAT were only available up to 2008. Therefore, for the purposes of extrapolation, it was assumed that R&D expenditures in 2009 were the same as the previous year.

[29] Since the aim of this study is to estimate only the U.S. economic contributions of TREC, the ideal variable to use when creating the cut-off sample for private companies would be total U.S. R&D expenditures in the computer sciences. However, since most companies do not report R&D expenditures by location or research area, RTI used total company R&D expenses as a proxy for preparing the sample.

Figure 5-1. Example Comparison of National Costs and Benefits for the Ad Hoc Track



5.4.1 Benefit-to-Cost Ratio

The BCR calculated in this analysis is the ratio of the NPV of benefits to the NPV of costs, which accounts for differences in the timing of cash flows (which has implications for the real value of $1 in one time period versus another).

Letting B_t be the benefits that accrued in year *t* to firms and academics and letting C_t be the total costs for TREC in year *t* including TREC investment costs and IR researchers' utilization costs, then the BCR for the Program is given by

$$BCR = \frac{\sum_{i=0}^{n} \frac{B1_{(t+i)}}{(1+r)^i} + \sum_{i=0}^{n} \frac{B2_{(t+i)}}{(1+r)^i}}{\sum_{i=0}^{n} \frac{C_{(t+i)}}{(1+r)^i}} \qquad (5.1)$$

where

t is the year in which benefits or costs occur,

n is the number of years the benefits and/or costs occur, and

r is the social discount rate.

In this study, r was set at 7%, the OMB-specified level.[30] Because costs and benefits occur at different time periods, both are expressed in present value terms before the ratio is calculated. Essentially, a BCR greater than 1 indicates that quantified benefits outweigh the calculated costs. A BCR less than 1 indicates that costs exceeded benefits, and a BCR equal to 1 means that the project broke even.

5.4.2 Net Present Value

The NPV of the investment in a project is calculated as

$$NPV = \sum_{i=0}^{n} \left[\frac{B1_{(t+i)}}{(1+r)^i} + \frac{B2_{(t+i)}}{(1+r)^i} - \frac{C_{(t+i)}}{(1+r)^i} \right] \tag{5.2}$$

where the terms have the same meanings as identified for Equation (5.1). Any project that yields a positive NPV is considered economically successful. Government projects that show a positive NPV when analyzed using OMB's 7% real discount rate are socially advantageous. A negative NPV would indicate that the costs to society outweigh the benefits, and an NPV equal to zero would indicate a breakeven point.

5.4.3 Internal Rate of Return

The IRR on an investment should be interpreted as the percentage yield on an R&D project over the life of the project, often multiple years. In mathematical terms, the IRR is the value of r that sets the NPV equal to zero in Equation (5.2) or results in a BCR of 1 in Equation (5.1). It should be noted that, in cases for which costs exceed benefits, an IRR cannot be calculated.

The IRR's value can be compared with conventional rates of return for comparable or alternative investments. Risk-free capital investments such as government bonds can be expected to yield rates of return under 5% in real terms, while equities seldom return more than 10% over an extended period of time. In cases such as TREC where government activities are associated with public goods, the IRR is also referred to as the "social rate of return." As described in Tassey (2003), academic studies of the diffusion of new technologies have found that IRRs of 100% or more have been associated with significant accomplishments with broad social benefit. Based on a variety of economic studies, the hurdle rate for rationalizing such public-good investments is in the 30 to 50% range.

[30]See OMB Circular A-94.

6. ECONOMIC ANALYSIS RESULTS

This chapter presents the results of an economic analysis conducted to evaluate the net benefits of NIST's investment in TREC resources and activities. Because not all economic impacts could be monetized, qualitative and quantitative metrics are presented. Specifically, the following economic impact measures are discussed:

- Qualitative metrics describing IR researchers' use of TREC resources (in Section 6.1).
- Qualitative metrics describing the impact of TREC on end users of IR products (in Section 6.2).
- Quantititatve estimates of the net economic benefits IR research organizations received from TREC (in Section 6.3).

As described in Section 5, the information used to support this analysis was largely collected through the 2010 TREC web survey. At the completion of the survey, RTI received 404 responses, 93 of which came from individuals indicating that they were employed by U.S. organizations.[31] Table 6-1 provides a breakdown of these 93 respondents by stakeholder group. Thirty percent of respondents (28 individuals) were located at U.S.-owned private companies conducting IR research. These companies represent 58% of the total 2008 R&D expenditures of those companies included in our private-sector extrapolation base, used to proxy for total private-sector IR research expenditures. Similarly, 66% of survey respondents (61 individuals) were employed by U.S. universities or academic research laboratories. These universities represented 47% of the total 2008 research expenditures of universities included in our academic extrapolation base, used to proxy for total academic-sector IR research spending.

Table 6-1. Total U.S. Respondents by Stakeholder Group

	Number	% of Total U.S. Respondents
Information retrieval service or software company	18	19%
Other software company	4	4%
University or academic research laboratory	61	66%
Government or institutional research laboratory[a]	4	4%
Other—for profit	6	6%
Other—nonprofit	0	0%
Total	93	100%

[a] Of the four government/institutional research laboratories that responded to the survey, only one provided valuation estimates that could be used in assessing benefits. To simplify the presentation of analysis results, the valuation data provided by this single respondent were added to the valuation estimates for academic respondents.

[31] One hundred fifty-three respondents indicated that they were employed by foreign organizations. One hundred fifty-eight did not indicate their employers' name or location. Our analysis used data from the 93 respondents who indicated they worked for U.S. organizations.

6.1 Qualitative Benefits of TREC: IR Researchers

TREC's impact was most strongly felt by IR researchers—both TREC workshop participants and other IR researchers who used TREC's test collections, evaluation methods, and research papers. Table 6-2 reports the results of this assessment for U.S. respondents. The vast majority (over 95%) of respondents indicated that they had used TREC test collections, evaluation methods, and research papers at some point in time. However, the percentage who had attended a TREC workshop was somewhat lower—approximately 83% of respondents.

After it was determined which resources the IR researchers used, IR researchers were asked a series of questions to determine how important these resources were to their research (Table 6-3). Of the respondents who attended TREC,

- 55% of the researchers who attended TREC workshops found them to be very important to their research, while 39% found them to be somewhat important to their research;
- 85% and 82% of those who used the test collections and evaluation methods, respectively, indicated that they were very important to their research; and
- 66% of those who read the research papers rated them as very important, while 31% rated them as somewhat important.

Although many survey repsondents likely participated in the survey because they benefited from TREC, and thus the sample was somewhat biased, it is an important finding that of those who used any TREC resources, over 50% found them to be very important to their research and less than 6% found them to be not very important to their research.

Respondents also identified which tracks they found to be most beneficial for their research (see Table 6-4). As shown in the far left column beside each track name, not all respondents used every track—in some cases the number ("N") of users was as low as 21 (genome), or 22% of respondents, while, in other cases, the number of users was as high as 62 (Web), or 65% of respondents. The percentages in Table 6-4 are based only on the number of people using each track. For example, of the 61 respondents who used resources from the Ad Hoc track, 77% believed the track was very imporant for their IR research.

Table 6-2. Use of TREC Resources

	Number of Respondents Not Using Product or Service	Number of Respondents Using Product or Service	No Response	Total Respondents
Workshops	15	77	1	93
Test collection(s)	1	91	1	93
Evaluation method(s)[a]	3	89	1	93
Research papers	0	91	2	93

[a] Evaluation methods refer to the performance metrics (such as Mean Average Precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

Table 6-3. Perceived Benefit of TREC Resources (as a Percentage of Users)

	Not Very Important for IR Research	Somewhat Important for IR Research	Very Important for IR Research
Workshops (N = 77)	6%	39%	55%
Test collection(s) (N = 91)	3%	12%	85%
Evaluation method(s)[a] (N = 89)	3%	15%	82%
Research papers (N = 91)	3%	31%	66%

[a] Evaluation methods refer to the performance metrics (such as mean average precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

Table 6-4. Perceived Benefit of Individual TREC Tracks

	Not Very Important for IR Research	Somewhat Important for IR Research	Very Important for IR Research
Ad hoc (N = 61)	5%	18%	77%
Blog (N = 43)	9%	51%	40%
Cross-language (N = 37)	22%	38%	41%
Enterprise (N = 29)	31%	41%	28%
Filtering/Routing (N = 46)	17%	33%	50%
Genome (N = 21)	33%	14%	52%
Interactive (N = 35)	20%	54%	26%
Legal (N = 32)	25%	38%	38%
Question answering (N = 57)	11%	35%	54%
Spam (N = 25)	44%	40%	16%
Speech (N = 26)	42%	27%	31%
Video/TRECVid (N = 27)	15%	22%	63%
Web (N = 62)	0%	26%	74%

Of particular note, the Ad Hoc track was both highly used (65% of respondents used information from this track) and was ranked by almost 80% of respondents using information from the track as being very important to their research. The Web track had similar metrics—67% of respondents used the information from this track, and 74% rated these resources as very important. In contrast, the Spam track was used by only 25 people, and of those, only 16% rated the associated resources as being very important, with 44% rating these resources as not very important. The Enterprise track received similar ratings with only 31% of respondents having used the resources and 28% rating these resources as not very important.

Table 6-5 summarizes additional perceptions of the benefits of TREC. Note that

- 75% of survey respondents (70 individuals) have published papers using TREC test collection data,
- 85% (79 individuals) have referenced a TREC paper in a peer-reviewed journal article they wrote or a paper they presented at a conference,
- 14% (13 individuals) have referenced a TREC paper in a patent filing,
- 47% (44 individuals) have used TREC test collections for tasks not studied at TREC (these tasks ranged from using TREC test collections to test algoritms to evaluating document download speeds), and
- 71% (66 individuals) have used TREC evaluation methods with non-TREC test collections to evaluate the performance of IR systems.

These responses suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study's economic benefits calculations.

Interviews also captured information about the improvement in human capital attributable to TREC. According to one researcher working in a web search services company, being able to hire individuals who have used TREC resources offers great value because they have hands-on experience with how IR systems work, making them more competent as employees.

Further, based on information collected during interviews and surveys, TREC had a substantive impact on the interest of individuals to pursue graduate education in IR and the human capital that developed as a result of that education. Two questions of academic respondents were asked to evaluate how they percieved TREC's impact on the pursuit of and return on graduate education. Table 6-6 provides the results of this question. We found that 67% of the 61 academic respondents believed that the number of individuals pursuing a doctoral or master's degree in a related field had increased as a result of TREC. Similarly, we found that 84% of academic respondents believed that participation in TREC workshops as a graduate student improved their employment prospects upon graduation.

Table 6-5. Use of TREC Resources (N = 93)

	Number	% of Total
Respondents publishing papers using TREC test collections	70	75%
Respondents using test collection data from TREC for tasks not studied at TREC	44	47%
Respondents using TREC evaluation methods to study the performance of an IR system using non-TREC test collections	66	71%
Respondents who have ever referenced a paper from TREC in a peer-reviewed journal article or a paper presented at conference	79	85%
Respondents who have referenced TREC papers in patent filing	13	14%

Table 6-6. Perceived Impact of TREC on Graduate Education (N = 61)

	Yes	No	No Response
Do you believe the number of individuals pursuing doctoral and master's degrees has increased as a result of TREC?	67%	28%	5%
Do you believe that participation in TREC workshops improved the employment prospects of graduate students in your department?	84%	7%	10%

6.2 Qualitative Benefits of TREC: End Users

This study's research shows that web search products, such as those develped by Google, Yahoo!, and Microsoft, have improved significantly over the last 10 years, and TREC played a significant role. End users of other IR products and services have also benefited from the costs savings and quality improvements in IR systems that resulted from TREC. However, this study determined that end users would be unable to differentiate the improvement in their computer performance and Internet speed from improvements in web search products.

IR researchers responding to our survey were asked to estimate the percentage improvement in web search products between 1999 and 2009 in terms of the time an average user would spend searching for a specific item. The survey also asked respondents to estimate the percentage of this improvement attributable to TREC.

On average, respondents to the survey estimated that end users of web search products would have been able to fill an information need 215% faster in 2009 than in 1999 as a result of improvements in web search engine performance. In other words, information needs could be filled in approximately half the time with newer web search engines. Respondents on average also estimated that 32% of this improvement was enabled by TREC Program activities.

To better understand what these improvements mean in real terms, they were reviewed relative to the number of hours individuals spent using web search applications during this time period (1999 to 2009). Table 6-7 reports that, based on data from the Online Publishers Associations's Internet Activity Index, U.S. Internet users spent approximately 8.6 billion hours using search applications over this time period.

Under the counterfactual scenario, without improvements in web search applications, U.S. Internet users likely would have attempted to fill the same information needs over this time period using web search engines that did not experience any improvement between 1999 and 2009.

Table 6-7. Time Spent on the Internet and Using Web Search Engines in the U.S., by Year

Year	Total Time Online (thousands of hours)	Time Spent Using Search Applications (thousands of hours)	% of Total
1999	8,386,405 (E)	176,809 (E)	2.1%
2000	9,550,538 (E)	221,491 (E)	2.3%
2001	10,876,266 (E)	277,466 (E)	2.6%
2002	12,386,021 (E)	347,587 (E)	2.8%
2003	14,604,340	487,631	3.3%
2004	17,397,446	721,630	4.1%
2005	18,265,259	833,673	4.6%
2006	21,337,053	1,069,135	5.0%
2007	27,272,623	1,300,331	4.8%
2008	28,953,731	1,503,726	5.2%
2009	31,343,245 (E)	1,702,202 (E)	5.4%
Total	200,372,927	8,641,682	4.3%

Source: OPA, 2010.

(E): The Internet Activity Index provides monthly reports of time spent online and time spent using search applications from May 2003 to May 2009 (because of a change in data collection methods, data collected after May 2009 are not directly comparable to data collected in prior months). Estimates of the total number of hours individuals spent online and the number of hours spent using search applications prior to May 2003 and after May 2009 were made by RTI assuming that hours spent online and hours spent using search applications grew at the same average rate each month as they did from May 2003 to May 2009.

Assuming their intended information needs remained unchanged, U.S. Internet users would have spent approximately 9.9 billion additional hours using web search products, for a total of 18.6 billion hours that would have been spent using search applications between 1999 and 2009. Assuming that 32% of the improvement in web search engines was enabled by TREC activities, without TREC, U.S. Internet users would have spent up to 3.15 billion additional hours using web search engines between 1999 and 2009.[32]

6.3 Quantitative Benefits of TREC: IR Researchers

As described in Section 5, a bidding system was used in the web survey to elicit respondents' estimates of how much value TREC generated for their organizations over the past 17 years. Tables 6-8 and 6-9 provide summaries of responses aggregated across private-sector and academic/nonprofit respondents, respectively. These tables report valuation estimates both before ("reported benefits") and after extrapolation ("extrapolated benefits") to industry-wide

[32]We did not quantify the economic impact of this benefit for several reasons. First, both the estimated percentage improvement in web search time and and estimated percentage attributable to TREC are highly subjective. Second, individuals might not have been willing to spend more time using web search engines to find a product; therefore, the estimated time savings might have been spent on other activities with higher or lower utility to the individuals.

Table 6-8. Total Benefits to Private-Sector Stakeholders

Year	Reported Benefits ($2009)	Additional Extrapolated Benefits ($2009)	Total Benefits ($2009)
1991	—	—	—
1992	—	—	—
1993	$2,257,431	$2,871,717	$5,129,148
1994	$2,257,431	$2,782,116	$5,039,548
1995	$2,257,431	$2,617,147	$4,874,578
1996	$2,262,431	$1,833,292	$4,095,724
1997	$3,281,929	$2,106,801	$5,388,730
1998	$8,611,443	$4,296,793	$12,908,236
1999	$7,585,337	$3,229,508	$10,814,845
2000	$7,585,337	$2,903,841	$10,489,178
2001	$8,710,560	$2,810,799	$11,521,358
2002	$6,045,803	$1,763,620	$7,809,423
2003	$5,900,160	$1,505,048	$7,405,209
2004	$5,645,799	$946,012	$6,591,812
2005	$5,687,748	$1,243,788	$6,931,536
2006	$5,533,068	$1,238,524	$6,771,591
2007	$5,549,085	$1,205,122	$6,754,207
2008	$5,534,085	$1,133,817	$6,667,901
2009	$5,534,085	$1,133,817	$6,667,901
Total	**$90,239,163**	**$35,621,762**	**$125,860,925**

estimates. Reported benefits are only the sum of benefits estimated directly by respondents to the survey. Extrapolated benefits represent an estimate of the additional benefits that acrued to the population of IR researchers who did not report benefits in the private sector and academia. In the case of private-sector benefits, extrapolating benefits increases total benefits by 39%. For academic and nonprofit respondents, which are less concentrated, extrapolating benefits increases total benefits by 83%.

To better understand the valuation estimates provided by respondents, post-survey interviews were conducted with key respondents. Although a variety of factors were stated as having influenced the valuation estimates, two main approaches were mentioned most often:

Table 6-9. Total Benefits to Academic and Government Laboratories Stakeholders

Year	Reported Benefits ($2009)	Additional Extrapolated Benefits ($2009)	Total Benefits ($2009)
1991	—	—	—
1992	$744,214	$432,697	$1,176,911
1993	$802,222	$488,887	$1,291,109
1994	$845,804	$518,041	$1,363,844
1995	$973,581	$633,348	$1,606,929
1996	$1,017,440	$780,705	$1,798,144
1997	$1,024,795	$700,740	$1,725,535
1998	$655,463	$472,916	$1,128,378
1999	$628,051	$435,309	$1,063,360
2000	$596,495	$571,810	$1,168,306
2001	$734,458	$681,743	$1,416,201
2002	$731,979	$606,703	$1,338,682
2003	$870,704	$696,445	$1,567,149
2004	$814,994	$709,004	$1,523,999
2005	$894,078	$833,735	$1,727,813
2006	$950,808	$948,195	$1,899,002
2007	$854,716	$864,028	$1,718,743
2008	$852,478	$957,093	$1,809,571
2009	$852,478	$957,076	$1,809,554
Total	**$14,844,757**	**$12,288,474**	**$27,133,231**

1. Some respondents estimated the costs (typically IR researchers' labor) that their organization would have spent to achieve the same level of IR system quality in the same time frame without TREC.
2. Some respondents estimated the costs (typically IR researchers' labor) that their organization would have wasted on unsuccessful research, training, and recruitment without TREC.

Respondents from private firms typically thought about the answer to the valuation question generally in terms of Approach (1) above—how much would they have spent to achieve the same level of IR system quality in the same schedule without TREC. Table 6-10 provides a summary of the types of cost metrics that private firms said they avoided as a result of TREC. Private-sector respondents indicated that they added up the number of labor hours that their company would have needed to spend on each activity listed in Table 6-10 to achieve the same

Table 6-10. Costs Avoided Because of TREC: Interpreting Valuation Estimates, by Stakeholder Group

TREC Output	Private Firms' Avoided Costs	Academic Institutions' Avoided Costs
Large test collections	Labor costs, intellectual property rights costs, and hardware costs to identify, develop, and store alternate test collections of *equal size/quality* to TREC	Labor costs wasted as a result of using test collections of *insufficient size/quality*
Standardized evaluation methods	Labor costs spent to achieve *similar research results* without TREC's standardized evaluation methods	Labor costs wasted on less effective research strategies, resulting in *lower quality research outcomes*
TREC workshops	Labor costs spent to learn new techniques (successes and failures) without TREC workshops and associated papers	Labor costs wasted working on previously tested (though not published) IR techniques

results without TREC as they did with TREC. Some private companies indicated that creating data sets of equal size to TREC's would have been the most costly activity; others (e.g., web companies) indicated that the TREC methodologies and workshops provided an incredibly valuable educational resource that would have been very expensive to gain otherwise. Specifically, many private firms indicated that without TREC they would have spent whatever was necessary to develop the same quality products and services in the same time frame.

Academics mostly indicated that their answers were based on Approach (2) above—how much time would they have wasted without the use of TREC resources. Most academic organizations could not have devoted the resources to develop sufficiently large test collections or robust methodologies like TREC. Table 6-10 provides a summary of the types of benefit metrics (in terms of reduced wasted labor) that academic organizations stated they considered when providing valuation estimates. Academic organizations placed particular value on the TREC test collections; very few academic researchers had the monetary resources to create test collections (labor costs), secure intellectual property rights (fees), or store (hardware costs) large test collections to be used for IR research. Those who did have such resources (e.g., from DARPA) indicated that they benefited more heavily from the standard evaluation methods developed by TREC and the workshops organized by TREC. As a whole, academic organizations indicated that without TREC they would not have been able to spend additional resources; therefore, they would have wasted a significant amount of time using lower quality data sets and evaluation methods and repeating unsuccessful research that others had already undertaken.

6.4 TREC Investment Costs and Participants' Utilization Costs

Total TREC Investment costs were estimated to be $41.7 million, which came from a variety of sources (Table 6-11):

Table 6-11. Total TREC Support Costs

Year	Total NIST Program Funding (thousands $2009)	Total DARPA Funding (thousands $2009)	Other Government Funding (thousands $2009)	Non-government Costs (thousands $2009)	TOTAL TREC Investment Costs (thousands $2009)
1991	—	$715	—	$38	$753
1992	—	$694	—	$19	$713
1993	—	$674	—	—	$674
1994	$579	$740	—	$203	$1,522
1995	$738	$456	—	$88	$1,282
1996	$1,221	$786	—	$121	$2,129
1997	—	—	—	$61	$61
1998	$981	$32	$538	$188	$1,739
1999	$805	$35	$869	$139	$1,848
2000	$589	$503	$678	$73	$1,844
2001	$523	$455	$540	$25	$1,544
2002	$570	$485	$971	$146	$2,173
2003	$583	—	$1,151	$146	$1,880
2004	$510	—	$1,038	$86	$1,634
2005	$833	—	$1,117	$193	$2,143
2006	$422	—	$1,275	$90	$1,788
2007	$637	—	$1,002	$29	$1,668
2008	$919	—	$850	$213	$1,982
2009	$1,279	$292	$75	$25	$1,671
Total	$11,189	$5,868	$10,104	$1,886	$29,046

Note: All dollar values were inflation-adjusted to 2009 dollars using the Consumer Price Index, U.S. City Average for all Items.

- NIST funding of approximately $11.2 million between 1991 and 2009,
- DARPA funding of $5.9 million, mainly as a part of the TIPSTER Program,
- Other government agency funding, including National Security Agency and the Central Intellegicence Agency, totaling $10.1 million, and
- Other nongovernment organizations' costs, such as those incurred by the University of Waterloo and LDC, of $1.9 million.[33]

[33]In addition to direct government funding of the TREC Program, other organizations have also incurred costs to develop test collections and evaluation methodologies. For example, entities such as the University of Waterloo and the LDC incurred costs to develop test collections that were used for TREC tracks such as the Spam track and the Ad Hoc track. These costs comprise the nongovernment costs category in Table 6-11. To estimate these

In addition to the costs associated with creating test collections and evaluation methodologies and running the TREC Program, costs were also incurred directly by the IR researchers who used TREC resources. For the purposes of this analysis, these are referred to as "utilization costs." Specifically, costs are associated with attending TREC workshops (such as the registration fees and travel costs).[34] Costs incurred by IR researchers to obtain TREC test collections were not included since the creation costs of these data sets were already part of the cost estimates. Estimates of the quantified utilization costs are provided in Table 6-12.

Table 6-12. TREC Participants' Utilization Costs

Year	Estimated Cost of Attending TREC Workshops ($2009)
1992	$18,699
1993	$23,089
1994	$24,717
1995	$27,094
1996	$28,843
1997	$38,094
1998	$42,257
1999	$51,037
2000	$53,664
2001	$67,828
2002	$72,304
2003	$72,529
2004	$79,405
2005	$93,622
2006	$87,226
2007	$77,676
2008	$45,683
2009	$66,063
Total	**$969,830**

costs, several organizations involved in creating TREC test collections and evaluation methodologies over the past 17 years were asked for information about the resources they devoted to creating these collections. To create each test collection, generally a document set was created by a university or an organization such as LDC. These document sets could take hundreds of hours to create, adding up to hundreds of thousands of dollars in some cases. Next, topics and relevance judgments had to be created. These were typically created by NIST. However, in some cases subject matter expertise was required, so the topics and relevance judgments had to be created outside of NIST. The study assumed it took approximately 20 hours to create these topics and test collections. Based on these conversations, total nongovernment costs associated with TREC test collections were estimated to be approximately $1.88 million since 1990.

[34]The fees paid by researchers to attend TREC conferences were used to pay for meeting costs, such as NIST conference staff overhead, meals, buses, etc. These costs were not included in the NIST investment costs described previously.

6.5 Summary Economic Benefit Estimates and Measures of Return

Table 6-13 assembles the complete time series of quantified costs and benefits for the period between 1991 and 2009, and Table 6-14 provides performance measures. Total extrapolated benefits were nearly $153 million for private, academic, and nonprofit organizations, and total unextrapolated benefits were $105 million. Inclusive of TREC Program and resource utilization costs, net extrapolated benefits were $123 million and net unextrapolated benefits were $75 million.

Table 6-13. Net Quantified Economic Benefits of TREC

	Cost					
Year	Total TREC Investment Costs (thousands $2009)	Total Utilization Costs (thousands $2009)	Total Unextrapolated Benefits (thousands $2009)	Total Extrapolated Benefits (thousands $2009)	Net Unextrapolated Benefits (thousands $2009)	Net Extrapolated Benefits (thousands $2009)
1991	−$753	—	—	—	−$753	−$753
1992	−$713	−$19	$744	$1,177	$12	$445
1993	−$674	−$23	$3,060	$6,420	$2,363	$5,723
1994	−$1,522	−$25	$3,103	$6,403	$1,556	$4,857
1995	−$1,282	−$27	$3,231	$6,482	$1,922	$5,172
1996	−$2,129	−$29	$3,280	$5,894	$1,122	$3,736
1997	−$61	−$38	$4,307	$7,114	$4,208	$7,015
1998	−$1,739	−$42	$9,267	$14,037	$7,486	$12,255
1999	−$1,848	−$51	$8,213	$11,878	$6,315	$9,980
2000	−$1,844	−$54	$8,182	$11,657	$6,285	$9,760
2001	−$1,544	−$68	$9,445	$12,938	$7,833	$11,326
2002	−$2,173	−$72	$6,778	$9,148	$4,533	$6,903
2003	−$1,880	−$73	$6,771	$8,972	$4,818	$7,020
2004	−$1,634	−$79	$6,461	$8,116	$4,748	$6,403
2005	−$2,143	−$94	$6,582	$8,659	$4,345	$6,423
2006	−$1,788	−$87	$6,484	$8,671	$4,609	$6,796
2007	−$1,668	−$78	$6,404	$8,473	$4,658	$6,727
2008	−$1,982	−$46	$6,387	$8,477	$4,359	$6,450
2009	−$1,671	−$66	$6,387	$8,477	$4,649	$6,740
Total	−$29,046	−$970	$105,084	$152,994	$75,068	$122,978

Note: All dollar values were inflation-adjusted to 2009 dollars using the Consumer Price Index, U.S. City Average for all Items.

Table 6-14. Performance Measures

Measure	Value
Unextrapolated Performance Measures	
Total quantified benefits (Discounted, Base Year = 1991)	$53,267,846
Total quantified costs (Discounted, Base Year = 1991)	−$15,916,193
Net present value of net benefits (NPV) (Base Year = 1991)	$37,351,653
Benefit-to-cost ratio (BCR)	3.35
Internal rate of return (IRR)	128%
Extrapolated Performance Measures	
Total quantified benefits (Discounted, Base Year = 1991)	$80,655,082
Total quantified costs (Discounted, Base Year = 1991)	−$15,916,193
Net present value of net benefits (NPV) (Base Year = 1991)	$64,738,889
Benefit-to-cost ratio (BCR)	5.07
Internal rate of return (IRR)	250%

Note: All dollar values were inflation-adjusted to 2009 dollars using the Consumer Price Index, U.S. City Average for all Items. Benefits and Costs were discounted using the 7% real social discount rate recommended by OMB.

The NPV of extrapolated net benefits is $65 million and $37 million for unextrapolated net benefits, applying the OMB-approved discount rate of 7%. The BCR ratio, which is the ratio of the NPV of total benefits to that of costs, was estimated to be 5.07 for extrapolated benefits and 3.35 for unextrapolated benefits. In other words, for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers. The IRR was estimated to be 250% for extrapolated benefits and 128% for unextrapolated benefits.

7. CONCLUSIONS

In 1990, IR research was largely being conducted by a handful of companies and universities whose techniques could not be easily compared. Companies operated in silos, developing new techniques on their own without incorporating ideas from academia or comparing their products with other companies' products. Furthermore, only a small number of academics had access to sufficiently large data sets and the funding necessary to store the data and conduct substantive research. However, even this group lacked metrics and evaluation methodologies needed to describe the performance of their IR systems.

TREC transformed this landscape by providing large test collections at relatively low cost, developing and publishing robust evaluation methodologies, and creating a competition through which researchers could objectively compare IR systems and discuss the results (successes and failures). TREC reduced the costs for IR research and opened the door to more robust IR system development. Without TREC, many IR researchers would not have been able to conduct the breadth and depth of research that TREC's resources enabled; in particular, creating a test collection was (and to a lesser extent, still is) a very expensive task, in terms of labor costs, hardware and software costs for storage, and copyright costs. Further, the competition that TREC created facilitated an atmosphere where specific points of failure were discussed; this is very uncommon even in academic circles, but TREC was able to motivate the research community to be very open in this setting. As a result of these forums for objective evaluation and information sharing, companies could see which IR techniques were most successful and integrate these findings into their products, thus benefiting these companies and their customers.

As described in Section 6, $16 million of discounted investments have made by NIST and others in TREC have resulted in $81 million in discounted extrapolated benefits or a net present value of $65 million. TREC's activities also had other benefits that were not quantified in economic terms. TREC helped educate graduate and undergraduate students, some who went on to lead IR companies and other who stayed in academia to teach and conduct research. TREC benefited IR product quality and availability—our research suggests that TREC motivated a large expansion in IR research that has enabled high quality applications such as web search, enterprise search, and domain-specific search products and services (e.g., for genomic analysis). More specifically, as described in Section 6, this study estimates that TREC's existence was responsible for approximately one-third of an improvement of more than 200% in web search products that was observed between 1999 and 2009.[1]

[1] In addition, RTI conducted further analysis of the survey data collected during this study to supplement the benefit-cost analysis results described in Section 6 of this report. Appendix C summarizes the results of this supplementary analysis, which focused on better understanding how the level of use of TREC resources and the value generated by using these resources differed by location and organizational type.

The quantified retrospective benefits of TREC, however, do not necessarily imply that the benefits of continuing TREC in the future would be of similar magnitude. In this concluding section, cautious estimates of the future benefits of TREC are provided, based on research, survey data collected, and interviews conducted as part of this study, and several suggestions for the future of TREC are ventured.

7.1 Implications for NIST and the Future Outlook on TREC

To investigate the potential future benefits of the TREC Program, interview and survey participants were asked to estimate the future benefits that they expect to receive from TREC over the next year (mid-2010 to mid-2011). Similar to how respondents were asked to estimate the benefits that they have received from TREC in the past using a contingent valuation approach, respondents were asked to estimate how much they would be willing to pay for "TREC-like services" over the next year if TREC were cancelled today. Box 1 provides the exact question wording that was used in the survey.

Box 1

Estimating the Benefits of Future TREC Resources Using Contingent Valuation

During the contingent valuation portion of the survey, respondents are asked the following question:

> "TREC was designed to foster IR research by offering a variety of resources (such as test collections and workshops) that support IR system evaluation. Suppose that NIST was considering discontinuing TREC next year. However, a different organization offered to perform the exact same services and the determination of research areas for specific tracks would be determined by an external panel of academic and private IR researchers.
>
> Would your organization be interested in purchasing these services?
> <Yes/No>
>
> Note to Academics: Assume that you would be able to include the cost of purchasing these TREC-like services in a grant as you would with other capital expenses.
>
> Would the value of the services offered by the TREC-like program be worth at least $27,000 to your research over the next year?
>
> <Yes/No>"

If respondents answer "yes," then the question is represented with a higher value and the process is repeated until the respondent answers "no." Alternatively, if the respondent answers "no," then the question is repeated with a lower value until the respondent answers "yes." This means of eliciting valuations from respondents is known as a bidding game (Freeman, 2003).

Respondents had to make certain assumptions about what TREC tracks would occur next year, who would attend the TREC workshop next year (and if they planned to attend), and the value of the papers that are produced at the TREC workshop in the fall of 2010. Such uncertainly means that the estimated benefits could be either higher or lower than those that will actually accrue to stakeholders. Given these caveats, this study estimates that the unextrapolated benefits from TREC over the next year will be approximately $5.2 million and the extrapolated benefits will be $7 million.[2] Assuming the TREC Program costs and utilization costs will be approximately the same as the costs in 2009 ($1.7 million), this results in a net societal benefit of between $3.5 and 5.4 million.

This net benefit estimate provides a compelling case for continuing TREC. As compared to our valuation estimates for previous years, our estimate of the value that will be observed in the next year (March 2010 through February 2011) is slightly below the average per-year benefit estimated from 1992 to 2009 (ranging from $5.5 million of unextrapolated benefits to $8.1 million of extrapolated benefits). This may be partially explained by the fact that a concentration of TREC benefits was generated in the late 1990s, likely resulting from the growth of personal computing and Internet usage.

As with our retrospective benefit estimates, this estimate of the benefits of TREC for 2010 do not include any end-user benefits or other spillover benefits outside standard IR research benefits (e.g., the research cost savings and the quality improvement benefits that private firms and academic institutions can capture). Further, this benefit estimate does not include valuation of the benefits to graduate students, who are able to learn from TREC resources that otherwise would not be available.

In addition to looking at the prospective benefits that stakeholders believe they will receive next year, IR researchers and experts also provided information on what they think TREC should do in the future. Several trends in survey responses emerged; 37 respondents indicated that TREC should expand into new tracks, 20 said TREC should develop new evaluation methods, and 17 said TREC should develop new data sets. Common suggestions were the following:

- Focus on more user behavior data (e.g., social data, Twitter, geographically based) to improve on the Interactive track.
- Continue to look at multimedia search techniques (e.g., pictures, video).
- Expand into more focused search areas (e.g., chemistry, drug design, evidence-based medicine).

More broadly, several respondents suggested that TREC should work with industry to increase their participation in the TREC workshops, as well as to solicit data that they might allow the TREC audience to use, thus increasing the usefulness of TREC results. One respondent

[2]Respondents' valuation estimates were extrapolated using the same extrapolation strategy discussed earlier in the report.

suggested that more time should be spent discussing the improvements in search techniques, instead of spending so much time talking about the methodologies used to compare system results in the TREC competition. Another respondent suggested that TREC should try to partner with a leading journal to expand the dissemination of TREC results and/or collaborate with another conference to increase participation.

TREC has proved to be a success, and our estimates suggest that TREC will be beneficial moving forward as long as TREC continues to focus on its key benefits—providing new data sets, developing methodologies to approach new IR topics, and providing a venue for IR researchers in all sectors to compare their IR techniques objectively and to share both successes and failures in a safe environment.

REFERENCES

Battelle, J. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Penguin Group.

Bourne, C. 1999. “40 Years of Database Distribution and Use: An Overview and Observation.” Available at <http://www.nfais.org/publications/mc_lecture_1999.htm>. As obtained on January 15, 2009.

Bourne, C., and T. Hahn. 2003. *A History of Online Information Services, 1963–1976*. Cambridge: MIT Press.

Buckley, C., A. Singhal, and M. Mitra. 1997. “Using Query Zoning and Correlation Within SMART: TREC 5.” In *Proceedings of the Fifth Text Retrieval Conference*, NIST Special Publication. Available at <http://singhal.info/trec5.pdf>. As obtained on January 15, 2009.

Buckley, C., and E. Voorhees. 2000. “Evaluating Evaluation Measure Stability.” *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 33-40. Athens, Greece.

Buckley, C. and E. Voorhees. 2004. “Retrieval Evaluation with Incomplete Information.” *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 25-32. Sheffield, UK.

Bureau of Economic Analysis (BEA). 2009. 2009 Input-Output (I-O) Accounts. Available at <http://www.bea.gov/industry/index.htm#annual>. As obtained on May 20, 2009.

Bureau of Labor Statistics (BLS). 2010. Current Employment Statistics Database. Available at <http://www.bls.gov/ces/>. As obtained on April 20, 2010.

Bush, V. 1945. “As We May Think.” *Atlantic Monthly* 176(1):101-108. Available at <http://www.theatlantic.com/doc/194507/bush>. As obtained on December 19, 2008.

Champ, P., K. Boyle, and T.C. Brown. 2003. *A Primer on Nonmarket Valuation*. Dordrecht: Kluwer Academic Publishers.

Chowdhury, G.G. 2004. *Introduction to Modern Information Retrieval 2e*. London: Facet Publishing.

Forum for Information Retrieval Evaluation (FIRE). FIRE Home page. Available at <http://www.isical.ac.in/~clia/>. As obtained on February 25, 2009.

Fox, E., and M. Koll. 1988. “Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems.” *Information Processing and Management* 24(3):257-267.

Freeman, A.M. 2003. *The Measurement of Environmental and Resource Values: Theory and Methods.* Washington, DC: Resources for the Future.]

Gartner. 2008. "Gartner Says Worldwide Enterprise Search Market Will Surpass $1.2 Billion in Total Software Revenue by 2010." Available at <http://www.gartner.com/it/page.jsp?id=596407>. As obtained on June 16, 2009.

Grisham, R., and B. Sundheim. 1996. "Message Understanding Conference 6: A Brief History." Available at <http://acl.ldc.upenn.edu/C/C96/C96-1079.pdf>. As obtained on January 15, 2009.

Grossman, D., and O. Frieder. 2004. *Information Retrieval, Algorithms and Heuristics*. Dordrecht, The Netherlands: Springer Publishing.

Harman, D., and G. Candela. 1990. "Retrieving Records from a Gigabyte of Text on a Minicomputer Using Statistical Ranking." *Journal of the American Society for Information Science* 41(8):581-589.

Hersh, W. 2003. *Information Retrieval.* New York: Springer Press.

Johnson, N. 2009. "Nielsen Online December 2008 Search Engine Share Rankings." Search Engine Watch. Available at <http://searchenginewatch.com/3632382>. As obtained on May 19, 2009.

Lancaster, F.W. 1979. *Information Retrieval Systems.* New York: John Wiley & Sons, Inc.

Levack, K. 2002. "Inxight Makes Unstructured Data Manageable with SmartDiscovery 3.0." E-Content Magazine. Available at <http://www.econtentmag.com/Articles/News/News-Feature/Inxight-Makes-Unstructured-Data-Manageable-with-SmartDiscovery-3.0-915.htm>. As obtained on May 19, 2009.

Manning, C.D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.

McGill, M., L. Smith, S. Davidson, and T. Noreault. 1976. "Syracuse Information Retrieval Experiment (SIRE): Design of an On-line Bibliographic Retrieval System." *ACM SIGIR Forum* 10(4):13-16.

Meadow, C., B. Boyce, and D. Kraft. 2000. *Text Information Retrieval Systems*. San Diego: Academic Press.

Milic-Frayling, N. 1999. "Information Retrieval from Unstructured Textual Data: Recent Developments and New Challenges." Available at <www-ai.ijs.si/MarkoGrobelnik/awamida99/milic.doc>. As obtained on April 28, 2009.

Moulton, L. 2008. "Enterprise Search Markets and Applications." Available at <http://gilbane.com/Research-Reports/Gilbane-Enterprise-Search-Report-June-08.pdf>. As obtained on April 19, 2010.

National Science Foundation (NSF). 2007. "Expenditures for U.S. Industrial R&D Continue to Increase in 2005; R&D Performance Geographically Concentrated." Available at <http://www.nsf.gov/statistics/infbrief/nsf07335/>. As obtained on April 28, 2009.

National Science Foundation (NSF). 2010. Academic R&D Expenditures: FY2008. Available at <http://www.nsf.gov/statistics/nsf10311/>. As obtained on April 19, 2010.

Online Publishers Association (OPA). 2010. Internet Activity Index. Available at <http://www.online-publishers.org/internet-activity-index>. As obtained on April 22, 2010.

Robertson, S. 2008. "On the History of Evaluation in IR." *Journal of Information Science* 34(4):439-456.

Salton, G. 1987. "Historical Note: The Past Thirty Years in Information Retrieval." *Journal of the American Society for Information Science* 38(5):375-380.

Schatz, B. 1997. "Information Retrieval in Digital Libraries: Bringing Search to the Net." *Science* 275:327-334.

Search Engine Watch. 2009. Top Search Providers: August 2009. Available at <http://searchenginewatch.com/3634991>. As obtained on April 19, 2010.

Tassey, G. 2003. *Methods for Assessing the Economic Impact of Government R&D*. NIST Planning Report 03-01. Gaithersburg, MD: National Institute of Standards and Technology.

U.S. Department of Commerce, Bureau of the Census. 2008. Service Annual Survey. Available at <http://www.census.gov/svsd/www/services/sas/sas_data/historical/sas-07.pdf>.

Varian, H. 2007. "Why Data Matters." Available at <http://googleblog.blogspot.com/2008/03/why-data-matters.html>. As obtained on April 14, 2010.

Varian, H. 2010. Personal communication with RTI International.

Voorhees, E. 2000. "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness." *Information Processing and Management* 36:697-716.

Voorhees, E. 2001. "Evaluation by Highly Relevant Documents." *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 74-82. New Orleans, LA.

Voorhees, E. 2008. Personal communication with RTI International. Data on TREC Participation.

Voorhees, E., and D. Harman. 2005. *Experiment and Evaluation in Information Retrieval*. Cambridge: MIT Press.

Voorhees, E. and D. Tice. 2000. Building a Question-Answering Test Collection. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 200-207.

White, M. 2007. *Making Search Work: Implementing web, Intranet and Enterprise Search*. London: Facet Publishing.

Wildstrom, S. 2005. "Desktop Search: The Game Is Afoot." *Business Week*. Available at <http://www.businessweek.com/magazine/content/05_02/b3915019_mz006.htm>. As obtained on May 20, 2009.

Wilson, T.D. 1981. "On User Studies and Information Needs." Journal of Documentation 37(1):3-15. Available at < http://informationr.net/tdw/publ/papers/1981infoneeds.html>. As obtained on June 16, 2009.

APPENDIX A: NIST/RTI TREC SURVEY

Economic Impact of the Text REtrieval Conferences (TREC)

A Survey Sponsored by the National Institute of Standards and Technology (NIST)

Introduction

The National Institute for Standards & Technology (NIST) has commissioned RTI International, a not-for-profit research institute, to conduct a retrospective impact assessment of the Text REtrieval Conference (TREC) Program and its sister program the TREC Video Retrieval Evaluation (TRECVid), which became an independent evaluation program in 2003.

The following survey is an important effort that seeks to provide insights into the economic and research contributions of these activities. It requests:

- information on your use of TREC resources and
- your estimates of the value of TREC resources to your organization's research.

Note that throughout the survey, we will refer to TREC and TRECVid collectively as TREC.

How long will this take?

Approximately 15 to 20 minutes

Why should I participate?

(1) Most importantly, to help TREC with future planning.
(2) For a chance to win an **Apple iPod Touch** (retail value of $400)!*

U.S. PAPERWORK REDUCTION ACT NOTICE: This questionnaire contains collection of information requirements subject to the Paperwork Reduction Act (PRA). Notwithstanding any other provisions of the law, no person is required to respond to, nor shall any person be subject to penalty for failure to comply with, a collection of information subject to the requirements of the PRA, unless that collection of information displays a currently valid OMB Control Number. The estimated response time for this questionnaire is 20 minutes. The response time includes the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this estimate or any other aspects of this collection of information, including suggestions for reducing the length of this questionnaire, to the National Institute of Standards and Technology, Attn., Gregory Tassey, gregory.tassey@nist.gov, (301) 975-2663. The OMB Control No. is 0693-0033, which expires on 10/31/2012.

CONFIDENTIALITY: Responses to this survey will be kept strictly confidential. At no time will any individual's name, any company or university name, their participation, or identifiable response be released by RTI to any third party, including NIST. The data that survey respondents provide will only be used to present aggregate analytical findings to NIST in the form of a final report that will be publicly released by the end of this calendar year following a peer review process.

QUESTIONS: If you have any questions about this survey, please contact Dallas Wood at (919) 541-8743 or dwood@rti.org [US Eastern Time], or Brent Rowe at (415) 848-1317 or browe@rti.org [US Pacific Time].

* Participants in this survey will be entered into a raffle drawing for an Apple iTouch (retail value $400). Only participants who <u>complete the survey</u> will be entered into the drawing. Please provide contact information in the final section so that we may contact you if you are selected.

1. Basic Information on the Use of TREC Products and Services

The following questions ask about how TREC has benefited your research.

1a Since 1992, TREC has hosted annual workshops, developed test collections, developed evaluation methods, and published research papers in order to foster information retrieval (IR) research. Please indicate how important each TREC product or service is for your IR research.

	Product or Service Not Used	Product or Service Used and Not Very Important for IR Research	Product or Service Used and Somewhat Important for IR Research	Product or Service Used and Very Important for IR Research
Workshops	☐	☐	☐	☐
Test collection(s)	☐	☐	☐	☐
Evaluation method(s)*	☐	☐	☐	☐
Research papers	☐	☐	☐	☐

* Evaluation methods refer to the performance metrics (such as Mean Average Precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

<<If respondent answers NOT USED for each aspect of TREC, direct them to goodbye screen. >>

Comments:

1b TREC has pursued a large number of tracks aimed at fostering research in specific IR areas. Typically, each track is associated with its own set of resources (workshops, test collections, evaluation methods, or research papers). In the table below, please **indicate how important these resources have been for your IR research.**

The years each track was held is indicated in parentheses below.

Track	Product or Service Not Used	Product or Service Used and Not Very Important for IR Research	Product or Service Used and Somewhat Important for IR Research	Product or Service Used and Very Important for IR Research
Ad hoc (1992–1999)	☐	☐	☐	☐
Blog (2006–present)	☐	☐	☐	☐
Cross-language (1997–2002)	☐	☐	☐	☐
Enterprise (2005–2008)	☐	☐	☐	☐
Filtering/Routing (1992–2002)	☐	☐	☐	☐
Genome (2002–2007)	☐	☐	☐	☐
Interactive (1994–2002)	☐	☐	☐	☐
Legal (2006–present)	☐	☐	☐	☐
Question answering (1999–2007)	☐	☐	☐	☐
Spam (2005–2007)	☐	☐	☐	☐
Speech (1997–2000)	☐	☐	☐	☐
Video/TRECVid (2001–present)	☐	☐	☐	☐
Web (1999–2004, 2009)	☐	☐	☐	☐
Other (___________)	☐	☐	☐	☐

Comments:

2. Detailed Information on Personal Use of TREC Products and Services

The following questions ask about how often and in what ways you use different aspects of TREC.

2a In the table below, please identify the years in which you participated in a TREC conference or the year from which you used other TREC resources. For example, if you participated in a TREC conference in 1994, select 1994. If you have ever used a test collection that was used in a 2005 TREC track, then select 2005.

Participation in TREC Workshops and Use of TREC Resources/Output

	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08
Participated in TREC workshop	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Used additional TREC resources (test collection, evaluation methods*, or research papers)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

*Evaluation methods refer to the performance metrics (such as Mean Average Precision) and experimental designs developed through TREC to evaluate the performance of IR systems.

Comments:

2b **Use of TREC Test Collections**

The questions below pertain to your use of the test collections associated with TREC. In some cases, you may not be able to provide an exact response. In those instances, please answer by providing your best estimate.

#	Question	Answer
I.	Have you published any research that uses the test collections from TREC??	**<Yes/No>**
Ia.	If yes, how many papers have you published in peer-reviewed journals?	
Ib.	If yes, how many papers have you presented at conferences?	
II.	Have you used test collection data from TREC for tasks not studied at TREC?	**<Yes/No>**
IIa.	If yes, please elaborate on how you used these data?	

Comments:

2c **Use of TREC Evaluation Methods**

The question below pertains to your use of the evaluation methods associated with TREC.

#	Question	Answer
I.	Have you used TREC evaluation methods to study the performance of an IR system using non-TREC test collections?	**<Yes/No>**

Comments:

2d **Use of TREC Research Papers**

The questions below pertain to your use of research papers associated with TREC. In some cases, you may not be able to provide an exact response. In those instances, please answer by providing your best estimate.

#	Question	Answer
I.	Approximately how many papers have you read that were written and published as part of the TREC Proceedings? (NOTE: Approximately 50 papers per year were published in the first 4-5 years, and approximately 100 papers per year were published over the last 10 years.)?	**Select One** **1-9** **10-49** **50-99** **100-249** **500-1,000** **1,000+**
II.	Have you ever referenced a paper from TREC in a peer-reviewed journal article or a paper presented at conference?	**<Yes/No>**
III.	Have you ever referenced a paper from TREC in a patent filing?	**<Yes/No>**

Comments:

3. The Value of TREC Products and Services to Your Organization's IR Research Activities

The questions in this section are designed to understand how your organization values TREC.

In the question below, please check the box that best characterizes your current employer. If you have worked at multiple organizations that have benefited from TREC, please focus either on your most recent employer or a single one of your previous employers that you believe best represents your experiences with TREC resources.

☐ Information Retrieval Service or Software Company

☐ Other Software Company

☐ University or Academic Research Laboratory

☐ Government or Institutional Research Laboratory

☐ Other (please specify)

3a How many FTEs does your organization currently employ in IR research and/or development?
FTEs

3b The following questions ask about the IR research activities pursued by the individuals employed by your organization. Given that you may not be familiar with the research activities of all these individuals, what percentage of the total FTEs you indicated in Q3a do you believe you would be able to speak for in your responses?
%

3c Do these researchers use TREC resources when conducting their research? **<Yes/No>**

If yes, how many years have TREC resources been used by your organization?

In what year did your organization begin using these TREC resources?

3d What percentage of your organization's research activities are devoted to the following research areas? Total should equal 100%

Research Area	Percentage of Research Activity
Desktop Search	%
Enterprise Search	%
Database Search	%
Web Search	%
Other IR Research	%
Non-IR Research	%
Total	**100%**

Comments:

3e TREC was designed to foster IR research by offering a variety of resources (such as test collections and workshops) that support IR system evaluation. Suppose that NIST was considering discontinuing TREC next year. However, a different organization offered to perform the exact same services and the determination of research areas for specific tracks would be determined by an external panel of academic and private IR researchers.

Would your organization be interested in purchasing these services?
<Yes/No>

<<If respondent is academic, show the following: Note to Academics: Assume that you would be able to include the cost of purchasing these TREC-like services in a grant as you would with other capital expenses.>>

Would the value of the services offered by the TREC-like program be worth at least **$XXX** to your research over the next year? **<Yes/No>**

<<The figure, $XXX, above is calculated differently for non-profit/academic/government organizations and for-profit organizations. Non-profit organizations will see an initial value that equals $2,000 (the cost of membership for Non-Profit Organizations in the LDC) plus 50% of the value of the FTEs they are responding for (calculated using responses above and BLS data). Private sector organizations will see an initial value that equals $20,000 (the cost of membership for Profit Organizations in the LDC) plus 50% of the value of the FTEs they are responding for (calculated using responses above and BLS data).

If the person answers that "YES" they value TREC this much, they will be presented with increasing values (in increments of 7, 1, 2, and 10x their FTEs) until they answer "NO". If they answer "NO", they will be presented with decreasing values (in increments of .25., .10, .01, and 0% of the value of their FTEs) until they answer "YES".>>

To better understand your response to this question, please elaborate on the factors you considered when estimating the value of TREC resources to your organization?

3f Over the past 17 years, TREC has provided a number of resources (test collections, evaluation methods, workshops, and research papers) to help foster IR research. Without the use of TREC resources, your organization might have

- spent **more time creating new data sets** (or using smaller or less robust data sets),
- spent **more time evaluating IR systems** and been less confident in evaluation results,
- spent **more time going down faulty paths,** and/or
- created **IR systems of lower quality.**

Earlier in this survey you indicated that researchers in your organization have used TREC resources for 0 years. Suppose that you were asked by a project manager or your organization's financial accountant to justify the use of TREC resources by specifying the value they contributed to your research.

Do you believe these resources contributed at least **$XXX** of value to your research? **<Yes/No>**

NOTE: We are not asking how much your organization would have or could have paid, but rather, what you should have paid, looking back. Consider the full impact that TREC has had on your research (including all the applicable benefits listed above)?

<< The figure above is calculated differently for non-profit and for-profit organizations. Non-profit organizations will see an initial value that equals $2,000 (the cost of membership for Non-Profit Organizations in the LDC) plus 50% of the value of the FTEs they are responding for (calculated using responses above and BLS data), multiplied by the number of years they have been utilizing TREC resources. Private sector organizations will see an initial value that equals $20,000 (the cost of membership for Profit Organizations in the LDC) plus 50% of the value of the FTEs they are responding for (calculated using responses above and BLS data), multiplied by the number of years they have been utilizing TREC resources.

If the person answers that "YES" they value TREC this much, they will be presented with increasing values (in increments of 7, 1, 2, and 10x their FTEs) until they answer "NO". If they answer "NO", they will be presented with decreasing values (in increments of .25, .10, .01, and 0% of the value of their FTEs over the number of years the resources were being used) until they answer "YES".>>

To better understand your responses, please elaborate on the factors you considered in making this estimate?

4. The Impact of TREC on IR System Quality

The purpose of the questions in this section is to determine how TREC has influenced the quality of IR systems being developed.

4a. First, let's focus on IR systems developed within your research group. Did the quality of the IR system(s) developed by your research group improve during the time your organization used TREC products and services? **<Yes/No>**

If yes, by how much did the quality of your IR system improve, on average, as measured by Mean Average Precision (MAP)? You may not be able to provide an exact response. However, please answer by providing your best estimate. % improvement in IR system performance

How much of this change would you attribute to TREC? % of the total change

Comments:

4b In addition, to understand how TREC has impacted the quality of IR systems more broadly, the following questions will ask you about the impact that TREC has had on the quality of web search (as measured by the time it takes to fill an information need) over the past decade. **<Yes/No>**

Do you believe that the time it takes to fill an information need using web search engines has decreased since 1999, holding improvements in hardware and non-search software constant? For example, if you used a web search engine from 2009 to search the Internet of 1999 using a computer made in 1999, do you think you could fill an information need in less time than if you had used a search engine from 1999?

	Select One	**%**
If yes, what percentage change (faster speed) do you think you would have observed, on average, using today's search engines on a 1999 computer looking at the 1999 web? (NOTE: Although you might find this question difficult to calculate, we are looking for your best estimate based on your expert knowledge of the field of IR.)	**0**	
	1-4	
	5-9	
	10-14	
	15-19	
	20-29	
	30-39	
	40-49	
	50-59	
	60-69	
	70-79	
	80-89	
	90-99	
	100-199	
	200-299	
	300-399	
	400-499	
	500-99	
	1000+	

	Select One	**%**
What portion of this time savings would you attribute to improvements in web search quality enabled by TREC's workshops and resources?	**0**	
	1-4	
	5-9	
	10-14	
	15-19	
	20-29	
	30-39	
	40-49	
	50-59	
	60-69	
	70-79	
	80-89	
	90-99	

Comments:

5. The Impact of TREC on Academic Research Activity (for Academic Respondents only)

As an academic, you are in the best position to determine whether TREC has had an impact on the education and job prospects of individuals pursuing advanced degrees engaged in IR research. The following questions are designed to help NIST better understand the impact of TREC on these individuals.

5a	Do you believe the number of individuals pursuing doctoral and Masters degrees has increased as a result of TREC?	**<Yes/No >**
5b	How many doctoral candidates in your department have written dissertations/theses that used TREC resources (such as test collections, research papers, etc.)?	

5c How many Masters students in your department have earned degrees that depended on theses or projects using TREC resources (such as test collections, research papers, etc.)?

5d Do you believe that participation in TREC workshops improves the employment prospects of graduate students in your department? **<Yes/No>**

5e

Please **rank** the following organizations in order of where your graduate students typically find employment after graduation. Mark the organization that employed the most of your students (=1) and so on to the organization that employed the fewest (=5).**Type of Organization**	**Ranking**
Information Retrieval Service/Software Company	**--**
Other Software Company	**--**
University or Academic Research Laboratory	**--**
Government or Institutional Research Laboratory	**--**
Other (please specify):	**--**

Comments:

6. Thank you for participating. Finally, please fill out some information about yourself and your employer.

This survey is confidential, and at no time will any individual's name, any organization's name, participation, or identifiable response be released by RTI to any third party, including the National Institute for Standards & Technology (NIST). This information will only be used to characterize the size of organizations benefiting from TREC's activities. All data will be held strictly confidential and only reported when aggregated with data provided by all respondents and reported in that form in a final report to NIST.

Please provide the following information about yourself.

6a Please provide the following information about yourself:

Name: ____________________

Title: ____________________

Employer: ____________________

E-mail: ____________________

Geographic location: **<Inside US/Outside US>**

Only participants who include their contact information above will be entered into the drawing to win an Apple iPod Touch (retail value of $400)!

6b <<For respondents that indicated they were in the private sector>> Please provide any of the following information that you are able to:

Stock trading symbol (ticker), if publicly traded:	____________	
2008 approx. company annual sales (if no ticker):	**Select One** **0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-99** **100-499** **500-999** **1,000+**	Million US$
2008 approx. company total annual R&D budget (if no ticker):	**Select One** **0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	Million US$
2008 approx. annual R&D budget for IR	**Select One** **0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	Million US$

2008 approx. company employment	**Select One** **<100** **100-249** **250-499** **500-999** **1,000-4,999** **5,000-9,999** **10,000-49,999** **50,000-99,999** **100,000+**	FTEs
2008 approx. number of IR researchers	**Select One** **1-3** **4-6** **7-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	FTEs

6b <<For respondents that indicated they were in academia>> Please provide any of the following information that you are able to:

2008 approximate funding for your department	**Select One** **1-3** **4-6** **7-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	Million US$

Percentage of department funding devoted to IR research	**Select One** **0-0.9** **1-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	%
2008 approximate number of full-time IR researchers/professors	**Select One** **1-3** **4-6** **7-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	FTEs
2008 approximate number of student IR researchers	**Select One** **1-3** **4-6** **7-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	FTEs

6b <<For respondents that indicated they operated in a government laboratory >>
Please provide any of the following information that you are able to:

	Select One	
2008 approximate research funding for your organization	**0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-99** **100-499** **500-999** **1,000+**	Million US$
Percentage of research funding devoted to IR research	**0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100**	%
Portion of IR Research Funding from Government Sources	**0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100**	%

	Portion of IR research funding from Non-Government Sources	**<u>Select One</u>** **0-0.9** **1-4** **5-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100**	%
	2008 approx. organization employment	**<u>Select One</u>** **<100** **100-249** **250-499** **500-999** **1,000-4,999** **5,000-9,999** **10,000-49,999** **50,000-99,999** **100,000+**	FTEs
	2008 approx. number of IR researchers	**<u>Select One</u>** **1-3** **4-6** **7-9** **10-19** **20-29** **30-39** **40-49** **50-59** **60-69** **70-79** **80-89** **90-99** **100+**	FTEs
6c	How can TREC best provide benefits to the IR research community and IR-relevant industry in the future? (e.g., new tracks, new evaluation procedures, etc.)		
6d	Would you be willing to participate in a 15-minute interview to discuss your responses to this survey?	**<Yes/No>**	

APPENDIX B: ORGANIZATIONS INCLUDED IN EXTRAPOLATION BASE

Table B-1. Companies Included in Extrapolation Base

Company
Amazon
AOL
Apple
Ask
AT&T Bell Laboratories
Autonomy
Boeing Company
Cirrus Logic
CL Research
Clearwell Systems Inc.
Cleary Gottlieb Steen & Hamilton
Google
Harris Corporation
Hewlett-Packard
IBM
Language Computer Corporation
LCC International
Lucid Imagination
Microsoft
Omniture
Oracle Corp.
Progress Software Corp.
SABIR Research
Sapient Corporation
Sun Microsystems
Texas Instruments
The Echo Nest
Thomson Reuters Corporation
VIStology, Inc.
Xerox
Yahoo

Table B-2. Academic Institutions Included in Extrapolation Base

Institution
Carnegie Mellon University
Arizona State University
Brown University
Case Western Reserve University
College University of New York, Queens College
Columbia University
Cornell University, all campuses
Dartmouth College
DePaul University
Drexel University
Emory University
Florida Atlantic University
Florida Institute of Technology
Florida International University
George Mason University
Georgia Institute of Technology all
Harvard University
Illinois Institute of Technology
Indiana University, all campuses
Iowa State University
Johns Hopkins University, The
Kettering University
Lehigh University
Louisiana State University, all campuses
Massachusetts Institute of Technology
Michigan State University
Millersville University, Pennsylvania
Milwaukee School of Engineering
New York University
North Dakota State University, all campuses
Northeastern University
Ohio State University all campuses
Oregon Health & Science University
Pennsylvania State University all campuses
Princeton University
Purdue University, all campuses
Rice University
Rutgers, The State University of New Jersey

(continued)

Table B-2. Academic Institutions Included in Extrapolation Base (continued)

Institution
Smith College
Southern Methodist University
Stanford University
State University of New York, Binghamton
SUNY Buffalo all campuses
Syracuse University, all campuses
Tufts University
U. VA all campuses
University of Illinois, Urbana-Champaign
University Michigan all campuses
University of California, Berkeley
University of California, Irvine
University of California, Los Angeles
University of California, San Diego
University of California, Santa Barbara
University of Colorado, all campuses
University of Delaware
University of Illinois, Chicago
University of Iowa
University of Louisville
University of Maryland, Baltimore County
University of Maryland, College Park
University of Massachusetts, Amherst
University of Miami
University of Minnesota, all campuses
University of North Carolina, Chapel Hill
University of North Carolina, Charlotte
University of North Texas
University of Pennsylvania
University of Pittsburgh, all campuses
University of Texas at Dallas
University of Texas, Austin
University of Utah
University of Washington
University of Wisconsin, Madison
University Texas, San Antonio
Utah State University
Virginia Polytechnic Institute
Yale University

APPENDIX C: ADDITIONAL ANALYSIS OF TREC BENEFITS DISTRIBUTION

C.1 Introduction

This appendix describes two types of additional analyses that were conducted to further explain the distribution of benefits stemming from NIST's TREC Program. First, RTI used the data set created during the TREC economic impact study to more fully analyze quantitative metrics of both TREC participation and the benefits generated from TREC, focusing on how these metrics differed by location and organizational type. Second, RTI used information retrieval (IR) techniques to analyze the flow of benefits among various TREC participants and other IR researchers, including specifically participants of TREC as well as their collaborators, some of who may or may have not participated in TREC. In particular, for the information flow analysis, we examined the relationships between researchers from both academia and industry.

The results of this analysis provide insights that NIST could use for future strategic planning purposes or specific prospective investment decision making.

C.2 Statistical Analysis of Survey Data

During RTI's study of the value of NIST's investment in the TREC Program, a large data set was developed that included information from IR researchers around the world. The data therein were used to conduct a benefit-cost analysis of the TREC Program. However, the breadth and depth of the data set offered the opportunity to uncover more information about the relationship between NIST and "consumers" of a specific program, in this case the Text REtrieval Conference. This appendix summarizes the results of additional analysis of the TREC survey data that RTI conducted, specifically aiming to address two key research questions:

1. Has the level of *use* of TREC resources varied by geographical location and by organizational type?
2. Has the *value* generated by the use of TREC resources varied by location and by organizational type?

C.2.1 Data Set Used for the Statistical Analysis

Data used in this analysis were collected through the 2010 TREC Web Survey,[1] which was distributed to IR researchers employed in four types of organizations:

- IR service and software companies,
- other software companies,
- universities and academic research laboratories, and
- government and institutional research laboratories.

This survey first asked respondents to provide information on how they have used TREC resources in their *individual research* (e.g., how often they have attended TREC workshops and

[1]A full copy of the survey instrument is provided in Appendix A.

used other TREC resources) and what impacts these resources have had on their research. Next, respondents were asked how TREC resources have been used in their *organization's research* (e.g., how long organizations have used TREC resources and when they first started using these resources) and how much value these resources have generated for their organizations. Throughout this statistical addendum, we use data on both TREC's impact on the individual respondent's research and TREC's impact on research pursued by the organization employing the respondent.

At the completion of data collection, RTI received 404 responses, of which 246 reported both the location in which their organization resides and its organizational type. Table 1 provides a distribution of these 246 respondents by organization location and type. A total of 93 organizations are from the United States and 153 are from other countries. As these data show, the majority of respondents were either employed by IR service and software companies or by universities or academic research laboratories (over 60% of the respondents in the sample of 246 are from universities or academic research laboratories). Because the number of respondents working for other software companies, government or institutional research laboratories, and other organizations is small, the cross-organizational comparisons presented throughout this addendum will focus on the differences between IR service and software companies and universities and academic laboratories. The main reason is that the mean values provided by these other organization types would not be meaningful.

Over 60% of the 246 respondents comprising our sample are located outside the United States. Although respondents working for non-U.S. organizations were not asked to identify the exact location of their organization, the Internet protocol (IP) address of the computer used by non-U.S.–based respondents was recorded and used to determine the geographic location of that respondent's computer. This information was included in this analysis as a proxy for the location of the respondent's organization to see if the use of TREC resources and the value they generate vary outside the United States by geographic location. Although data were collected from non-U.S. respondents in Africa, Asia, Europe, the Middle East, North America, Oceania, and South America, over 80% of non-U.S. respondents were located in either Europe or Asia. Because the number of respondents in remaining areas was relatively small, it was decided that discussing information at the organizational level for these areas would not lead to useful results regarding the average use and value of TREC resources for organizations in these areas and could potentially lead to revealing sensitive information about individual respondents. Therefore, a disaggregation of non-U.S. respondents is only provided in our analytical results for Europe and Asia. Table C-1 provides the distribution of non-U.S. respondents located in these two areas by organizational type.

Table C-1. Distribution of Respondents by their Organization's Location and Type

Organization Location	Organization Type	Number of Respondents
U.S.		
	IR service or software company	18
	Other software company	4
	University or academic research laboratory	61
	Government or institutional research laboratory	4
	Other	6
	All U.S.	93
Non-U.S.		
	IR service or software company	10
	Other software company	3
	University or academic research laboratory	113
	Government or institutional research laboratory	22
	Other	5
	All Non-U.S.	153
Asia		
	IR service or software company	1
	University or academic research laboratory	29
	Government or institutional research laboratory	4
	All Asia	34
Europe		
	IR service or software company	7
	Other software company	1
	University or academic research laboratory	65
	Government or institutional research laboratory	15
	Other	3
	All Europe	91

.

C.2.2 Use of TREC Resources

The use of TREC resources was measured in three ways:

- the number of years that *the respondent* participated in a TREC workshop (*WorkPart*);
- the number of years that *the respondent* used additional TREC resources (e.g., test collections, evaluation methods, research papers) (*ResUse*); and
- the number of years that researchers at *the respondent's organization* used TREC resources in pursuing IR research activities (*NumYears*).

A summary of the data collected for each measure is presented for all respondents in Table C-2. As these data show, respondents employed by U.S. organizations indicated they have individually used TREC resources for longer than respondents from non-U.S. organizations on average (*WorkPart*). For example, respondents employed by U.S. organizations reported attending TREC workshops for an average of 3.7 years while non-U.S. organizations reported attending TREC workshops for an average of 2.1 years (nearly half as long). Similarly, respondents employed by U.S. organizations reported using other TREC resources for an average of 7.8 years (*ResUse*), while respondents employed by non-U.S. organizations reported using these resources for only 5.4 years (over 30% less time).

However, it is important to note that although U.S. respondents report having used TREC resources and attended TREC workshops for longer periods of time than their non-U.S. counterparts, the intensity of utilization among U.S. and non-U.S. respondents has been relatively similar in recent years. To investigate the intensity of use over time, we developed an aggregated "utilization score" by combining the *WorkPart* and *ResUse* metrics as follows:

- If respondents indicated they participated in a TREC workshop in a specific year (e.g., 1992), their answer was recorded as a 1 for that year. If they indicated they did not participate that year, their answer was recorded as a 0.
- If respondents indicated they had used TREC resources in a specific year (e.g., 1992), their answer was recorded as a 1 for that year. If they indicated they did not use TREC resources that year, their answer was recorded as a 0.
- Combining responses to these questions at the individual level yields a sum ranging from 0 to 2 that represents an individual's utilization score for a given year.

To analyze differences among respondents by country—U.S. and non-U.S respondents—we then averaged these individual utilization scores across all respondents of each type for each year of TREC's history. These annual data provide us with a convenient metric for comparing the intensity of TREC resource utilization over time.

Table C-2. Measures of the Use of TREC Resources

		Mean				Std. Dev.				Range			
Variable	**Definition**	**U.S. (n)**	**Non-U.S. (n)**	**Europe (n)**	**Asia (n)**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**
WorkPart	Number of years that a representative from the organization participated in a TREC workshop (years)	3.77 (93)	2.12 (153)	1.88 (91)	1.97 (34)	4.99	2.85	2.54	2.18	0–17	0–14	0–12	0–7
ResUse	Number of years that a representative from the organization used additional TREC resources (e.g., test collections, evaluation methods, research papers) (years)	7.76 (93)	5.35 (153)	5.29 (91)	4.62 (34)	5.94	4.56	4.45	3.08	0–17	0–17	0–17	0–12
NumYears	Number of years that researchers at the respondent’s organization used TREC resources in pursuing IR research activities (years)	9.16 (73)	6.78 (134)	6.64 (80)	5.70 (31)	5.34	4.14	4.13	2.84	0–17	1–17	1–17	2–15

Figure C-1 presents the results of the calculation described above for U.S. and non-U.S. respondents for each year, from 1992 to 2008. As this graph shows, U.S. respondents used TREC resources much more heavily in the early years of TREC. However, since 2000, utilization of TREC resources among non-U.S. respondents has increased significantly, and in 2008 non-U.S. respondents used TREC resources more heavily, on average, than U.S. respondents for the first time.

In addition to the use of TREC resources by individual researchers, respondents employed by U.S. organizations also indicated that researchers at their organizations used TREC resources for longer on average (*NumYears*). Specifically, these respondents indicated that researchers at their organizations had used TREC resources for 9.2 years, while respondents employed by non-U.S. organizations indicated that researchers at their organizations had used TREC resources for 6.8 years on average.

To see how using TREC resources may differ by organization type, Tables C-3 and C-4 present the three use metrics by location for universities and academic research laboratories, the largest organization type of user in the data sample, and IR service and software companies. Similar to the data presented above, these data show that respondents employed by U.S. organizations typically used TREC resources for longer than their non-U.S. counterparts on average.

Figure C-1. TREC Resource Utilization: 1992 to 2008

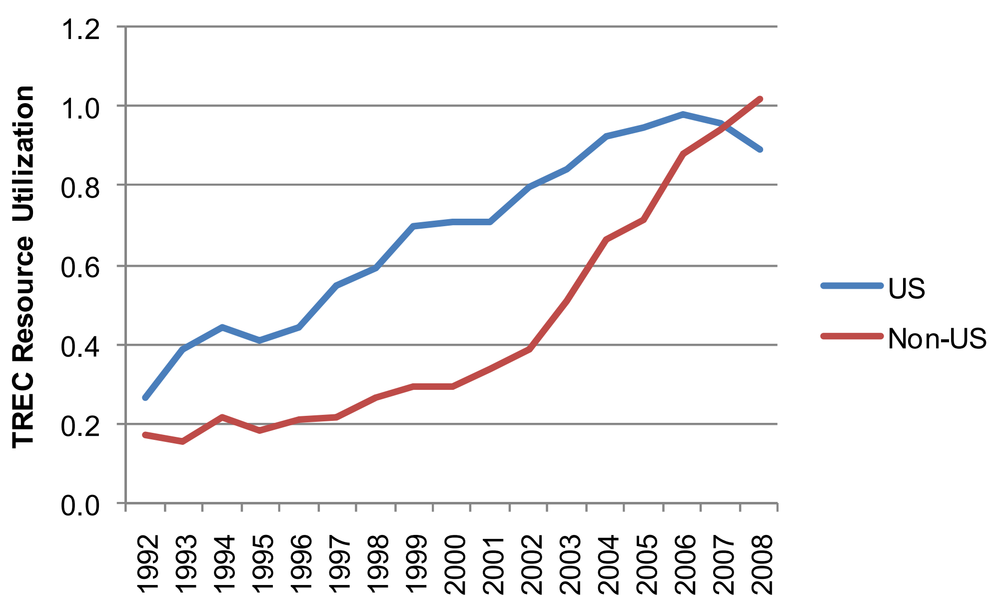


Table C-3. Measures of the Use of TREC Resources by IR Service and Software Companies

		Mean				Std. Dev.				Range			
Variable	**Definition**	**U.S. (n)**	**Non-U.S. (n)**	**Europe (n)**	**Asia (n)**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**
WorkPart	Number of years that a representative from the organization participated in a TREC workshop (years)	3.33 (18)	3.0 (10)	0.86 (7)	(D)	5.40	4.88	1.86	(D)	0–17	0–14	0–5	(D)
ResUse	Number of years that a representative from the organization used additional TREC resources (e.g., test collections, evaluation methods, research papers) (years)	8.17 (18)	6.0 (10)	4.14 (7)	(D)	6.12	5.93	3.72	(D)	0–17	0–17	0–10	(D)
NumYears	Number of years that researchers at the respondent's organization used TREC resources in pursuing IR research activities (years)	7.77 (13)	6.14 (7)	2.25 (4)	(D)	4.94	6.41	0.50	(D)	0–17	2–16	2–3	(D)

(D) signifies that so few individuals responded that data cannot be disclosed without revealing individual information.

Table C-4. Measures of the Use of TREC Resources by Universities and Academic Research Laboratories

		Mean				Std. Dev.				Range			
Variable	**Definition**	**U.S. (n)**	**Non-U.S. (n)**	**Europe (n)**	**Asia (n)**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**
WorkPart	Number of years that a representative from the organization participated in a TREC workshop (years)	4.18 (61)	1.83 (113)	1.71 (65)	1.86 (29)	5.24	2.45	2.19	2.18	0–17	0–14	0–8	0–7
ResUse	Number of years that a representative from the organization used additional TREC resources (e.g., test collections, evaluation methods, research papers) (years)	8.13 (61)	5.48 (113)	5.48 (65)	4.79 (29)	6.34	4.59	4.48	3.17	0–17	0–17	0–17	0–12
NumYears	Number of years that researchers at the respondent's organization used TREC resources in pursuing IR research activities (years)	9.98 (50)	6.70 (102)	7.00 (60)	5.42 (26)	5.32	4.06	4.26	2.35	1–17	1–17	0–17	2–10

However, when comparing usage of TREC resources across U.S. organizations, respondents working in universities and academic research laboratories reported using TREC resources for longer than respondents in IR service and software companies. For example, respondents working for U.S. universities and academic research laboratories indicated that they attended TREC workshops for 4.2 years on average, and respondents working for U.S. IR service and software companies indicated they had only attended TREC workshops for 3.3 years on average. Several potential reasons exist for why respondents in private firms may have used TREC resources less than those in academic settings. For example, researchers in private-sector firms may not have participated in TREC workshops as often as academic and nonprofit organizations because participation in TREC workshops could have resulted in disclosure of some proprietary tools and techniques.

C.2.3 Value Received from TREC Resources

To estimate the value that organizations and individual researchers received from using TREC, we used three metrics:

- the value respondents indicate their organizations would attribute to (would be willing to pay for) TREC resources over the next year (*FutValue),*
- the historical value respondents indicate their organizations would attribute to (the benefits they received from) past use of TREC resources (*HistValue*), and
- the importance of TREC resources to the respondent's individual research (*Impt*).[2]

A summary of each measure of value is provided in Table C-5 for all organizations. As these data show, both the historical and future values of TREC resources have been greater for U.S. organizations than for non-U.S. organizations. For example, respondents working for U.S organizations indicated that their organizations would be willing to pay an average of $90,000 for access to TREC resources next year (*FutValue*), while respondents working for non-U.S. organizations indicated that they would only be willing to pay an average of $20,000. Similarly, respondents working for U.S. organizations estimated the value TREC resources have generated over the history of their use to be $1.6 million on average (*HistValue*), while their non-U.S. counterparts estimated the historical value of TREC resources to be $0.29 million on average. However, data provided by respondents on the level of importance (*Impt*) suggest that researchers across all geographic areas believe that TREC resources were approximately equally important to their individual research.[3]

[2]This metric was calculated as the sum of responses to the importance of workshops, test collections, evaluation methods, and research papers (0 = not used, 1 = used but not very important, 2 = used and somewhat important, 3 = used and very important).
[3]No statistical significance was found between the values shown in Table C-5 for the *Impt* metric.

Table C-5. Measures of the Value of TREC Resources

		Mean				Std. Dev.				Range			
Variable	**Definition**	**U.S. (n)**	**Non-U.S. (n)**	**Europe (n)**	**Asia (n)**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**
FutValue	The amount a respondent's organization would be willing to pay for access to TREC resources in the coming year ($ millions)	$0.09 (75)	$0.02 (131)	$0.02 (77)	$0.04 (30)	$0.35	$0.05	$0.03	$0.07	$0–2.8	$0–0.3	$0–0.2	$0–0.3
HistValue	The amount of value a respondent's organization would attribute to the use of TREC resources since they began using those resources ($ millions)	$1.61 (74)	$0.29 (132)	$0.33 (78)	$0.23 (30)	$4.83	$1.01	$1.27	$0.45	$0–27.6	$0–9.9	$0–9.9	$0–2.1
Impt	Sum of responses to the importance of workshops, test collections, evaluation methods, and research papers (0 = not used, 1 = used but not very important, 2 = used and somewhat important, 3 = used and very important)	10.20 (91)	9.96 (148)	9.73 (88)	9.88 (32)	1.91	2.35	2.47	2.42	4.0–12	2.0–12	2.0–12	2.0–12

To illustrate how the value of TREC resources may differ by organization type, Tables C-6 and C-7 show the measured value realized by universities and academic research laboratories and by IR service and software companies, by location respectively. For all three measures of value, U.S. organizations of both types report, on average, marginally greater value than non-U.S. organizations. The most pronounced difference observed is that respondents working for U.S.-based IR service and software companies estimated the historical value generated by TREC resources to be significantly higher than respondents working for U.S. universities and academic research laboratories. For example, respondents employed by U.S. IR service and software companies estimated the historical value of TREC resources to be an average $6.5 million, while respondents in U.S. universities and academic research labs estimated the historical value of TREC resources to be $0.4 million. This large difference may be driven by several very high estimates for the historical value of TREC resources. That is, there was a wide range for the estimates provided by individuals employed by U.S.-based IR service and software companies. Still, a difference in the value estimated by private versus academic organizations would make intuitive sense. Private companies are much more interested in seeking a monetary return on their investment than academic organizations, which measure return on investment primarily using non-monetary metrics, such as increased scientific stature bestowed on the organization.

C.2.4 Concluding Remarks

RTI's additional analysis of the TREC survey data resulted in several key findings:[4]

- U.S. organizations have typically used TREC resources (workshops as well as test collections, evaluation methods, and research papers) more than non-U.S. organizations.
- Universities and academic research laboratories have used TREC resources for a longer time period, on average, than IR service and software companies (particularly in the United States).
- U.S. organizations have received more benefit from TREC (in terms of stated historical valuation of benefits) than non-U.S. organizations have received.
- IR service and software companies have received more benefit from TREC (historical valuation) than academic organizations have received.
- Looking forward, U.S. organizations place a higher value on (are willing to pay more for) the use of TREC resources in the coming year than do non-U.S. organizations.

[4]Each of the comparative findings presented in this report were also verified statistically (i.e., through statistical tests of differences between means).

Table C-6. Measures of the Value of TREC Resources by IR Service and Software Companies

		Mean				Std. Dev.				Range			
Variable	**Definition**	**U.S. (n)**	**Non-U.S. (n)**	**Europe (n)**	**Asia (n)**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**	**U.S.**	**Non-U.S.**	**Europe**	**Asia**
FutValue	The amount a respondent's organization would be willing to pay for access to TREC resources in the coming year ($ millions)	$0.29 (14)	$0.05 (6)	$0.05 (3)	(D)	$0.75	$0.06	$0.09	(D)	$0–2.7	$0–0.2	$0–0.2	(D)
HistValue	The amount of value a respondent's organization would attribute to the use of TREC resources since they began using those resources ($ millions)	$6.54 (13)	$0.32 (7)	$0.12 (4)	(D)	$10.19	$0.51	$0.22	(D)	$0.1–27.6	$0–1.4	$0–0.5	(D)
Impt	Sum of responses to the importance of workshops, test collections, evaluation methods, and research papers (0 = not used, 1 = used but not very important, 2 = used and somewhat important, 3 = used and very important)	10.12 (17)	9.0 (10)	8.14 (7)	(D)	2.26	2.62	2.54	(D)	5.0–12	5.0–12	5.0–12	(D)

(D) signifies that so few individuals responded that data cannot be disclosed without revealing individual information.

Table C-7. Measures of the Value of TREC Resources by Universities and Academic Research Laboratories

		Mean				Std. Dev.				Range			
Variable	Definition	U.S. (n)	Non-U.S. (n)	Europe (n)	Asia (n)	U.S.	Non-U.S.	Europe	Asia	U.S.	Non-U.S.	Europe	Asia
FutValue	The amount a respondent's organization would be willing to pay for access to TREC resources in the coming year ($ millions)	$0.03 (51)	$0.02 (100)	$0.01 (58)	$0.04 (25)	$0.11	$0.05	$0.02	$0.08	$0–0.7	$0–0.3	$0–0.2	$0–0.3
HistValue	The amount of value a respondent's organization would attribute to the use of TREC resources since they began using those resources ($ millions)	$0.41 (52)	$0.33 (100)	$0.39 (58)	$0.26 (25)	$0.75	$1.14	$1.46	$0.49	$0–4.4	$0–9.9	$0–9.9	$0–2.1
Impt	Sum of responses to the importance of workshops, test collections, evaluation methods, and research papers (0 = not used, 1 = used but not very important, 2 = used and somewhat important, 3 = used and very important)	10.23 (61)	10.06 (109)	9.83 (63)	10.04 (27)	1.93	2.27	2.52	2.05	4.0–12	2.0–12	2.0–12	3.0–12

These findings suggest that NIST's investment in TREC had the impact desired by all NIST investments—to support U.S. industries and the broader R&D community by providing a more robust technology infrastructure. TREC offered a significant improvement in the technology infrastructure for IR research, as detailed in the main NIST TREC report, and U.S. private-sector companies took the most advantage of the opportunities offered and the information generated. Future NIST studies should aim to further investigate the relationship between NIST investments and the processes by which benefits accrue to specific types of organizations and individuals.

C.3 IR Analysis of Flow of Benefits among TREC Participants and Collaborators

In addition to the benefit-cost analysis that we performed, we were also interested in two additional research questions:

- How did information presented at TREC flow among participants of TREC?
- How did information presented at TREC flow among collaborators of those researchers who participated in TREC?

To answer the first question, we performed a bibliographic analysis of TREC papers cited by other TREC papers. To further add to the data to answer the first question and to fully address the second question, we went beyond the TREC proceedings and searched the literature to examine links between TREC researchers and their collaborators. In particular, we searched for examples of collaborations between individuals from academic organizations and private-sector companies.

Our analysis used a variety of tools and analysis steps. First, we converted all the TREC proceedings into text format (from Postscript, image PDF, and searchable PDF files) unless they were already in text format and performed text mining with custom software code to extract basic statistics, such as numbers of authors, numbers of papers/author, and affiliations. We explored the use of the IBM AlphaWorks LanguageWare text analytics tools[38] to extract patterns and validate the data sets obtained with the custom software code. Additionally, to address the second question, we used Microsoft Research Asia's free Academic Search engine[39], which allowed us to validate and further investigate coauthor relationships and visualize the results. We used a combination of all these tools to examine a large number of cases where information was assumed to flow between authors from academic and commercial IR entities and vice versa through an analysis of the types of papers that those authors wrote together. This examination was not exhaustive, because that would have required developing a more complete set of metrics, which is a task that lies beyond the scope of this add-on project. We instead focused on sample relationships that described likely scenarios. Whether those scenarios are general remains an open question. But their mere existence is interesting.

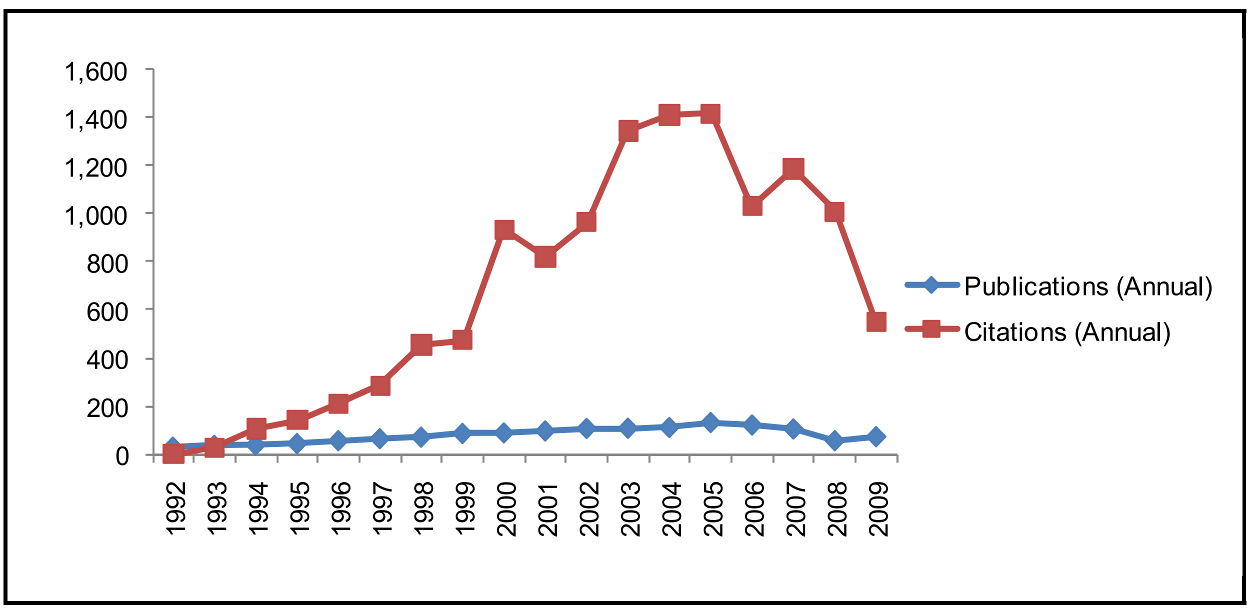
[38] See information at http://www.alphaworks.ibm.com/tech/lrw/download.
[39] See information at http://research.microsoft.com/en-us/projects/academic.

One way to measure the impact of a published work is to count the number of citations of that publication. However, we chose to use a more powerful measure, namely the identification of coauthorships between researchers presenting in TREC and their collaborators. Some of those collaborators might have also contributed to papers alongside other researchers, some of whom might have presented at TREC or not. That way we can trace how information from one author could have transferred from person to person and from type of entity to another. Specifically we were interested in learning how information presented by TREC authors in academic institutions made its way, through coauthorship linkages, to investigators in commercial IR companies, and vice versa.

C.3.1 Data Set Used for IR Analysis

The data used in this analysis included the 1,366 publications from 2,118 authors contained in the entire set of proceedings from TREC 1992 up to TREC 2009, as well as a corpus of 7,228,685 publications from 155,505 authors.[40] TREC authors were cited 16,024 times in the literature. The latter data set included all academic areas, not only IR researchers, but we wanted to include as many possible ways of identifying linkages between authors who presented at TREC and their collaborators. Figures C-2 and C-3 show the annual and cumulative number of TREC publications and citations to TREC publications.[41]

Figure C-2. Annual Number of TREC Publications and Citations

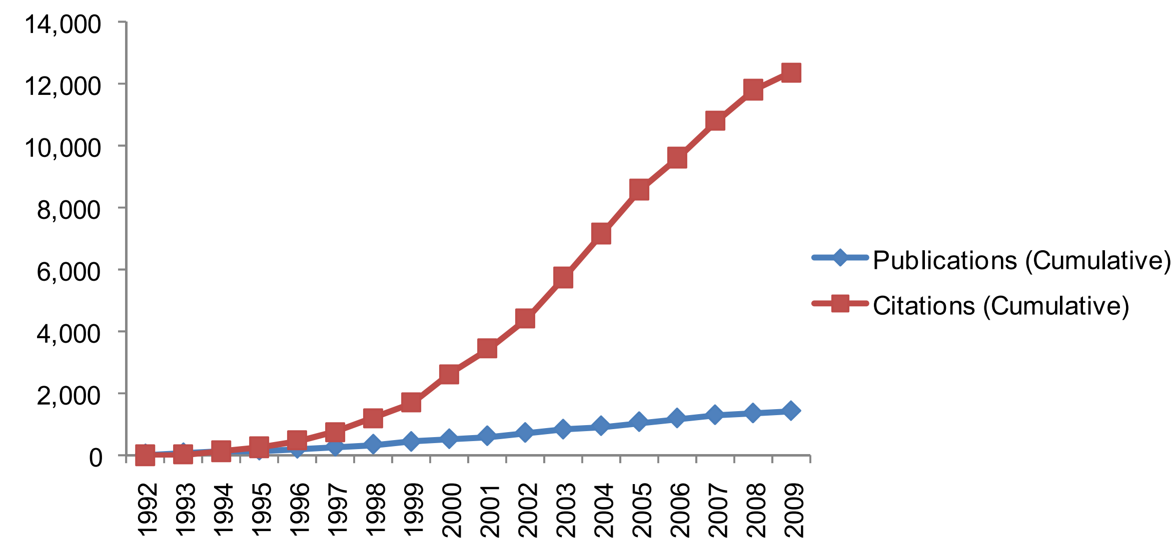


Source: Data from TREC Program Office and http://academic.research.microsoft.com/Conference/422.aspx.

[40] http://academic.research.microsoft.com/

[41] Of note, TREC paper account for the third highest number of publications and citations behind SIGIR— which had 2,751 publications and 45,982 citations between 1971 and 2010—and the International Conference on Information and Knowledge Management (CIKM)—which had 2,196 publications and 15,689 citations between 1992 and 2009. Te ACM/IEEE Joint Conference on Digital Libraries (JCDL)—which had 1,138 publications and 4,485 citations between 1996 and 2010—follows closely behind TREC.

Figure C-3. Cumulative Number of Publications and Citations for TREC



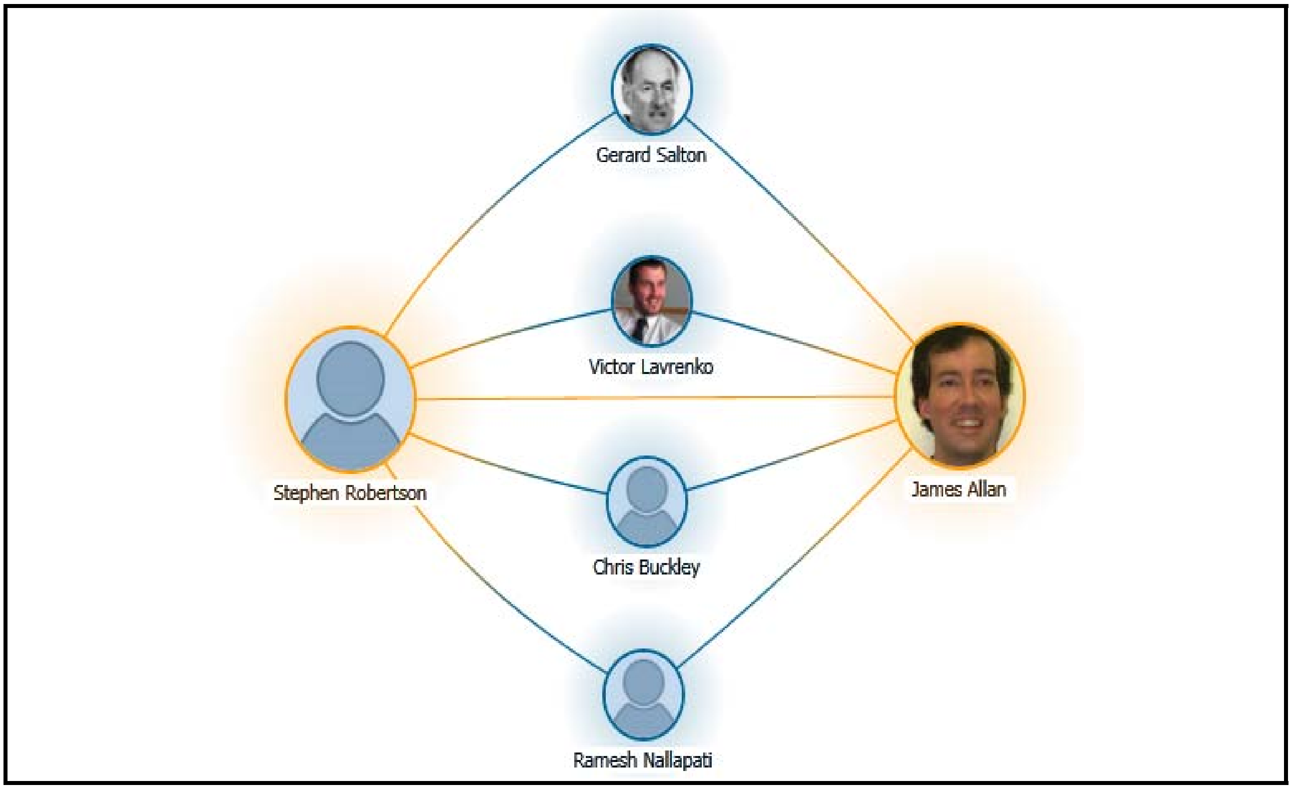
Source: Data from TREC Program Office and http://academic.research.microsoft.com/Conference/422.aspx.

C.3.2 Information Flow Between TREC Participants

Given our desire to understand how the flow of information between TREC participants has been expressed in explicit coauthored papers, we obtained a sample of TREC authors from the two sectors of interest—academic and commercial IR companies—and examined the coauthor paths that join them.

An example is the coauthor path between Stephen Robertson (Microsoft, 32 TREC publications) and James Allan (University of Massachusetts, 26 TREC publications). As Figure C-4 shows, in addition to the one paper coauthored by Robertson and Allan (who happen to be the authors with the highest number of TREC publications), a linkage exists between them through Chris Buckley (Cornell University, 24 TREC publications). Indeed, Robertson and Buckley coauthored 2 publications, and Buckley and Allan coauthored 12 publications.

Figure C-4. Coauthor Path Between Stephen Robertson and James Allan

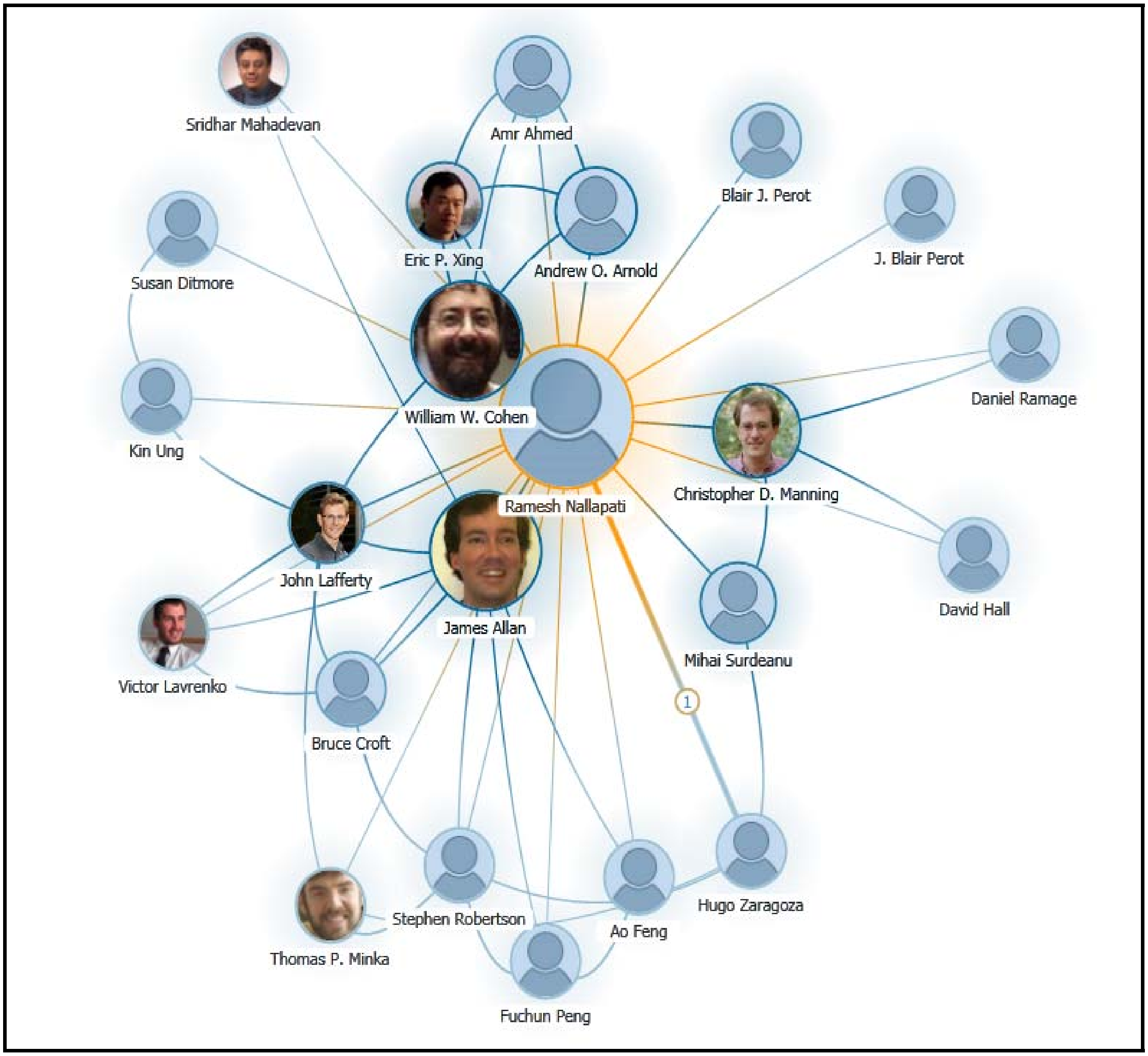


Source: http://academic.research.microsoft.com/VisualExplorer.aspx#1720436&1426902. Note: Pictures are available for some individuals for whom a picture was submitted to Microsoft Academic Search.

The coauthor path between Robertson and Allan also goes through Ramesh Nallapati (Stanford University); Nallapati has one publication coauthored with Robertson and eight publications coauthored with Allan. Although Nallapati has never published in TREC, his publication graph includes many collaborators who have, as shown Figure C-5. It is sensible to conclude that some of the knowledge shared between Allan and Nallapati has made its way into publications by Nallapati and his collaborators.

We found many other cases where TREC authors, both academic and private-sector researchers, were linked indirectly by coauthor paths through other TREC authors. A more quantitative analysis would require a full examination of the data set to obtain metrics that describe the type of relationship between the authors weighted by time. Indeed, the current analysis does not include the fact that there is a time component in the flow of information. Clearly information transfers over time from person to person, so a more thorough analysis would provide a measure of the level of influence of one author’s contributions as a function of when the coauthored publication appeared in the literature.

Figure C-5. Coauthor Graph for Ramesh Nallapati



Source: http://academic.research.microsoft.com/VisualExplorer.aspx#803040.

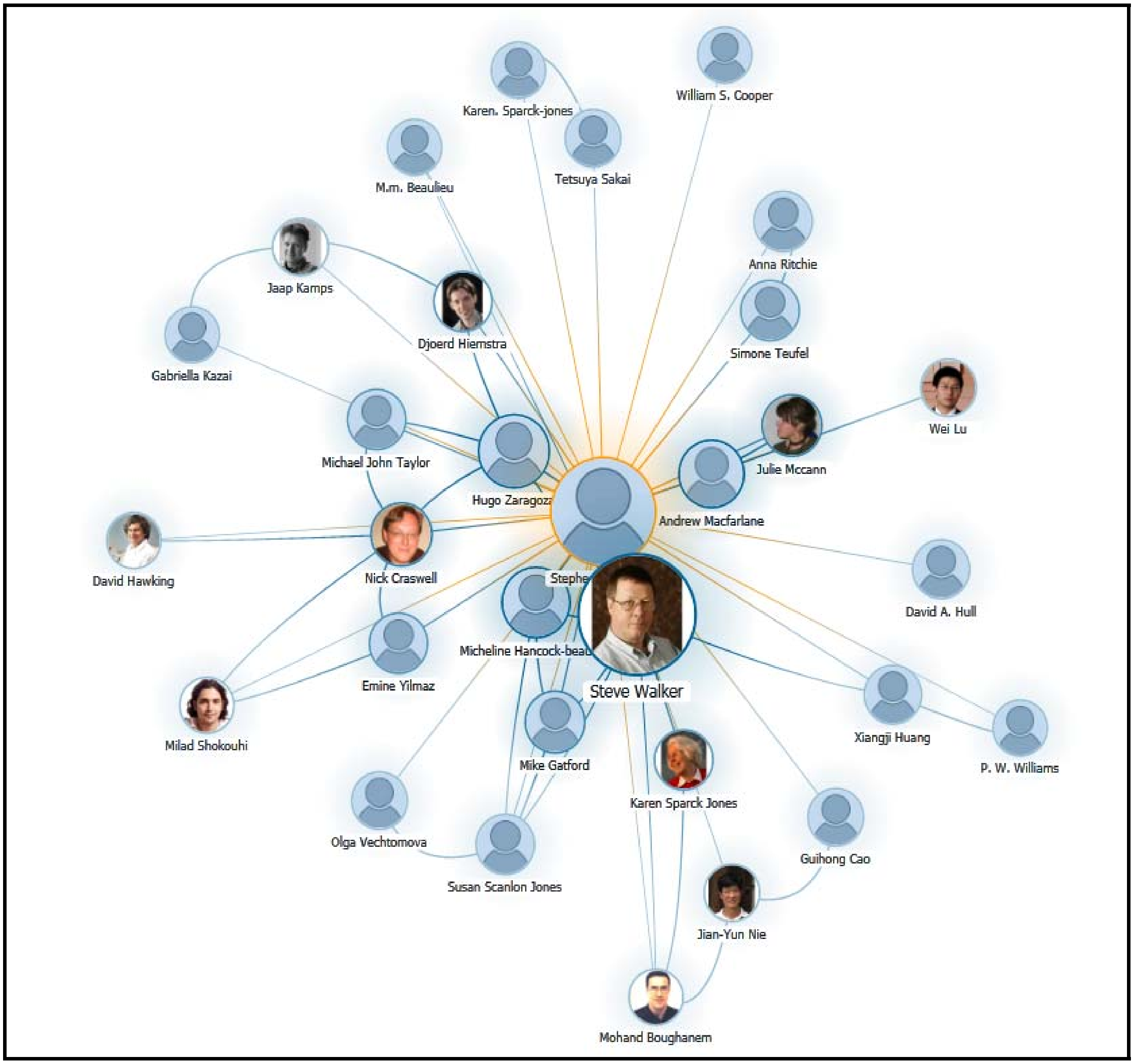
C.3.3 Information Flow Between TREC Participants and Their Collaborators

Building from coauthor relationships, we wanted to obtain more general metrics that describe the impact of TREC, as evidenced by the coauthorships between TREC researchers and their colleagues.

Degree Centrality and Hubs of Authority

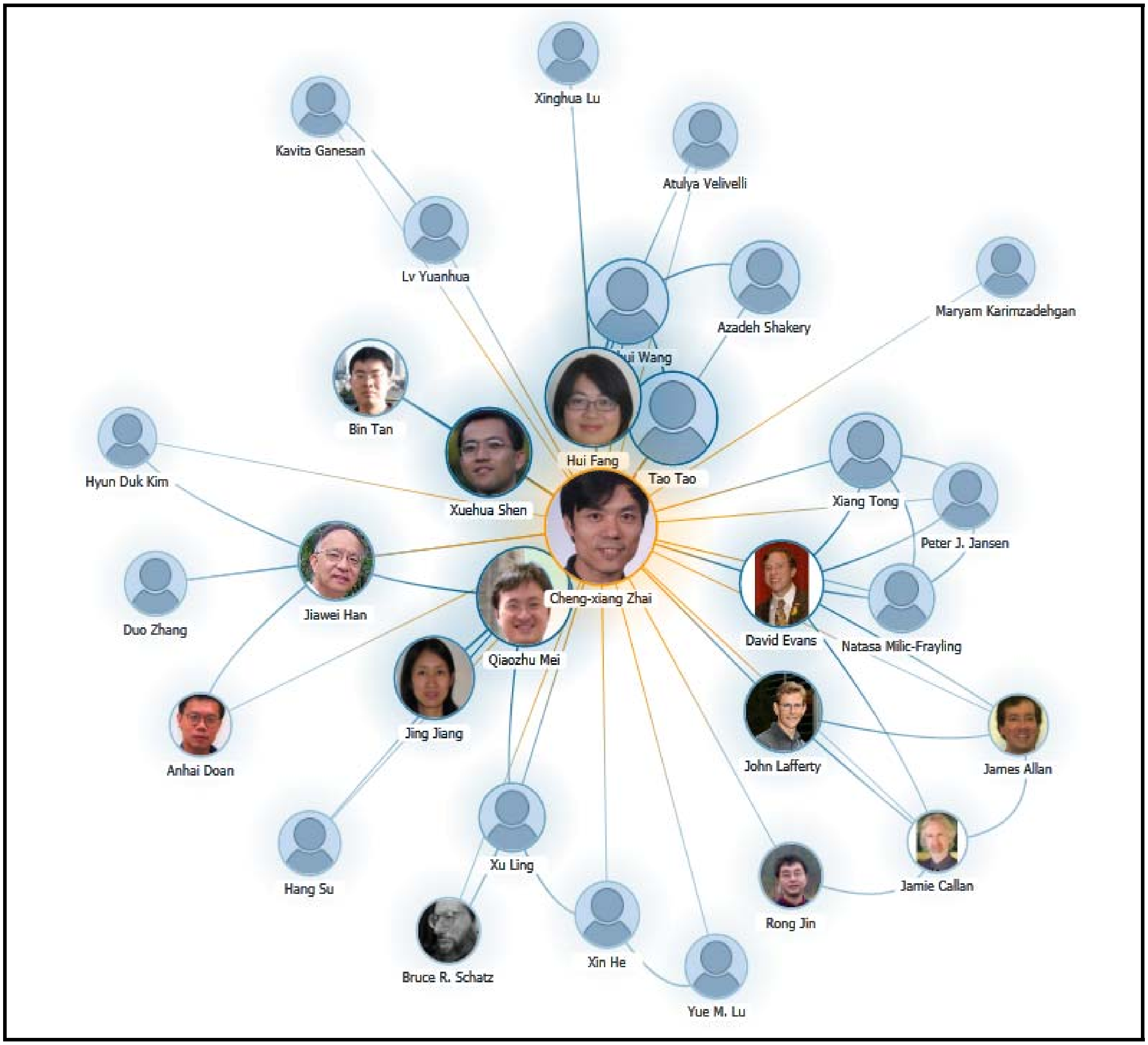
Degree centrality is the number of coauthorships in which a TREC researcher has participated. We found several cases with high degree centrality, which indicates that the author is an active participant in the general research community and is indicative of being a "connector" or "hub" in the coauthorship network and, therefore, may have an advantaged position in the network in terms of influence. Note that higher degree centrality does not necessarily imply that the author is the most connected person in the network, because an author may have a large number of coauthored papers, the majority of which were done with authors who were not as prolific. For example, Figure C-6 shows the coauthor graph for Stephen Robertson, and Figure C-7 shows the coauthor graph for Cheng-xiang Zhai. Both show high degree centrality.

Figure C-6. Coauthor Graph for Stephen Robertson



Source: http://academic.research.microsoft.com/VisualExplorer.aspx#1720436).

Figure C-7. Coauthor Graph for Cheng-xiang Zhai



Source: http://academic.research.microsoft.com/VisualExplorer.aspx#1710090.

C.3.4 IR Analysis: Concluding Remarks

Some of the key findings of the IR analysis include the following:

- We found several situations where TREC authors from academic institutions collaborated with authors from commercial IR companies.
- Some of the most prolific authors of TREC papers are also considered "hubs," indicating their high level of influence in the network.
- In many cases, we found individuals with high degree centrality who never presented at TREC, but whose position in the coauthor network indicated their role as "hubs" or

centers of authority; therefore, we could conclude that information flowed through them to other authors.

These findings support the suggestion that NIST's investments in TREC have had a far-reaching impact in the IR field, and likely beyond, by providing a central point where information can be shared not only among TREC investigators, but through them, with other investigators in the field. The level of impact was expressed most importantly not only in terms of citations but in terms of past and present collaborations, which indicate a more intimate transfer of information from researcher to researcher.