O'REILLY®

2nd Edition

# Enterprise Search

ENHANCING BUSINESS PERFORMANCE

Martin White

O'REILLY®

# Enterprise Search

Is your organization rapidly accumulating more information than you know how to manage? This updated edition of *Enterprise Search* helps you create an enterprise search solution based on more than just technology. Author Martin White shows you how to plan and implement a managed search environment that meets the needs of your business and your employees. You'll learn why it's absolutely vital to have a dedicated staff manage your search technology and support your users.

New material for this second edition includes SharePoint 2013 search, managing open source search development, website search, designing the search user, and assessing search performance. Chapters now include a Further Reading section for computer science and information science students.

Topics include:

- 10 critical success factors to assess organizational search maturity
- Essential skills needed to support a successful search application
- How to specify and manage open source search development
- How to manage SharePoint 2013 search
- Methods to assess the business impact of search
- Best practices in user interface design
- The importance of search for websites
- What to include in a search strategy

**Martin White** is an intranet and information management strategy consultant, and has been running Intranet Focus Ltd since 1999. He's also been the Visiting Professor at the Department of Information Studies (now the iSchool) at the University of Sheffield, and was elected a Fellow of the Royal Society of Chemistry in 2006.

"Martin White has come up with the most useful book about enterprise search on the market. A must-read for anyone trying to be successful with search."

**—Jeff Fried**
CTO, BA Insight

"Martin's roadmap has demystified the complexity of all those ingredients needed to make finding information more effective in the workplace."

**—Elaine Toms**
Professor of Information Science,
University of Sheffield

WEB/SITE DIRECTORIES

**US $29.99** **CAN $34.99**
ISBN: 978-1-491-91553-0

9 781491 915530 52999

Twitter: @oreillymedia
facebook.com/oreilly

SECOND EDITION

# Enterprise Search

## *Enhancing Business Performance*

*Martin White*

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

### Enterprise Search

by Martin White

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

**Editor:** Allyson MacDonald
**Production Editor:** Shiny Kalapurakkel
**Copyeditor:** Jasmine Kwityn
**Proofreader:** Sonia Saruba
**Indexer:** Stephen Ingle, Word Co.
**Interior Designer:** David Futato
**Cover Designer:** Karen Montgomery
**Illustrator:** Rebecca Demarest

October 2015: Second Edition

**Revision History for the Second Edition**
2015-10-09: First Release

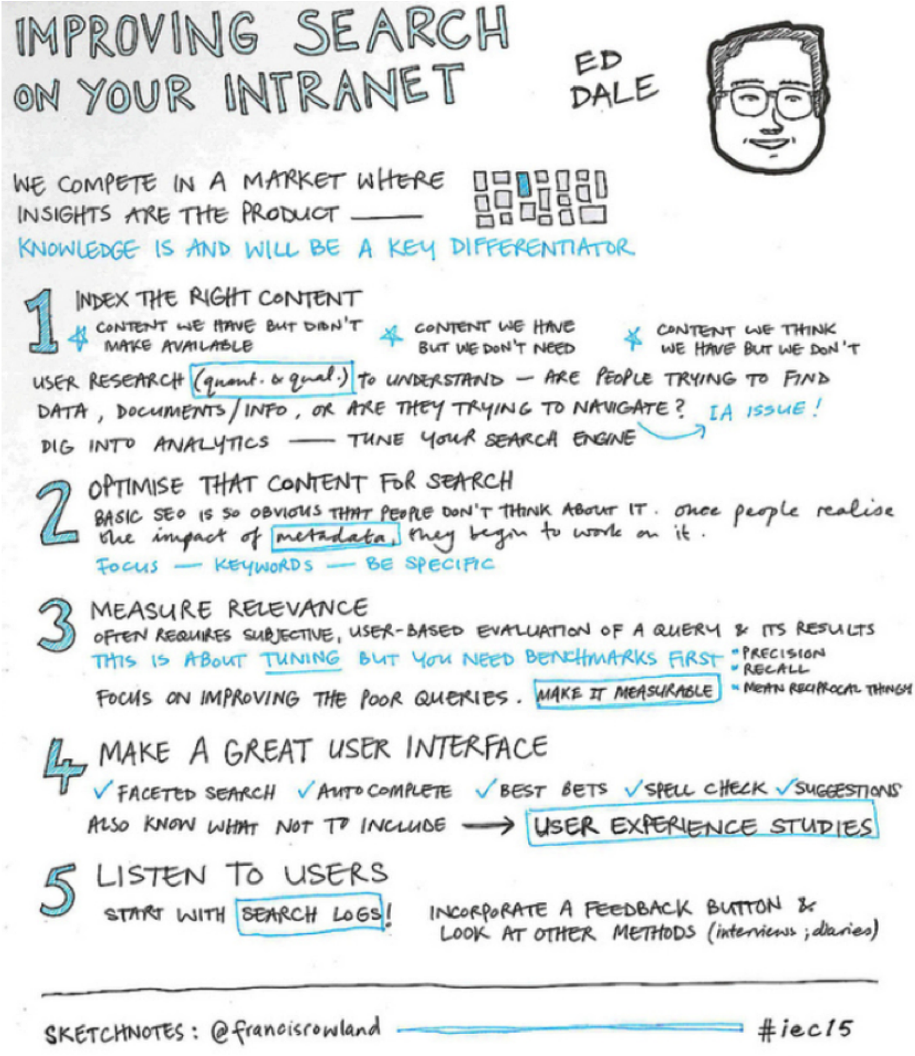
See *http://oreilly.com/catalog/errata.csp?isbn=9781491915530* for release details.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-91553-0

[LSI]

# Table of Contents



IMPROVING SEARCH ON YOUR INTRANET

ED DALE

WE COMPETE IN A MARKET WHERE INSIGHTS ARE THE PRODUCT —

KNOWLEDGE IS AND WILL BE A KEY DIFFERENTIATOR

1 INDEX THE RIGHT CONTENT

- CONTENT WE HAVE BUT DIDN'T MAKE AVAILABLE
- CONTENT WE HAVE BUT WE DON'T NEED
- CONTENT WE THINK WE HAVE BUT WE DON'T

USER RESEARCH (quant. & qual.) TO UNDERSTAND — ARE PEOPLE TRYING TO FIND DATA, DOCUMENTS / INFO, OR ARE THEY TRYING TO NAVIGATE? IA ISSUE!

DIG INTO ANALYTICS — TUNE YOUR SEARCH ENGINE

2 OPTIMISE THAT CONTENT FOR SEARCH

BASIC SEO IS SO OBVIOUS THAT PEOPLE DON'T THINK ABOUT IT. once people realise the impact of metadata, they begin to work on it.

FOCUS — KEYWORDS — BE SPECIFIC

3 MEASURE RELEVANCE

OFTEN REQUIRES SUBJECTIVE, USER-BASED EVALUATION OF A QUERY & ITS RESULTS

THIS IS ABOUT TUNING BUT YOU NEED BENCHMARKS FIRST

- PRECISION
- RECALL
- MEAN RECIPROCAL THINGY

FOCUS ON IMPROVING THE POOR QUERIES. MAKE IT MEASURABLE

4 MAKE A GREAT USER INTERFACE

✓ FACETED SEARCH ✓ AUTO COMPLETE ✓ BEST BETS ✓ SPELL CHECK ✓ SUGGESTIONS

ALSO KNOW WHAT NOT TO INCLUDE → USER EXPERIENCE STUDIES

5 LISTEN TO USERS

START WITH SEARCH LOGS!

INCORPORATE A FEEDBACK BUTTON & LOOK AT OTHER METHODS (interviews; diaries)

SKETCHNOTES: @francisrowland — #iec15

*This sketch by Francis Rowland summarizes the intranet search success factors proposed by Ed Dale (E&Y) at the IntraTeam Event, Copenhagen, February 2015.*

# Preface

No matter who they work for or what work they do, most people seem to want their internal search application to work like Google. When you ask what it is about Google Search that makes it so desirable, the response is usually about the very simple interface, the speed with which results are presented, and the high probability that the first page of results will provide at least a good start in the information discovery process. That's quite an achievement for a company whose business is advertising and not search. The investment in research and development at Google reached $10 billion in 2014, and most of this is spent on nearly 20,000 staff. This effort was largely responsible for Google selling close to $60 billion of advertising in 2014. Does your organization spend 13% of its revenue on search?

There is more to the Google success story than technology. Website owners and contributors spend a considerable amount of time and effort to make sure that Google indexes their content and presents it at the highest possible position in a list of search results. However, internally, there are never any rewards for making sure that information is of the highest quality and presented in a way that will make it easy for the technology to work its retrieval magic. There is rarely more than one lonely person with the responsibility for supporting the search application and making sure that it is tuned to meet user requirements. Investment in search is never seen as a priority.

There is another aspect of Google Search that is not fully appreciated. For almost any search, the same information will be presented from multiple sources, be they restaurant reviews, airline flight times, or the distance from the Earth to the Moon. The real reason that people want Google is that they trust it to deliver *something*, even if not *everything*. For most purposes, something is good enough. If Google cannot find British Airways flight times from London to New York, then you can check out the airline or the airport or a multitude of other sources.

Inside an organization, information that cannot be found is information that, in effect, does not exist. It has vanished. Permanently. No one will ever see it again. There is a chance that a call to a colleague might result in a document with the anticipated title, but can you be certain it is the latest version? *Something* is not good enough. Meanwhile, the colleague might be annoyed to be interrupted yet again by other colleagues asking about the same document. In the course of writing this edition, I interviewed over 20 senior managers in a global business about their requirements for a proposed intranet upgrade. Within a minute of the introductions, without exception, they started to talk at length about the poor quality of the intranet search, and several wanted Google. I explained to them that the search application they had was significantly more powerful and had a wider range of functionality than a Google enterprise appliance. It came as quite a shock!

The fundamental problem is that organizations do not see information as an asset. They know how many desks there are, how much money is in the bank, the names of every employee and customer, and the depreciated value of buildings and IT hardware. They have no idea of the amount of information they have. The CIO might quote a total storage volume, but that is not the same as the amount of credible, trustworthy information. It is usually not until the intranet is migrated from a perfectly usable content management system (CMS) to SharePoint that the organization finds that it has perhaps 500,000 documents hidden (and I use the word advisedly) in the application.

CEOs, managers, and other leadership personnel are now beginning to appreciate that information that cannot be found and shared might well be putting their organizations at risk. All directors have a responsibility to minimize the risk profile of their organizations, but rarely does information risk appear explicitly on the risk register. If it did, the market demand for search would escalate exponentially. *Information management* and the currently more popular term of *information governance* are gradually moving center stage. *Knowledge management* still has a role to play, but as with Big Data, it is not easy to translate into improvements in revenues and profits. The impact of information is much easier to assess.

Google is certainly impressive, but it is not quite the information access panacea that it seems to be. Searching scholarly articles that have very few links is very challenging, which is why Google offers Google Scholar. Google knows that one size does not fit all, but that is a message lost on far too many IT managers. Entering search terms into the Google search box seems very easy, but why then are there books with several hundred pages of advice on how to get the best from Google? People need to be trained in how to search. Even Google is not totally intuitive.

This second edition is almost twice the length of the first edition, published in late 2012. The increase in size is not because there have been dramatic changes in technology, but rather, because in the previous edition I passed over subjects on search based on my own assumptions that they were already known. The feedback from readers was that I had skimmed over some topics, notably website search and user interface design. Moreover, both open source search and the search application in SharePoint 2013 have increased the awareness of what search can offer.

If you follow the guidance I offer in this book, then I am confident that you will achieve one of the most important attributes of Google—users will trust your enterprise search applications to deliver significant amounts of relevant information on which they can base decisions that will benefit their organizations and their careers. In my view, search is a decision support application. If you agree with me, then making a business case for the resources needed to capitalize on the existing investment in technology will be very easy, and could potentially save your organization from unnecessarily purchasing new technology "because the current technology is broken."

This book is technology-light. I have described the core elements in search technology in a way that does not require a degree in computer science, primarily because I don't have one. I am an information scientist. There are chapters on open source search and on SharePoint search but they are written for business managers and not for developers and systems administrators.

In the final analysis, good search depends on good content and good people. People who have the time to write and tag high-quality content, people to support search, people who have been trained to use search applications to their maximum potential, and people who are willing to provide an appropriate level of investment.

## How to Use This Book

This book has been written to help business managers, and the IT teams supporting them, understand why effective enterprise-wide search is essential in any organization. The focus is on how to ensure that search applications deliver the information needed to make sound business decisions that enhance business performance. This second edition is twice the size of the first edition, which was published in 2012. Virtually every chapter has been revised, and chapters on open source search, SharePoint search, user interface design, website search, and search governance have been added.

One of the most visible changes to this edition is that each chapter includes a list of books, papers, reports, and web resources for readers who wish to dig deeper into search technology and its applications.

The first two chapters set the scene, explaining why effective enterprise search is essential to any organization. Chapters 3 and 4 then provide a low-technology description of how search works. After an overview of the search business in Chapter 5, there are individual chapters on open source search (Chapter 6) and on SharePoint SP2013 search (Chapter 7), as both are widely used for enterprise search purposes.

Chapter 8 is all about search governance, and in particular, the skills needed to provide support to enterprise search. In almost every case I have come across of search failing to meet the requirements of the business, it is because there is not an appropriate level of skilled support for an application that is used probably every day by most employees.

The process of specifying, selecting, and implementing a search application is set out in Chapters 9 to 14, including substantial chapters on how to define user requirements and the design of user interfaces. Search performance is not just about speed but about employees feeling that the search application meets their expectations and has a significant impact on business performance. Chapter 15 covers how to assess technical performance, discovery, satisfaction, and business impact.

Corporate websites are also one element of an enterprise search strategy and are the subject of Chapter 16, followed by chapters on eDiscovery (Chapter 17) and content analytics (Chapter 18). In Chapter 19, I have been brave enough to suggest how search technology will develop over the next five years, by which time I will have retired!

The book sets out all the information you need, or need to collect, to write a search strategy. Every organization writes strategies to its own house style. Appendix A provides an A–Z of all the topics that need to be included. Finally, there is a list of 14 critical success factors, up from 10 in the first edition (Appendix B), a Glossary, a list of books and blogs on search (Appendixes A and B), and a list of search vendors (Appendix E). Hopefully, this will be all you need to make enterprise search a success for your organization.

## Safari® Books Online

*Safari Books Online* is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

## How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/1VHpGLW*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

## Acknowledgments

I could not have written this book without the generous support of many colleagues over quite a number of years. In particular, I must thank Stephen Arnold, Paul Clough (Information School, University of Sheffield), Jed Cawthorne, Ed Dale (E&Y), Jeff Fried (BA Insight), Charlie Hull (Flax), Miles Kehoe (New Idea Engineering), Helen Lippell, Andrew McFarlane (City University), Agnes Molnar, Valentin Richter (Raytion), and Professor Elaine Toms (Information School, University of Sheffield). Agnes, Charlie, Jed, Jeff, and Elaine read through substantial portions of the book at a draft stage and made invaluable contributions.

Other colleagues who in various ways have shaped my understanding of search technology and search good practice include Denise Bedford, Thomas Borch, Susan Feldman, David Hawking, Jane McConnell, Kristian Norling, Sam Marshall, Peter Morville, James Robertson, Tony Russell-Rose, Michal Szlanowski, and many members of the Findwise team.

Janus Boye (JBoye), Erik Hartman (Hartman Event), Kurt Kragh Sørensen (Intra-Team), and Val Skelton (UKeiG) have given me many opportunities to run search workshops at their events. These have been invaluable in learning from the experiences of enterprise search managers across Europe. Katherine Allen and the team at Information Today Europe have supported my vision for the Enterprise Search Europe conference with enormous enthusiasm and skill. I have had the honor of being a Visiting Professor at the Information School, University of Sheffield, since 2002 and everyone on the academic staff has given generously of their time and expertise.

I am grateful to the Aberdeen Group, AIIM, Alta Plana, and Findwise for the use of information from their surveys. Over the last decade, I have carried out many enterprise search consulting assignments, but I am not in a position to list the organizations involved. Each of these assignments has given me additional insights into the technology and use of enterprise search.

I would like to thank Allyson MacDonald at O'Reilly Media for commissioning this second edition and guiding it from spreadsheet to bookshelf. It is a privilege to be an O'Reilly author.

It has not been easy for my wife, Cynthia, when people ask her what I do for a living. Being an information scientist is fascinating for me but difficult for Cynthia to describe. She has been immensely supportive during 12 career changes and 8 books.

This book is dedicated to the memory of W. Gordon Graham, who was Chairman of Butterworths when I was involved in launching Lexis in the UK. Gordon was a distinguished publisher and an early and passionate believer in digital publishing. Although I left Reed Publishing in 1984, he continued to be a valued friend and mentor. He died early in 2015.

Search has been a part of my life since 1970 when I was using 10,000-hole optical coincidence cards at the British Non-Ferrous Metals Research Association. The principles I learned at that time in terms of indexing, constructing a query, relevance, recall, and precision have sustained me throughout the last 45 years. In many respects, nothing has changed, and yet everything has changed.

# CHAPTER 1
# Searching, Finding, Deciding

No one searches for information just for the fun of it. Behind the process of searching is a requirement to find the information needed to make a decision. How often have you had to make a decision without feeling totally confident that you had found all the relevant information you needed?

On June 10, 2000, the London Millennium Footbridge across the River Thames was opened. Shortly thereafter, it was obvious there was a problem. It became known as the "Wobbly Bridge" because it started to sway as pedestrians began using it. Within a couple of days, it was closed down, reopening in 2002 with a new damping system. In the course of investigating the issue, it was found that this was a known problem. The paper describing it had been published in the journal *Earthquake Engineering and Structural Dynamics* in 1993, but the designers of the bridge were not aware of it.

What might be the outcome for your organization in a similar situation?

Reading this book and taking appropriate action will help everyone in your organization make better decisions. Some of these decisions will be about implementing and managing search applications. Other decisions will have a direct impact on the organization. Enterprise search is a decision support application. It is a business-critical application that needs to be planned, resourced, and managed at a level commensurate with the decisions that it is supporting. That is why the strapline for this book is "Enhancing Business Performance." It is easy to devote a lot of effort to analyzing search logs and producing visually attractive dashboards showing an increase in use of a search application. The increase is almost certainly due to the fact that the volume of information in the organization is increasing and so each individual item becomes more difficult to find.

There are only two important success metrics for search. The first is that users trust it to find all the information they need. The second is that it makes an impact on business performance. The objective of this book is to help your organization improve both the level of trust and the impact on performance.

## Every Day Is a Decision Day

Every day people make the news headlines because they made the wrong decision. The financial meltdown of 2008 was arguably an information problem. Loans had been made to people purchasing homes without adequate security. The pressures of making sales targets led to an inadequate review of the circumstances of the people asking for loans, and senior managers in the banks had no information about the scale of the problem.

Once upon a time, you could at least walk into your office in the morning and feel reasonably certain about the decisions you would need to make. With the arrival of 24/7 mobile access, reductions in staff, and difficult economic and market conditions, you may well get a call at any time during the day from colleagues just about to walk in to a prospective client and have just realized that they did not have a critical piece of information about the client or the proposal they were making.

That puts the pressure on to find information that could have a very positive impact on the bottom line. Fortunately, your organization has invested in an enterprise search application! You enter a few keywords into the search box, sit back, and within a few seconds discover that there either seems to be no information at all on the query you have made, or you find that there are over 3,000 documents and you only have a few minutes to hunt through them to provide a response.

When we are dealing with decisions that are based on some standard business processes, such as setting up a project or writing a monthly report, we often rely on browsing through the information architecture of an intranet, shared file collection, or a document management system to find the information we need.

In 2014, PwC published a global survey of decision making in organizations with a particular focus on decisions that could change the business direction of the organization and increase revenues by millions of dollars. The survey was conducted among over 1,000 senior executives around the world. It showed that almost half of these executives make a “big” decision every month and a further third make a “big” decision every three months. The timing of these decisions is almost always opportunistic and not to a specific schedule.

These “big” decisions are supported by many other employees in the organization making decisions that will eventually feed into the briefing given to a senior executive.

Search becomes critical when there is time pressure and a need for an immediate solution. When using the Web for information, we may well be unaware about how often we use Google or Bing to find a website, then use the search function on the website itself, and finally go back to Google or Bing to check that we have not missed out on a better deal or a more recent source of information.

The situation inside an organization is just the same.  We expect it to be as easy to use as Google and at least as effective, providing us with the information we need on the first page of results. Anything less, and the search application is regarded as a failure. Google and Bing have huge scale, and an immense amount of development has gone into providing search experiences across the Web. Searching for information inside a single organization seems like it should be easier, but in reality it is much more challenging.

## Information as a Corporate Asset

Many companies attach asset numbers to all of their property, be it a wastepaper bin or a complex machine tool. All those assets are logged in a database and their residual financial value will be given on the balance sheet of the company. The balance sheet will also show the financial assets of the business.

No matter how hard you look, there will be two corporate assets missing from the balance sheet. One of these is the employees, though at least there will be a record in the annual report of how many employees there are, possibly categorized by location or gender. But what about the information assets of the business, and the knowledge assets possessed by each employee? International accounting standards do not allow for information to be capitalized as an asset because there is no definitive way of calculating its value. The value of a piece of information is unique to an individual at a particular point in time. In search terms, it has a different *relevance* (I'll have much more to say about relevance later in this book).

In addition to recording every physical asset they own, companies entrust employees to manage these assets and make decisions about when and how they should be replaced or upgraded. In most companies, no one owns information as a corporate asset, even though there may be someone with the title of Chief Information Officer. There is now a growing concern among senior managers about the sheer scale of corporate information resources with the arrival of the concept of Big Data. With hundreds of applications being used each day inside even a modest-sized company, the amount of data and information that is being collected is often poorly understood. Worse, because of the low cost of storage, nothing is ever deleted. As a result, the rate of growth is a combination of new information and old information, with the assumption that all information has a value (of course, that value can only be realized if the information can be found!).

The term *unstructured information* is widely used to describe documents, emails, blogs, and other text information, and more recently, rich media applications. In fact, this information does have a structure, in that there is usually a title, an author, a date, and perhaps section headings and tables. The term came into use to distinguish these categories of information from *structured databases* in which data is stored in defined fields such as Address Line1, Address Line 2, Town, and so on. For many years, the UK search vendor Autonomy emphasized the fact that unstructured information represented 80% of the total information assets of the organization. No evidence was ever presented for this assertion, which seemed to be based solely on the Pareto principle. More important, there is no relationship between volume and value.

Until recently, enterprise search was used primarily to search unstructured text, and it therefore needed to be able to cope with the issues of language and semantics. For example, consider these two sentences:

> Noah loaded boxes into the van.
> Noah loaded the van with boxes.

In the case of the first sentence, the number of boxes could be any number of two or more. In the second sentence, there is the implicit message that the van was totally full of boxes, though we cannot be sure.

The textual differences between the sentences are very small but semantically very important. In almost every conversation we have, we are constantly checking whether we have fully understood what others are saying, perhaps asking for clarification from time to time. In the case of a document that might have been written several years ago, we cannot have this type of conversation, and yet we expect a search engine to be able to read and understand the document, and then be able to say with certainty that the document contains information that is relevant to the search we have carried out and list it in the top few results.

## Information Quality

Search software is both very smart and very stupid. Search applications can index information at an amazing speed and deliver the results of a search within a second or so of being asked a query. However, no search application yet built can distinguish between high-quality and low-quality information. It treats both just the same. When users criticize a search application, they make the assumption that the problem lies with the technology. Almost always the problem lies with people, not technology.

Although employees may spend much of the day in creating information, there is usually very little guidance on the process of information creation and curation. Ensuring that every content item has an informative title, a date, and author will make a difference, even more so if there is a review process that at least on a yearly basis ensures that the item is still fit for purpose. This is a particular problem when companies have merged. A search on a topic may well find information dating back to before the merger, leaving the employee to make a judgement on whether the information is still valid.

In the second instance, it is a lack of people to create an adequate depth of experience in the search support team.

Without addressing these people issues, there is no point in investing in a new search application. All the new technology will do is highlight a lack of content quality and management support even more quickly than in the past.

## Information: Important and Yet Invisible!

It is only over the last few years that surveys have been undertaken about the challenges of finding information inside organizations. In this respect, 2014 was a very good year, as the Association for Information and Image Management (AIIM) conducted its first ever survey of search implementation and Findwise published its third Enterprise Search and Findability Survey. The AIIM respondents were mainly based in North America, and the Findwise respondents were mainly European organizations. Although the questions asked in the two surveys were slightly different, a very consistent picture emerged from these and other surveys.

Quotes from some of these studies include:

> *For 71% of the organizations polled search is vital or essential, yet only 18% have cross-repository search capabilities. The IT department takes responsibility for search in 52% of organizations but only 25% feel that it should be so. 38% have not tuned or optimized their search tool at all, and half of the responding organizations allocate less than half an FTE to support search applications.*
>
> *Better decision-making and faster customer service are the top benefits from improved search tools. Only 14% were required to make a financial business case for search investment. 42% consider that they have achieved pay-back from their investment in search tools within 12 months or less, and 62% achieved pay-back in 18 months.*
>
> —AIIM, "Search and Discovery—Exploiting Knowledge, Minimizing Risk" (2014)

> *86% of respondents reported that it was either very important or important to improve the ability to find information in their organizations but only 38% had a strategy for doing so. 48% said that employees found that it was very difficult or difficult to find the information they needed and less than 5% reported that their employees were very satisfied with the search applications.*
>
> —Findwise, "Enterprise Search and Findability Survey 2014"

> *Only 23% of organizations have enterprise-wide search and a further 16% have a more limited level of implementation. 22% have no plans to implement enterprise-wide search. Only 11% are very satisfied with search, a figure that has changed little over the last five years.*
>
> —NetStrategyJMC, "The Digital Workplace" (2014)

The situation seems to be getting worse, not better. The 2015 edition of the Digital Workplace Trends Report showed that the percentage of respondents who reported that they were either very satisfied or even moderately satisfied was lower than in the 2014 edition, and was in fact at the lowest level since the survey was launched in 2009. The rate of growth of information stored digitally in an organization seems to be overwhelming the capability of the organization to provide adequate search solutions.

The headline summary of all these surveys is that information is becoming more important, more difficult to find, and yet there is an inadequate level of investment in improving information quality and the resources to enhance search effectiveness.

# Enterprise Search

It is time for some definitions. This book is entitled *Enterprise Search*, so what is *enterprise search*? Here's one possible definition:

> An enterprise search application enables employees to find all the information that the company possesses without the need to know where the information is stored.

The position I take in this book is that enterprise search is not about selecting and installing a single search application that will index every item of information and data owned by the organization.

This is my definition:

> Enterprise search is a managed search environment that enables employees to find information they can rely on in making decisions that will achieve organizational and personal objectives.

Many companies already have one or more search applications, either operating as a discrete search application or embedded into another enterprise application. Trying to replace all of these with one HAL-like enterprise search application is not a realistic strategy. One of the buzzwords in enterprise search is *federated search*, which is an attempt to provide one single search box linked to one single search application that has an index of every single item of information and data that the organization has created.

The sales pitch is that a single enterprise search application can break down information silos. The issue is whether that will have value. In many organizations, the information silos reflect different areas of business or technology. Searching across all these silos could result in having to work through a very long list of search results, many of which seem to be highly relevant but in fact are highly relevant to a particular silo. There can also be some significant access permission management problems that need to be solved. This is not to say that federated search is never an effective approach, but the decision to implement federated search needs very careful consideration of both the benefits and risks. These are  discussed in Chapter 4.

The downside of this definition is that it excludes search applications on the corporate website. These should certainly be included in any enterprise strategy, even though they may be owned by corporate communications and use a hosted search application. The skills needed to support website search are the same as those for internal enterprise search, and there will be benefits from taking an integrated approach to their management.

## Search and "Information Retrieval"

Information retrieval can be regarded as the science (largely mathematics) behind search. It is a branch of information science and dates back to the mid-1950s. It has been defined as follows:

> Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, web pages, online catalogs, structured and semi-structured records, and multimedia objects.

There are two different perspectives of information retrieval research: the first considers the computer technology of information retrieval, such as ways of building efficient indexes and finding ways to handle multiple languages; and the second is user-based, looking at search user interfaces and how people go about constructing search queries. Although there are some very distinguished university departments of information science around the world (many now called information schools, or iSchools), few teach information retrieval in any depth as an undergraduate course, and this means that the annual output of graduates with skills in search implementation is very low indeed. Computer science departments, of which there are many more, also pay little attention to the science and technology of enterprise search, even though many of the major IT vendors, such as IBM, Oracle, HP, and Microsoft, have a long history of carrying out information retrieval research, as, of course, does Google.

The scale of the science behind search can be seen in the fact that the standard textbook on the subject, *Modern Information Retrieval* (Addison-Wesley) by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, is 700+ pages in length and includes a bibliography of nearly 2,000 references. Marti Hearst's *Search User Interfaces* (Cambridge University Press) is 300+ pages and has around 500 references in its bibliography to research papers on user interface design. *Elasticsearch: The Definitive Guide* is 700+ pages.

Search and "Information Retrieval" | 7

Sadly, there seems to be a gulf between the information retrieval community and the enterprise search community. Some information retrieval conferences do include sessions where papers from the commercial search world are presented. For some years, there was an Enterprise Search Summit in New York, but this has been discontinued. There is an Enterprise Search Europe conference where the emphasis is very much on enterprise search implementation and management. There are signs that the situation is now starting to change and in the future, much closer ties are likely to develop between the information retrieval community and search software vendors and users.  A good example is the annual Search Solutions conference organized in London by the Information Retrieval Specialist Group of the British Computer Society. Hopefully the rapid adoption of open source search solutions will catalyze the launch of new conferences in the next few years.

## A Short History of Information Retrieval

One of the problems facing anyone interested in search is that there are a number of parallel domains:

- Enterprise search
- Site search of a public website
- Internet search
- Search engine optimization
- Information retrieval

Internet search and search engine optimization fall outside of the scope of this book, but information retrieval certainly does not, and yet is probably a totally unfamiliar topic to most search managers.

The term was first used by Calvin Mooers, a pioneer in the early history of the development of computer technology in the 1950s and 1960s. Mooers made the point that one of the challenges of information retrieval is that the person contributing the information to a system has no idea of when the information will be found by a searcher and what it will be used for.

In 1959, he coined Mooers's law and its corollary:

> An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it.
>
> Where an information retrieval system tends not to be used, a more capable information retrieval system may tend to be used even less.

Mooers's law is a reflection that an information retrieval system will not be used if in doing so there is a chance that the information found may put the user to some inconvenience.  In 1989, J. Michael Pemberton rewrote this law as follows:

> The more difficult and time consuming it is for a customer to use an information system, the less likely it is that he will use that information system.

This is the form that tends to be quoted today.

The corollary needs a brief explanation. What Mooers realized was that if users have a poor experience with an information retrieval system, they are unlikely to try again in another system even if they are told it will produce better results. They are more likely to send an email or call someone on the telephone.

Although much of the early work on information retrieval algorithms was undertaken in the United States, Stephen Robertson and Karen Sparck Jones, working at Microsoft Research, Cambridge, and the University of Cambridge also developed some important elements of current search applications. It was in 1972 that Sparck Jones published a paper on what would become known as inverse document frequency (IDF). Her intuition was that a search term that occurs in many documents is not a good discriminator, and should be given less weight in assessing relevance than one that occurs in a few documents. IDF was a heuristic implementation of this intuition and was a very significant development in information retrieval. In the 1990s, Robertson, with Sheila Walker, went on to develop a weighting model called BM25, which remains a core element of most search applications, including Microsoft SharePoint. One of the best introductions to BM25 can be found on the Microsoft TechNet site, along with a good overview of a range of other ranking models.

Gaining some understanding, even without a full appreciation of the mathematics, is important to be able to appreciate the fundamental principles of information retrieval and the challenges of being able to rank results in order of relevance.

Information science is just one of the sciences behind search—mathematical probability, computational linguistics, and computer science also come into play. Around the world there are probably several hundred academic departments offering courses in information science, of which information retrieval is a core topic. In particular, there are nearly 60 iSchools specializing in information science and information retrieval. However, research in information retrieval is often based on well-defined sets of documents and other content items.

## A Short History of Search

Search came into prominence with the advent of the web search services in the 1990s, notably Alta Vista, Google, Microsoft, and Yahoo!. However, the history of search technology goes back much further than this. Arguably, the story starts with Douglas Engelbart, a remarkable electrical engineer whose main claim to fame is that he invented the computer mouse, which is now a standard control device for personal computers. In 1959, Engelbart started up the Augmented Human Intellect Program at the Stanford Research Institute (SRI) in Menlo Park, California. One of his research students was Charles Bourne, who worked on whether it would be possible to transform the batch search retrieval technology developed in the 1950s into a service based on a large mainframe computer that users could connect to over a network.

By 1963, SRI was able to demonstrate the first “online” information retrieval service using a cathode ray tube (CRT) device to interact with the computer. It is worth remembering that the computers being used for this service had 64 K of core memory. Even at this early stage of development, the facility to cope with spelling variants was implemented in the software. Other pioneers included System Development Corporation, Massachusetts Institute of Technology, and Lockheed. The main focus of these online systems was to provide researchers with access to large files of abstracts of scientific literature to support research into space technology and other large-scale scientific and engineering projects.

These services were only able to search short text documents, such as abstracts of scientific papers. In the late 1960s, two new areas of opportunity arose, which prompted work into how to search the full text of documents. One was to support the work of lawyers who needed to search through case reports to find precedents. The second was also connected to the legal profession, and arose from the US Department of Justice deciding to break up what it regarded as monopolies in the computer industry (targeting IBM) and later the telecommunications industry, where AT&T was the target. These actions led IBM in particular to make a massive investment into full-text search, which by 1969 led to the development of STAIRS (Storage and Information Retrieval System), which was subsequently released as a commercial IBM application. This was the first enterprise search application and remained in the IBM product catalog until the early 1990s.

However, it is important to keep in mind that not all the developments were taking place in the United States. For example, a team at the United Kingdom Atomic Energy Authority took the lead in using mini-computers to support online services in the mid-1970s.

The problem with STAIRS was that at least in its initial versions it could only search for words that appeared in the document. What researchers wanted was to find information about concepts that were not present as words in a document, especially if they were working for the security services. Dynamite can be used for mining but also to make a bomb, and they needed a search system that would present results that included dynamite to a search query on bomb making. One of the innovators in the mid-1980s in developing concept searching was Advanced Decision Systems (ADS). Verity was the name of the company that was spun off from ADS to bid (successfully) on a US Department of Defense/US Air Force project. A feature of the Verity Query Language was the capability to weight topics against a taxonomy tree. Verity was also able to offer real-time indexing. Verity became one of the most innovative companies in enterprise search, and in 2003, it acquired the enterprise search business of Inktomi Corp., relaunching the application as Ultraseek. Then, in 2005, Verity was acquired by the UK-based search company Autonomy. The story of the enterprise search industry continues in Chapter 5.

## Search Is a Dialog

Earlier in this chapter, I remarked on how in conversations we are constantly engaged in a dialog to ensure that we understand what the people we are talking with are trying to convey. It is very important to understand that search is a dialog. We tend to see search as a “first strike” application; just putting a search term into Google or Bing will provide all we need in the first page of results. The reality is that Google sometimes needs to correct our spelling mistakes or prompt us with “did you mean” suggestions. On the page, there are filters that we can use to narrow down our search, and on public search sites, there is paid-for advertising that also offers solutions to our problems.

We often go into a large department store to find a birthday present, and yet I have never come across a store with a Birthday Present Department. We may look at the store directory (the information architecture) for ideas, but if we are in a hurry, we may also go to the Information Desk (the search box) for advice. There we will be asked the age and gender of the person for whom we are buying a present, and what their interests are, in order to suggest one or more specific departments we might wish to explore. Once in the Sporting Goods section, we may have another conversation with a floor manager about which is the best set of soccer goalkeeping gloves.

The importance of this conversation is that it is an example of what is often referred to as *exploratory search*. There is a tendency to focus the search team’s efforts on optimizing the delivery of specific documents in search results (or as best bets) because the information architecture of the intranet or document management system is not optimal. Exploratory search, where there is a strong requirement for the search system to support a dialog with the user, is much more challenging to develop, test and implement. Search really becomes useful when it makes it possible to go on a journey through the information resources of an organization, starting with perhaps a vague idea of the initial query and progressively refining the search query, or even starting all over again, on the basis of the information found.

The challenge with search, as is the case for the staff of the Information Desk, is that all users are different, with their own individual perceptions of what would make a good birthday present and what would represent value for money. In the business environment, the challenge is to find a way of meeting the individual expectations of each staff member without having to provide individual search applications. Indeed, the aim is to make them think that it does actually work just the way they want it to.

## Search Must Be Managed

For over a decade, I have been providing consulting services in management of intranets, and one of the most common issues is who should be taking responsibility for intranet development and operation in a company. An intranet, like search, is a very high-touch application, with most, if not all, employees using the intranet every day. The information on an intranet will be authored by most departments in the company, but clearly the people managing the application need to report to a manager who has the budget to support the intranet. The end result is that an intranet can be owned by corporate communications, HR, IT, or even marketing on the basis that an intranet is just another website.

In the final analysis, it should not matter who owns search, and the same situation applies to an intranet. Both should be managed within an overall information management policy and an information management strategy, but very rarely are. Some years ago, I went to run an intranet workshop for a major UK organization for which the effective management of information was probably its main competitive advantage. When I arrived, I noticed that all the cars were reverse parked, and it looked very neat and tidy. It transpired that the organization was concerned about safety and at the end of the day did not want staff reversing out of a parking space and either crashing into another car or staff walking to their cars. The parking policy was published on the intranet and at the reception desk, and it was made clear that a very dim view would be taken if the policy was not followed.

However, this organization did not have any policies about the management of information, so almost every document was written in a different format, often with no owner or even a date on the document. The quality of the search experience is directly related to the quality of the content. The old adage of garbage in, garbage out (GIGO) applies to search more than any other application. Someone has to take responsibility for information quality within an overall information management strategy. This is ideally written around an information life cycle, of which the following is just one example. The use of the term *document* is just a convenience and could be any item of information, from a personal profile to a video file.

The following is an example of an information life cycle:

*Phase 1: Create*
: This is the process of creating documents in a way that enables the document to progress through the stages of the information life cycle. These might include establishing document categories, writing good titles, and adding metadata. There could also be a quality assurance process.

*Phase 2: Store*
: There are many places that documents can be stored, including local and shared drives, document management applications, Lotus Notes applications, and intra-

nets. A set of criteria needs to be established so that employees know where documents should be stored so that they can be located and accessed by any employee with permission to do so.

*Phase 3: Authenticate access*

One of the differences between website search and enterprise search is that only certain groups of employees are able to see specific items of information. In addition, in some sectors, governments impose export license controls that mean access to corporate information repositories while outside the country could be restricted. Managing access permissions is a challenge for search managers.

*Phase 4: Discover*

Information can be found by searching through repositories, browsing through folder structures and intranet navigation, and through alerting services such as wikis and blogs. Each has a role to play in the discovery process. The process can be facilitated by good usability and the design of intuitive lists.

*Phase 5: Use*

The employees need to be able to feel confident in the quality of the information they are using, and that the information is valid for the particular use to which it will be put.

*Phase 6: Control*

This is not the same as access authentication. There could be implications from Freedom of Information or Data Protection legislation on the use that can be made of the information.

*Phase 7: Share*

The employee needs to be certain that the information can be shared internally, and if required, with third parties and with the public. Users of these documents have to be confident that the information they contain can be trusted to be reliable, and that if needed, the documents are available in a number of different languages.

*Phase 8: Review*

As documents are shared, others may have views on the accuracy and value of the document. A system must be established for undertaking the review process, and if needed, creating a new version of the document. A possible decision could be that the document is disposed of to prevent inadvertent use at some time in the future.

*Phase 9: Record*

Some documents will need to be retained in a secure environment for an agreed on period of time. Details of retention periods need to be established, and must take into account legal and regulatory requirements, as well as product and service lifetimes.

*Phase 10: Dispose*

Disposal is the final stage of the information life cycle. At this point, the document has no further value to the company and can be deleted from all systems without any risk to the future integrity of the company.

If the discover phase is broken in any way, then the information that has been laboriously created and stored cannot be used and cannot be shared. It has become invisible to people who could benefit from it. If you take the view that much of an organization's knowledge is also in a documented form, this knowledge also cannot be used and shared.

The review phase is also important, as it is this process that maintains the quality of the information, perhaps even enhancing it with additional metadata in the light of a change in business direction and/or a review of search logs and user requirements.

# Why Search Is Important

The biggest single challenge that any search manager faces is making a business case for a level of investment in search that is appropriate to the requirements of the company. The process of making a business case is covered in depth in Chapter 9, but for now, the following subsections will take a brief look at the many business benefits of good enterprise search.

## Capitalizing on Information Investment

Every day, most employees will have spent time on creating information—everything from writing a business plan, sending an email, or reporting on a visit to a customer. The process of creation may well be of the order of an hour a day, or 12% of the working year. If this information cannot be found and used by other employees, then that time has been wasted twice over, as other employees may have had to create the information all over again. There is also information from external sources, such as market research reports, that has been purchased and will have a company-wide value beyond the original purchaser.

## Reacting to Business Opportunities

At a time when business growth is static, finding new business opportunities is of the highest importance. When an opportunity does arise, the speed with which the company can find examples of relevant experience or size can be the difference between winning the business and being a poor second.

If a business opportunity arrived on your desk today, how quickly could you respond with a proposal that had low risk and a good financial margin? An enterprise search application can reduce the research time from days to hours, if not minutes, making the best use of staff expertise.

## Knowledge and Expertise Discovery

It is important not to focus just on information—knowledge must also be considered. Knowledge cannot be written down, as it is context specific and changes day by day as new knowledge is gained. Typically, companies have employee turnover rates of 10% a year. In a company with 5,000 employees, this means that, on average, every working day two people arrive at the company to build their careers and enable the company to meet its objectives.

How certain are you that these employees will be able to track down people with the expertise and experience they may need to make an immediate and effective contribution? Enterprise search can play an important role in finding them, though this is by no means as easy as many vendors would have you believe.

## Bringing New Staff on Board More Quickly

New employees want to make a positive contribution as quickly as possible. They do not have the time or the inclination to work through the navigation of the intranet or the folder structure in the document management system, nor do they know the names of people who might be useful to them as they begin work.

Employees taking on new roles and responsibilities will be in just the same position, but possibly with a greater need to get up to speed as the expectation will be that they know exactly where all the relevant information will be located. If only!

## Speeding Up the Process of Acquisition

One of the most significant benefits of enterprise search is that once the deal has been done, employees in the acquired company need to have immediate access to the information resources and employee knowledge base of their new employer. In addition, the business case for the acquisition will have been based on the skills and knowledge that the acquired business will bring.

In those crucial early days, enterprise search can make a substantial contribution to the rapid and successful integration of the acquired company by quickly indexing the information resources of the acquired company.

## Supporting Mobile Workers

Many of these employees will be working outside of the office, dealing with customers, prospects, and suppliers. They will need information as the meeting is taking place to confirm the details of a product or the name of a subject matter expert in the company.

Mobile users will use enterprise search on their smartphones or tablets to find information on a close-to-instantaneous basis and close the deal.

## Reducing Workplace Stress

Routine tasks are rarely routine. New policies emerge and new forms are devised to capture information. Of course, what is a routine task for a long-serving employee is not routine for someone new to the company or the role. In both cases, there never seems to be enough time to complete the tasks.

Embedding search into a task can ensure that as the task is undertaken, the most recent information is presented to the employee by the enterprise search application working in the background as a search-based application.

# Summary

All the evidence suggests that organizations are ill prepared for the rate of growth of information they are experiencing. Because information is not seen as a business asset, with an associated information management strategy, organizations have no view on the scale of the problem. As a result, no one is taking ownership of the problem because it's not being perceived as such.

If information cannot be found, then the effort and investment in creating and storing it are wasted. The work may need to be duplicated (if there is time to do so) or a decision made on what can be found, even if it does not represent the best of what the organization has in terms not only of information but also of knowledge.

Seeing enterprise search as the quest for a single search application that can index all organizational information is not the solution. Enterprise search is about creating a managed search environment that enables employees to find the information they need to achieve organizational and/or personal objectives. There will be many different business cases that need to be addressed within this managed search environment, each contributing to the overall investment case.

## Further Reading

At the end of this book, Appendix B sets out a core library of books and reports on information retrieval and enterprise search. Most of the individual chapters also include a "Further Reading" section listing out more specialized sources of information that have either been referred to in the chapter or will provide additional information and guidance.

A comprehensive list of books on search technology and implementation can be found at *http://www.intranetfocus.com/enterprise-search/books-and-reports*, and a list of blogs on search can be found at *http://www.intranetfocus.com/enterprise-search/enterprise-search-blogs*.

Association for Information and Image Management (AIIM), "Search and Discovery —Exploiting Knowledge, Minimizing Risk", September 12, 2014.

Chun Wei Choo, "Information Culture and Organizational Effectiveness," *International Journal of Information Management* 33 (2013): 775–779.

Paul H. Cleverley, Simon Burnett, and Laura Muir, "Exploratory Information Searching in the Enterprise: A Study of User Satisfaction and Task Performance," *Journal of the Association for Information Science and Technology*. Published online in Wiley Online Library (wileyonlinelibrary.com, 2015). DOI: 10.1002/asi.23595

Stephen Dale, "Content Curation: The Future of Relevance," *Business Information Review* 31:4 (December 2014): 199–205.

Christopher Durugbo, Ashutosh Tiwari, and Jeffrey R. Alcock, "Modelling Information Flows for Organisations: A Review of Approaches and Future Challenges," *International Journal of Information Management* 33 (2013): 597–610.

Deloitte Access Economics, "The Economic Impact of the Information Glut,", 2011.

Findwise, "Enterprise Search and Findability Survey 2014", April 2014.

Nigel Ford, Introduction to Information Behaviour, Facet Publishing, 2015.

Gartner Group, "2013 Strategic Road Map for Enterprise Information Management", November 2013.

Information Governance Initiative (IGI), "Information Governance Initiative Annual Report 2014," August 2014.

Jane McConnell, "The Organization in the Digital Age", 2015.

Peter Morville, *Ambient Findability* (Sebastopol, CA: O'Reilly, 2005).

PricewaterhouseCoopers (PWC), "PwC's Global Data & Analytics Survey 2014: Big Decisions", 2014.

Andrew Thatcher, Ana C. Vasconcelos, and David Ellis, "An Investigation into the Impact of Information Behaviour on Information Failure: The Fukushima Daiichi Nuclear Power Disaster," *International Journal of Information Management* 35 (2013): 57–63.

Zi Yang, Ying Li, James Cai, and Eric Nyberg, "QUADS: Question Answering for Decision Support," July 2014.

# CHAPTER 2
# Benefits and Challenges

Most people think that search is easy. All you have to do is type a word or two into the search box on Google or Bing. In a fraction of a second, thousands, if not millions, of results are ready to review. You don't know and don't care about how this was accomplished, and for searching the Internet, that's acceptable. Even if you knew all about PageRank, BigTable, Markov chains, and the teleportation matrix, it would be of no value in using Google, and the situation is similar with Bing. The nice thing about searching the Web is that we are easily satisfied. Even if you don't find quite what you are looking for, you will find something close enough to be useful and forget about the initial disappointment.

Enterprise search is much more challenging. From the evidence presented in Chapter 1, it is clear that there is a significant dissatisfaction with enterprise search applications. One of the reasons for this is the height of the satisfaction barrier. If you are looking for a specific document or specific information and cannot find it, then your satisfaction is zero. Finding something roughly similar is rarely good enough to risk your career on.

When it comes to enterprise search, it really does make a lot of sense to know something about how search works. However, before we start to look inside the technology (we'll cover this in Chapter 3), we will look at some typical experiences with enterprise search. Then, in Chapter 10, we'll consider some ways in which we can define the search requirements of our employees.

## A Day at the Office

Over the first coffee of the day, you've been looking through the overnight emails, and found one from your manager asking you to prepare the section on corporate social responsibility for the annual report. As your company has acquired Advanced Energy Corporation and Building Benchmark Services in the course of the last 12 months, your manager has suggested it would be a good idea to check out what their approach has been to corporate social responsibility in case there are lessons to be learned.

As this is the first time you have been asked to write this section, your initial action is to see what can be found on Google just to make sure you know exactly what corporate social responsibility is all about. In just over 0.2 seconds, Google comes back with over 10 million results. Impressive! The first result is from Wikipedia listing all the various different terms used for corporate social responsibility.

Next, you turn to the intranet search box and enter *corporate social responsibility*. The initial response is to ask you whether you want to search the intranet, the document management system, or all sources. Your immediate reaction is to wonder why you have to know where information is before you search for it, and then to wonder why anyone would choose to search in a specific application rather than all the applications that the company has invested in.

For now, you decide to search in all the applications. After perhaps 15–20 seconds (much slower than Google when it was searching the world!), you get some results back and are faced with one or more of the following scenarios:

*There Are 3,245 Results*

Your first reaction will be that you do not have the time to look through 3,245 results, but it won't be a problem, as you expect all the relevant results will appear in the first couple pages of results. However, as you look through the initial set of results, the titles are often meaningless, such as "Doc1" or "6635RTS." Looking at the summaries of the results, you then realize that the words *corporate*, *social*, and *responsibility* are highlighted, but in many cases not as a phrase, and it dawns on you that you have been presented with results that contain any of the three words —you haven't narrowed it down to documents concerned with only corporate social responsibility.

*There Are 9 Results*

Next, you search for "corporate social responsibility"—but you're surprised by what turns up. Surely there must be more than nine results? Where are the rest of them? Then you think back to the Wikipedia entry and remember that *corporate social responsibility* is often shortened to the acronym *CSR*, and none of the results show any reference to CSR. Now you are faced with the problem of how to search for documents that contain either the phrase *corporate social responsibility* or *CSR*. This is starting to be more difficult than you imagined. You move to the Advanced Search option and start again with a new search: "corporate search responsibility" CSR.

*There Are 230 Results*

The problem now is that many of the results seem to be about construction regulations. This is when you realize that CSR also stands for Construction Safety Register, which is an important for the Construction Division of your company but has no relationship at all to corporate social responsibility. You have no idea how to narrow down results for CSR to include only those about corporate social responsbility.

*There Are 400 Results*

However, most of these results have different versions of the same document. Preparing the official corporate social responsibility document usually means going through many versions before the final document is approved. Because you are searching the document management application as well as the intranet, all these versions are now visible, and it is impossible to differentiate the final versions from the interim versions.

*There Are 425 Results*

None of the results seem to be previous corporate social responsibility reports. What has happened is that the report has been added to the company's website, but no one has been tasked with putting it up on the intranet, as there is a link on the intranet to the corporate website in case anyone wants to look at it. Of course, everyone knows this, but the search application does not.

*There Are 390 Results*

But none of these results seem to be from Advanced Energy Corporation and very few are from Building Benchmark Services. Looking in more detail at the results that have BBS somewhere in the URL, you see that *corporate* has been highlighted but the adjacent word is *citizenship*, and you begin to realize that Building Benchmark Services has a section in its annual report titled "Corporate Citizenship" rather than "Corporate Social Responsibility." Now you have to think about how to create a new search query to bring in the concept of citizenship. As for Advanced Energy, it is probable that for some reason the search application is not indexing the Advanced Energy application that contains the work the company undertook on corporate social responsibility. Now you will have to track down who was responsible for this activity, hoping that they are still employed by your company.

It has been a good day—you think you've found all the relevant information. The search application gave you 20 really useful documents from the 83 it listed out, including the statements from Advanced Energy Corporation and Building Benchmark Services. You spend the rest of the day writing up a statement about your company's approach to corporate social responsibility and email it to your manager. You take an early train home.

On the journey home, your manager calls you and wants to know why the outcomes of the project on CSR that the Project Prospero team has been working on for the last few months is not included in your analysis. You promise to check and log on to the corporate desktop through your iPad. Rerunning the search fails to disclose anything about a report from Project Prospero, and indeed there is nothing about Project Prospero.

The next morning you call a friend in the Project Management Office and through her track down Simon, the project manager. That is when you discover that he and a group from Legal have been working on CSR issues using a TeamSite in SharePoint 2013. This can only be accessed by people who are part of the group, and because you are not a member, the search application did not show you documents from the TeamSite. Simon is more than willing to add you to the TeamSite, and your manager accepts your explanation.

But you resolve not to put your trust in the search application again. Ever!

This example illustrates the typical problems that arise with enterprise search. Managing them requires a combination of high-quality content, search technology selected with care, and a team of people supporting the technology and users.

## Recall, Precision, and Relevance

Recall, precision, and relevance appear many times in this book, so it is time for some definitions. The first two seem easy to define:

- *Recall* is a measure of the number of relevant documents returned as a percentage of all the relevant documents in a collection.
- *Precision* is a measure of the extent to which the set of documents returned from a search are relevant.

However, the problem with the definitions for these two terms is that they are defined in terms of relevance, which is a personal judgment on the value of a piece of information. In an ideal world, any result from a search should produce a list of all, and only, relevant documents. This is impossible because there is no way of knowing, at least outside of a test collection, how many relevant documents are in a given collection. Another problem is that relevance is defined in absolute terms; either a document is relevant or it is not. In reality, things are fuzzy.

Search vendors often talk about "accurate" results. This is nonsense. Information can be accurate in terms of totally representing a known and agreed fact—for example, that Horsham (where I live) is a town in West Sussex. Except that there is a town called Horsham just outside of Philadelphia and another in the state of Victoria, Australia! If a representative from a search vendor tells you that the results from her company's search software are more accurate than that of the competitors, always ask for a definition of "accurate" and a demonstration! Remember that any definition of "accurate" has to define a scale of conformance.

## Why Can't Our Search Be Like Google?

Many companies that are dissatisfied with their current enterprise search application want to know why it is not as good as Google. This is a very good question, and deserves to be answered. The slick answer is that if they allocated around 13,000 engineers to supporting enterprise search in their company, then it probably would be as good as Google.

The technology behind Google Search is extremely complex and mostly hidden from view. The technology behind Microsoft Bing and other web search sites is equally complex and even more hidden from view.

The story starts in 1997 when Jon Kleinberg, a research scientist at the IBM Almaden Laboratories in Silicon Valley, started to look at how the hyperlinks between pages and sites on the World Wide Web could be used to enhance search performance. The algorithms were powerful, but as IBM was not in the web search business, the outcomes of the research were not of direct value to IBM (they were, however, taken up to some extent by Yahoo!).

At around the same time, Sergey Brin and Larry Page were working on what would become Google. They announced the outcomes of their work at a conference in Australia in 1998. The underlying principle of Google's PageRank algorithm is that if a web page is important, then it is pointed to by other important pages. This needs to be read carefully. It is not just the number of links, but the links from important pages, and that means a great deal of analysis must be performed on the results from the web crawls. This concept of reverse citation was not invented by Brin and Page, but comes mainly from the work of Eugene Garfield and his Science Citation Index, which he developed in the late 1950s. The mathematics of PageRank is very complex, but the computational effort is perhaps greater and led Google to develop its BigTable. It was implemented in 2005, having taken seven man-years of research effort.

The combination of PageRank and BigTable is only part of the story. Google is constantly trying to improve search performance and has a team of over 13,000 staff in research and development, 40% of the total staff complement.

The end result is a very powerful web search capability, but in using Google we sometimes forget just how much work we have to do to find the information we are looking for. Sometimes we strike lucky and the information is on the first page or two of results. On other occasions, we may spend a considerable amount of time following false leads. It can be very instructive to start a stopwatch at the beginning of an important Google search to note just how long it took to reach a satisfactory conclusion.

As I have noted in the Preface to this book, Google and all of the other web search services benefit from the work that all website owners undertake to ensure that content on their site can be found by the robot crawlers and indexed. The effort (and often consulting fees) that goes into search engine optimization may well be larger than for internal enterprise search applications. Web teams worry about the quality of the content, and may well have guidelines on titles and on the formatting of website content. Rarely are there similar guidelines for internal content.

The clarity of the search box is admirable, but at the outset of a search, Google does not offer any suggestions about whether it might be better to use one of its specialized search applications. If you search Google for *ferrocene*, a complex organo-metallic compound that I spent some time at university working with, then there is no indication at the beginning of a list of over 500,000 search results that Google also offers Google Scholar, offering 150,000 results from the research literature. One of ongoing debates in enterprise search is the extent to which an enterprise search application can be made as intuitive as Google. The question I would ask is whether Google is really intuitive. There are a number of books on how to get the best out of Google Search, most of them running to several hundred pages.

If the senior management team remains adamant in wanting the enterprise application to be as good as Google, then the actions they need to take include:

- Include a requirement in the job description of all employees that they create information to agreed organization standards, and the time that they spend is seen as a core element of their working day.
- Also included should be the time to review, revise, and if appropriate, replace the document if it is no longer current or is incorrect in any way.
- Set out and monitor compliance with standards on content quality, especially regarding the quality and consistency of titles and the way in which the information is structured. Particular attention should be paid to ensuring that words that could be search query terms are included.
- Each document should refer to as many other relevant documents as possible so that the target document can be placed in context and other documents consulted without the need to conduct a further search.
- The search applications should be supported by a team of people with specific experience in managing search applications, have a strong background in information retrieval, a deep understanding of the organization, and the capability to

conduct research projects across the organization in the process of enhancing search quality.

- A member of the board of directors should have search as a specific high-priority objective, reporting at each board meeting about the extent to which organizational risks are being ameliorated or increased through the performance of the search applications.
- An annual survey of employees should be conducted about the extent to which the search applications are meeting staff requirements, with the performance bonus of the director linked directly to agreed annual targets for search improvement.

This, of course, is "search nirvana," but as you continue reading you will see that the advice given throughout this book matches this list.

## Web Search Gives You Options

Whenever you carry out a Google search, there are always other options available, and you usually have a Plan B if you cannot find what you are looking for. Even if you find information on Google, you will probably do a quick evaluation to see if you trust it, taking into account the website, the age of the document, the formatting (a PDF always looks more impressive than a Word document) and perhaps the organization publishing the information.

In the case of enterprise search, you have nowhere else to search. If you can't find a document, you don't know if it's because it doesn't exist or because for some technical reason the search application cannot find it. In addition, security management ensures that users only gain access to documents that they have permission to see, though the way in which security management is applied is rarely transparent.

Even Google cannot ensure that you are only presented with high-quality information. Just because there are a lot of links to a page from an important site does not mean that the information is of high quality. Quality, like relevance, is relative and personal. If you have a document of poor quality but cannot find anything better, then miraculously the quality of that document improves considerably.

## Information Quality

Throughout this book, there will be references to the fact that poor search performance is a result of poor information quality. It is quite difficult to define *information quality*.

Typical problems include the following:

*Poor titles*

PowerPoint files are frequent culprits, with titles such as *Project Prospero – update*, with no indication of what the scope is of Project Prospero or the date of the presentation. In most cases, the author pays little attention to the title because the document is being written for a known audience of colleagues. No consideration is made of the potential use of the information in the document by colleagues who have no idea what "Next steps in the TDG deployment plan" is about.

*No author information*

Many enterprise searches are about people. Now that you know Simon has been involved in corporate social responsibility work, you may want to find other documents that he has prepared. Simon is a member of the Governance Group, but none of the presentations include author information—the Governance Group knows exactly who prepared and gave the presentation, so why add a name? It is the same problem as the case for titles.

*Inconsistent metadata*

Good quality and consistent metadata is essential in achieving good quality search. In general, searching a document management application is a delight because documents are set within folders and so acquire the basic folder metadata (e.g., *FY2015/16*) automatically. A document management system will often refuse to let users file a document without adding certain metadata, such as their name, department, job title, and some subject terms. Web content added through a web CMS or documents added through SharePoint are a different matter.

*Ambiguous date formats*

One of the most challenging metadata problems in enterprise search is the requirement to carry out date and date-range specific searches. The PowerPoint file on Project Prospero was finalized in June 2011, but one of the search results listed a 2012 version. This is because Simon wanted to change a couple of words in the presentation and then filed it away. This adds a 2012 date to the file metadata, but for all practical purposes, the document is identical to the 2011 version. The management of dates is something that is far more important in enterprise search than in web search, because we usually have good memory mechanisms for maintaining a chronology. We might recall that there was an excellent presentation on Product A at the 2013 sales conference in Miami, so even without knowing the title or the author, we can quickly search for the presentations from the conference and find the one that we remember.

Another challenge is that the date format for North America is month/day/year, but for most of the rest of the world it is day/month/year. Most search engines normalize the date to a common format so that it is possible to search for all documents published (or updated!) in, for example, March 2012. The complication arises when the results set out the date in the original date format.

In addition to the problems just described, you must also consider the following:

*Document structure*

It can also be helpful to ensure that documents have informative chapter headings and informative subheadings. Even if the search application is not giving these weight as metadata, remember that someone finding the document may want to scan through it at speed to find a particular piece of information.

*Language*

In the world of search, text is not about the meaning of words but the meaning of sentences and the meaning of sections of documents.

Even within English, the same word can have very different meanings. If an American asked me to slate a meeting, I'd know that I would need to set a date and perhaps the attendees and agenda. But if a Brit asked me to slate a meeting, I'd ask which meeting she wanted me to criticize. How can the same word have totally different meanings? The US usage is derived from a French word meaning "to splinter," which is what slate does when it is mined. The UK social usage is derived from an Old Norse word *sletta* meaning "to slap."

Understanding the meaning of social language is going to be increasingly important in the future as social media applications become widely adopted. The search application will need to be aware of acronyms, slang, and the use of shortened forms of names.

Searching for information in multiple languages is also going to be increasingly important, not only because of social media in a local language, but because companies are beginning to appreciate that the concept of English as a corporate language is not consistent with an ethical approach to employees and their cultural values. An even bigger challenge is searching for information in documents that have been written in the author's second or even third language. When speaking in a second or a third language, there is an opportunity for people to check that they have correctly understood one another. That will not be the case for a written document.

The problem becomes more challenging when searching video transcripts. Although the ability of voice-to-text conversion is improving very rapidly, the problem of how to render a person's name remains a challenge.

*Additional metadata concerns and taxonomies*

Metadata adds structure to unstructured information. I cannot think about metadata without recalling Henry Reed's poem, "Naming of Parts". The most well-known metadata schema is referred to as the Dublin Core, which defines 15 metadata elements for simple resource discovery: title, creator, subject and keywords, description, publisher, contributor, date, resource type, format, resource identifier, source, language, relation, coverage, and rights management.

Consistent metadata is very important in achieving good search performance. Findwise's "Enterprise Search and Findability Survey 2014" indicated that using metadata, a taxonomy, and a content life cycle all made a significant difference to search performance. It does matter in practice if an item of information is not tagged with every one of these 15 categories. A critical element of a search strategy is to decide which of these should be added. There is an important element of balance here because the workload associated with in-depth tagging is probably beyond any enterprise team. Decisions must be made, and tested out, about which elements would be of particular value to search.

There is an excellent introduction to metadata and taxonomies published by Pebble Road, a consulting company based in Singapore. Also based in Singapore is Patricke Lambe, the author of *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*.

## Search Maturity

There is no technical solution to the problems of poor information quality. Unless these quality issues are addressed head-on, spending more money and time on introducing a new search technology will have no impact on search satisfaction.

The 2014 Findwise survey lists the main obstacles that users found in trying to find information:

- Lack of appropriate metatags
- Search results not relevant
- Don't know where to look
- Not all content sources are searchable
- Information is outdated
- Search skills are lacking
- Poor navigation functionality
- Access restrictions to content that could be of value
- Search process takes too long

None of these will be solved by a new search application.

AIIM's "Search and Discovery—Exploiting Knowledge, Minimizing Risk" provides a useful approach to assessing search maturity, suggesting that these are the key elements:

- An agreed corporate search strategy
- A specific budget for search
- An acknowledged owner for search-related issues
- Dedicated and trained staff supporting search
- An agreed corporate taxonomy or vocabulary of terms
- A metadata standard across different repositories

The survey indicated that almost 60% of respondents had none of these elements. The 2014 Findwise survey showed that organizations with a taxonomy, metadata standards, an information management life cycle policy, and a search strategy had levels of search satisfaction 10–15 percentage points higher than where these fundamental elements of search solutions were not present.

The challenge for IT departments owning search applications is that they rarely have staff with the information and library science skills needed to create taxonomies and metadata schemes. Over the years, corporate libraries were closed down as not being relevant to the needs of the business. Now that businesses need the skills of librarians and information specialists, they are in short supply.

## Summary

The speed and performance of web search with Google and Bing set levels of expectation for enterprise search that cannot be met. It is not just about the technology but about the categories of content that are published on the Web, and that even something only marginally close to what we were looking for may be adequate. Enterprise search is also about searching for information in many different applications, not just in different servers, and that adds significantly to the scale of the problem.

Nevertheless, there are solutions available, but the resources and skills of the search support team are perhaps even more important as a success factor than the technology.

## Further Reading

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search, Second Edition* (Boston: Addison-Wesley, 2011).

Marcia J. Bates (editor), *Understanding Information Retrieval Systems: Management, Types, and Standards* (Boca Raton, FL: Auerbach Publications, 2011).

G. G. Chowdhury, *Introduction to Modern Information Retrieval* (London: Facet Publishing, 2010).

Massimo Franceschet, “Page Rank: Standing on the Shoulders of Giants,” *Communications of the ACM* 54:6 (June 2011): 92–101.

Patrick Lambe, *Organizing Knowledge: Taxonomies, Knowledge and Organisational Effectiveness* (Cambridge, UK: Chandos Publishing, 2007).

Pebble Road, *Organizing Digital Information for Others* (Singapore: Pebble Road, 2012).

Ian Ruthven and Diane Kelly (Editors), *Interactive Information Seeking, Behaviour and Retrieval* (London, Facet Publishing, 2011).

Stephen Robertson, “Understanding Inverse Document Frequency: On Theoretical Arguments for IDF,” *Journal of Documentation* 60:5 (2004): 503–520.

Stephen Robertson and Karen Sparck Jones, “Simple Proven Approaches to Text Retrieval,” University of Cambridge Computer Laboratory Technical Report 356 (includes amendments up to 2006), 1994.

Mark Sanderson and W. Bruce Croft, “The History of Information Retrieval Research,” *Proceedings of the IEEE (Special Centennial Issue)* 100 (2012): 1444–1451.

Ali Shiri, *Powering Search: The Role of Thesauri in New Information Environments* (Medford, NJ: Information Today, Inc., 2012).

# CHAPTER 3
# Search Technology

There are three fundamental components of any search application. First, an index of the information in a set of documents is created. The words in a search query then need to be matched against this index, and all of the documents that match the words in the query must be compiled into a list. Finally, this list of results is ranked in descending order of relevance. To any one unfamiliar with the technology of search, it all seems so simple. But there is much more involved in the mechanics behind the search process. The reality is that any search application consists of a set of modules, each of which carries out a specific task in the search process. Some of these modules may be brought in by the search vendor, and others will be developed internally. The same is true of open source software development.

Users should not have to know anything about search technology to be able to use it effectively, but understanding the elements of search technology is important in the selection, testing, and management of a search application. This is because one or more of these modules may be especially important in meeting a specific user requirement. It is very much a question of the whole only being as strong as the weakest link in the chain. If there are some limitations in the way that content is indexed, then it does not matter how elegant the user interface looks—it could be that information critical to the operations of the organization remains invisible.

To risk a generalization, most enterprise applications are built around a structured database with highly structured data collected through a well-established set of business processes. The data are collected using forms, data entry screens, or from sensors of various types. It is usually relatively easy to track down where a problem has arisen. With search, the issues are all about fuzziness and approximations, and tracking down why a particular document has not appeared high up on the list of relevant documents for a specific search can be very difficult to work out. Any advanced book on search will look more like an advanced book on applied mathematics with a substantial amount of computational linguistics.

In this chapter, the building blocks of a search engine are described in sequence, focusing on what might be regarded as backend technologies. Chapter 4 looks at the way in which queries are managed and results presented. This is a somewhat artificial distinction for the purposes of structuring the book.

## Content Gathering

A search application gathers in the information to be indexed in a number of ways. It can crawl through web pages and file servers, applications can be set up in a way that any new information is automatically sent to the search application for indexing, and RSS feeds can also be indexed. Decisions must be made about which servers are going to be indexed and at what frequency. In an ideal world, it would be good to index information the moment that it is added to a server, but this is not practical for two reasons. The first is that many applications auto-save as a document is being created and without careful management, multiple versions of the same document may be indexed. The second is that crawling and indexing are bandwidth and processor intensive and both have cost implications.

All crawlers are not created equal. If information is missed or incorrectly passed back to the indexer, no amount of subsequent processing is going to find it. This is one of the major implementation issues, because the proof of concept might work well, but as the scope of the collections to be indexed increases, the chances of crawler failure increase substantially.

Another approach is to write a script that identifies when a new document is added to the server (or when an existing document is updated) and then push the new or updated document to the indexing engine of the search application for processing.

Once the search application has indexed a document and knows its location, it will assume that the location does not change and that the document is always accessible. However, locations do change, so work then needs to be carried out to make sure that the new location is recorded in the search application. The document may not always be accessible—for example, the server may fail in some way or be taken down for service. This is when companies discover holes in disaster recovery plans, as it may not be obvious to a local IT manager that business-critical information resides on a particular server.

A criterion for search performance is *freshness*, which is a measure of the time taken to update the index from the point at which the document was added to a repository. You might think that only news items need to be added to the index as quickly as possible, but it could be just as important to expedite adding project reports so that the expertise gained is available with minimum delay to an engineer who might be able to use the information from that project to make a very competitive bid for a new customer.

It is very difficult to make any back-of-the-envelope assumptions about server performance for search applications, especially in distributed server architectures. Although most large IT departments will have someone with a specialization in capacity planning, all the standard rules don't work with the types of indexes and processing being undertaken by a search application. Scaling up from small collections is also fraught with risks.

## Connectors

In enterprise installations, the information to be indexed resides on a number of different application servers. For example, the intranet may be on a single web server, but the people search application is on an Oracle HR application. Indexing the Oracle HR application may require the use of a connector. The best way to think of the role of a connector is to see it as a sophisticated travel plug. Just like a power adapter converter allows international travelers to use otherwise incompatible outlets, the connector enables the content of the Oracle database to be read by the search application. Most search applications will come with a number of connectors integrated into the software, but these will not cover every eventuality. Many search integration specialists, notably Findwise, Raytion, and Search Technologies, also offer connectors.

Connectors also tend to be version specific, so that a change in the version release of the Oracle HR application could require a new connector to be installed and tested. Connectors can be expensive to purchase and need constant attention to make sure that they are working correctly. They also connect applications developed by two different vendors, and tracking down where the problems lie if the connector seems not to be working correctly can be a difficult discussion that has to be managed by getting the appropriate experts from both vendors around the same table. It can all be very time consuming.

Although connector failures have the potential to significantly impact search performance, they can be very difficult to track down. Several months might pass before it becomes apparent that a particular server has not been indexed because of a configuration problem.

## Document Filters and Language Identification

Unfortunately, the information you need to index will very rarely be contained in plain text. Microsoft Office documents might be stored in many different versions of Office, and the Office suite also includes Excel spreadsheets, PowerPoint presentation files, and perhaps Visio process diagrams. PDFs and Lotus Notes databases also need to be considered. All PDFs are not created equal, and there are probably more than

20 variants depending on the software supplier. Things get really interesting when working with PDFs that have very high image content (a briefing note from an investment house comes to mind), and deconstructing these to create index terms can prove to be quite difficult.

Each document format must be reduced to plain text, and any non-content control characters (e.g., the code that produces justified text in Microsoft Office) need to be removed. *Plain text* is actually a misnomer because although documents are often referred to as *unstructured information*, a document does contain a great deal of structure, such as a title, date, author, summary, contents, page numbers, and an index. Search engines are able to place a different weight on words that appear in a title or executive summary versus words appearing in the body of the document, so it is important to be able to retain core elements of the document structure.

It's important to note that some document formats are more difficult to manage than others. An example would be a PDF version of an Excel spreadsheet where it is important to retain the row and column information, without which the document might just be converted into a set of cells without any context. Microsoft Visio and Microsoft Project files can also be a challenge. If it is important to be able to find people working on projects by indexing Microsoft Project files, this requirement needs to be highlighted in the early stages of defining the scope of the search application.

Finally, not all the content will be in English. Even in the United Kingdom it could also be in Welsh. The ways in which a search application will detect the language of the information are quite complex, but at least the search application does have a significant amount of content to work with. The problem is much more challenging with queries. Is a query about *sante* (the word for *health* in French) related to someone looking for articles in French on the subject of health, or could the user be looking for something else entirely? For example, it might be a misspelled attempt to search for information about the city of Santa Fe or even about Santa Claus. In addition, confusion may arise because many international organizations have two different acronyms for their titles, notably the Organisation for Economic Co-operation and Development (OECD), which in French is l'Organisation de Coopération et de Développement Économiques (OCDE).

# Parsing and Tokenizing

At this point in the process, the complexities of language must be addressed by the search application. The text file output from document conversion now must be converted into tokens by a process known as either parsing or tokenizing. In most cases, each word is a token, but hyphens, apostrophes, and capital letters are just some of the problems a search application has to deal with. Is the name *McDonald* the same as *MacDonald*? From the point of view of the person concerned, certainly not, but a user may only have heard the name mentioned in a conversation and might not know how the person's name is spelled. The search application should be able to offer the user the option of using either a specific spelling, reminding the user that there are two different spellings or allowing the user to query for *M?cDonald* in which *?* is a so-called wildcard that could be an *a* or could be nothing at all. The parsing process has to be good enough to ensure that any valid query term can be matched to its position in the search application index.

The chances of doing this to perfection at the time the search application is installed are very low indeed, and this is just one of the reasons why there has to be a search support team looking through the search logs to identify where there seems to have been a mismatch between a query and the parsing process.

Other examples include:

- Hyphenated forms of words—for example, to ensure that a search for *oil-free compressors* also finds documents in which the word has been written as *oilfree*
- Numbers, especially in cases when they may refer to products
- Periods in abbreviations—for example, OCED and O.C.E.D. should be recognized as being semantically one and the same
- Capitalized words, where *apple* and *Apple* have different meanings, though, of course, if *apple* is the first word in a sentence, then it will be *Apple*

The critical issue about tokenizing is that it has to match the query transformation. If *I.B.M.* is converted into *IBM*, but a query on *I.B.M* does not transform to *IBM*, then no match will be made. One of the tasks at both implementation and later is to identify acronyms and other terms where the tokenizing rules and the query transformation rules may not meet in the middle.

## Stop Words

In all languages, there are many words that do not have any value as subject terms, such as *the*, *of*, *for*, and *about* in English, and these can be stripped out of the index, though the storage savings is not significant. Search application vendors either generate their own dictionaries of stop words or buy them from a specialist supplier. Care needs to be taken that what seems to be a word of no value, such as *next*, is considered in the context of the business. In the United Kingdom, there is a store group called Next which uses *next* as its brand name. A stop list that removed this word would be a disaster for any company that did business with Next.

The classic example that illustrates why it's important to take care in crafting a stop word list is the phrase from Shakespeare's play *Hamlet*, "To be or not to be," which consists of words that would normally be defined as stop words.

In enterprise applications, it is very easy to overlook the use of special characters, particularly punctuation that is used for product numbers or project codes. As with all aspects of search, the way that the index is managed needs to match the way that the query is transformed or there will be no match.

## Stemming and Lemmatization

These are two slightly different approaches to group together words that have a common stem. A user searching for *ships* may also be interested in results on *shipping*, so the stemmer has to reduce both to *ships* but also tag the occurrence in the index so that when the result is displayed, the words are shown in their full form. The original work on stemming was carried out at the University of Cambridge by Dr. Martin Porter in 1979. He developed an algorithm that enables a computer to undertake the stemming process through a set of rules. He later developed a specific version of this algorithm for English, known both as Snowball and Porter2.

This algorithm, with some small modifications, works well for all Romance languages, but other languages (e.g., Finnish or Arabic) pose additional problems. Good stemming will help ease user frustration by reducing the number of results that seem to match the query term but have actually no relevance at all. The abundance in the English language of synonyms that have very different meanings does not help matters.

Porter's approach is based around a set of rules, and is sometimes described as affix-removal, but there is also a statistical approach that uses a very large corpus of text to derive a morphology for a language. The advantage of this approach is that it can be used for almost any language, although the challenges of stemming non-European languages are considerable.

Lemmatization involves removing inflectional endings from words, through the use of morphological analysis, in order to reduce words to their base or dictionary form, which is known as the lemma.

There is much discussion about which of the many approaches to stemming and lemmatization works best in providing good search effectiveness, but without any definitive outcomes. When selecting a search engine, there is no point in stipulating the stemming and lemmatization approaches that are required, but it is important to consider whether there are some common terms used in your organization that might give rise to frustration over seemingly relevant documents being presented which in fact are irrelevant. For example, in the United States, the term *gas* is used in favor of *petrol*, which makes life difficult for international oil companies that inevitably are also in the gas business.

It is also important to anticipate future search requirements—for example, the initial search implementation, which will almost certainly be on English language documents, might eventually need to be extended to other languages. Even if there is no immediate intention of searching for documents in Russian, there should be a comfort factor in knowing that the search engine can process Russian language content and that at the proof-of-concept stage, including Russian language content would be advisable.

## Dates

Date management is important in enterprise search, especially in multinational companies that have offices in both North America and Europe. ISO Standard 8601 sets out the sequence as yyyy–mm–dd, and time as hh-mm-ss using the 24-hour clock. Many enterprise searches involve looking for the most recent document, a document issued within a particular time period, or the first time a document was issued. The initial challenge is that many documents do not have fixed dates. A document can go through many versions, each of which has a different date and time, and the most recent may not be the latest released/approved version.

A very common problem is that normal practice in North America is to write the date in a mm-dd-yyyy or mm-dd-yy format, so that 5/3/12 is the third day of May 2012. In most of the rest of the world, that date representation would be interpreted as the fifth day of March 2012. Most search engines normalize dates to the ISO standard, but the issue then is how the date is displayed in the search results. Ideally, it should be as either dd-mmm-yy or mmm-dd-yy, both of which are unambiguous. This may seem a very small detail and not worth two paragraphs, but it can cause very major problems in finding a specific document containing time-dependent information, such as month-end sales results.

## Phrases

Many search queries are phrases. *Corporate social responsibility* was used as an example for this reason. Phrases are often tagged by the search application to identify if the words in a phrase are nouns, verbs, adjectives, or other parts of speech. Vendors build up these databases from working with customers and by buying phrase dictionaries. However, this approach can also slow down the query process, and there are a number of other approaches that can be used. Because the index knows the proximity of every word in a document to every other word, in theory, it can match the phrase just through a proximity analysis. This can result in a large index and attendant performance issues.

Phrases also bring us into the world of *n*-grams. Any sequence of two words is described as a bigram and three words as a trigram. *n*-grams are of fundamental importance in many aspects of the search process, especially in the management of search across multiple languages, but a detailed description of their roles is outside the scope of this book. The core books on information retrieval do all this in great detail.

## Processing Pipeline

A number of search vendors refer to this collection of content processing steps as the content pipeline or the processing pipeline. The concept of a processing pipeline originates from software engineering, where the output from one process becomes input of the succeeding process. The sales pitch tends to be around the speed and efficiency with which the content processing can be carried out because each step of the process has been optimized for the steps immediately before and after any given step.

In principle, this is good news, especially for implementations where there is a significant number of new documents added each day. An example might be transcripts of call center conversations. It does mean that upgrading or customizing one particular step may have some implications for other steps in the process.

## Building and Managing the Index

Search applications use an inverted index to store all the tokens and other metadata generated by text processing. In principle, an inverted index is the same as the index to a book, but instead of just giving a page number, it is capable of not only telling the reader that the word *intranet* is on page 23 but that it is the third word on line 4. Similarly, there could be an entry for *strategy* on page 23 as the fourth word on line 4, so there would be a strong possibility that the page, if not the document, is about intranet strategy. If the entry for *strategy* was on page 22 as the seventh word on line 5, then that possibility could be smaller. This is a very simple example, but it illustrates the point that search inverted indexes contain a significant amount of positional information.

There are two components of the index: a document-level index and a word-level index. The word-level index with its positional information is described in the previous paragraph. The document-level index is a count of the number of occurrences of the term in the document.

Search applications have to cope with a significant amount of processing. There needs to be very fast random access to any point in the index, and indeed to multiple points at the same time. The index is also a very dynamic database. The number of changes to the HR database even in a very large company will be a very small percentage of the total number of records held in the database. Adding new employees usually means that they are replacing employees who have left the company, though even in tough economic conditions there will also be some new employees. Overall, the rate of growth of the database will be fairly small.

That will not be the case with a search index. Each day, each employee might spend an hour a day creating content that will need to be indexed, and that excludes emails. All that information has to be processed and indexed on a timely basis. Performance management for a search application is very important indeed, and each vendor will have its own approaches to the ordering of the index, and the way in which new (or changed) information is incorporated into the index. Users are very intolerant of response delays, largely as a result of the investment that Google, Bing, and other search services have made in web search, so even a 20-second delay for returning a set of results can seem like a lifetime to a user.

In addition to containing a pointer to every word in every document, the index will also contain all the positional information and tagging about phrases. The result is that the size of an inverted index could be the same size, if not larger, than the total size of the repositories that have been indexed. However, the indexes are usually compressed to the point that typically they may be around 30%–50% of the size of the repositories. For search, it is important not only to compress the size of the index (as this will reduce memory requirements and can speed processing), but also to restore the index to present the results. As with so much of the technology of search, the difference between the search products available on the market is often down to the basic processes of text processing, indexing, query processing, and results delivery.

Throughout the working day, new content is being indexed. How will this be included in the index? Some vendors are able to rebuild the index very quickly indeed and then switch incoming queries to the most recent version of the index. Other vendors will build a temporary index that is then searched in parallel with the main index; the results set is then integrated into a single sequence. This approach can slow down the overall process of the search as the integration takes place.

If a document is deleted because it is no longer relevant, that will almost certainly not remove the content from the index. In one major insurance company, salaries of senior managers were posted to the intranet. Even though the document was removed, a search still revealed this confidential information. The process of deleting a content item from the search index varies between applications, but is often nowhere near as easy as indexing it in the first place.

## Security and Access Control Lists

In many organizations, certain information can only be seen by specific employees, by employees in particular roles (e.g., HR), or by employees dealing with specific customer accounts. A search application must be able to recognize these access limitations to avoid the confidentiality of information being compromised. Safeguards must be put in place to prevent unauthorized downloads of confidential documents, and their existence should be concealed completely from the search results except for authorized users.

These authorization permissions are managed through access control lists (ACLs) in an Active Directory environment. An ACL defines which files (documents, databases, videos, and any other content indexed by the search application) can be accessed by each individual employee. The complexity of managing ACLs in a corporate environment with several hundred applications is difficult enough, but scaling this to potentially millions of documents is an altogether more challenging requirement.

This reduces the complexity of the ACLs, but assumes that the organization is quickly able to identify and change group memberships should a member of staff leave the organization. In the time after someone has been given notice to leave, even if it is at the end of the day, the potential damage that could be caused in downloading confidential documents could be very significant.

Another approach is SAML, developed by the Security Services Technical Committee of the Organization for the Advancement of Structured Information Standards (OASIS). SAML is an XML-based framework for communicating user authentication, entitlement, and attribute information. The current version is SAML 2.0, which dates back to 2005, and is therefore a widely understood and very stable protocol.

Another aspect of access control in search is that there is very often considerable benefit in being able to restrict searches to specific collections of documents. For example, users might only want to see documents in a specific language or only look at SharePoint repositories because they know that the information they are looking for will be in a SharePoint server. All this granularity of access presents substantial management challenges, especially because a difficult situation will arise if users find out that there is content that they are restricted from but is accessible to a colleague doing the same job in a different subsidiary.

There are two basic approaches that can be used to manage access to information to which not all employees should have access.

Based on the authorization ACL, users are shown results only from those collections that they have permission to access. This is based on a collection management policy of putting limited access documents into specific collections. This approach has limited impact on retrieval speed, but only works for those situations in which controlled documents fall into relatively few categories and therefore relatively few collections. As the number of collections increases, the retrieval performance will decrease because of the amount of processing that is taking place prior to results display. This is often referred to as the early binding approach to security management

Users are asked to provide authorization credentials after submitting a search request. The search engine then filters the results to display only those for which the user has authorization. This works well for document-level security, but has some major performance issues when implemented. This is because the search application has to undertake a check of the document ACL for each document, which results in a substantial performance overhead. If the servers concerned are being heavily used, then the search application may have to time-out the request for a result and users will never know whether they have not seen a document on the basis of security or system performance. This is often referred to as the late-binding approach.

In theory, there is a third approach in which all results are presented to the user, but there is a link to confidential information which requires a user to provide additional authentication before the content is displayed. Apart from the fact that the user is now aware of the existence of confidential information to which they may not have access, users that do have access now have to provide additional authorization for each of the documents concerned, which can be a very tedious process. In practice, this is not an approach that has any benefits.

The secure access management challenges just outlined are complex enough, but become even more difficult to manage when there is a federated search environment and the authorization processes and ACLs are not uniform across all collections and/or search applications.

From the viewpoint of the search application, managing the delivery of information according to an ACL is quite straightforward, if somewhat processor intensive for very large repositories. The organization must be able to uniquely identify employees or groups of employees; add, modify, or remove identifiers at short notice; and have a rigorous information security classification model for all the content to be indexed by the search engine.

One security loophole that is often overlooked is that some search applications provide users with suggested query terms as they enter a query into the search box. These may not be security trimmed. A search for *redundancy* might come up with a suggestion for *redundancy strategy planning*, even though the document cannot be accessed by the user.

## Entity Extraction

The concept of entity extraction is to be able to use the search application to automatically identify personal names, locations, and other terms that can then be used as query terms without the need to manually index these terms. The technical term for this process is *named entity extraction*, and it analyzes not just individual words but also sequences of words to determine index terms that could be of value in responding to queries. For organizations using English, they are working with a language that has over one million words, a result of invasions and the scale (at one time!) of the British Empire. The result is a language full of synonyms and polysemes. Fortunately, words do not appear in isolation (other than in tables and charts!), so an analysis of a sentence will help substantially in determining the meaning of a word.

The mathematics of entity extraction is largely based on the mathematics of Markov models, which describe a process as a collection of states and transitions between states, each of which can be given a probability. Although an understanding of Markov models, hidden Markov models, and the Viterbi algorithm are not a requirement for a search support team, it does illustrate the extent to which search is based on mathematics. These and many related mathematical models will be used in different ways by each search vendor and will lead to subtle differences in search performance. These can only be assessed through careful testing at the proof-of-concept stage.

The following is a list of some of the typical entities that can be extracted:

- Credit card number
- Currency
- Date
- Distance
- Email
- Location
- Longitude/latitude
- Nationality
- Number
- Organization
- Person
- Phone number
- Postcode/zip code
- Time
- URL

This extraction can be accomplished in four different ways, though many search applications will use all three in a blended approach:

- Statistical models provide a means of recognizing never-seen-before names and providing good answers when words can have multiple meanings. Analyzing the correlation with the other words helps identify the correct context, such as deciding when the word *Paris* is used as the name of a person or a city.
- Telephone numbers and credit card numbers have standard formats, so as these are indexed, a set of rules can be used to determine the category of the entity. This could be extended to other entities, such as part numbers. For example, bd436678 might be part of a rule specifying that any character string starting with *bd* and a six-digit number is a part number.
- Dictionaries and gazetteers will support the extraction of places and groups of places, so that a search for *EU* will also offer a search for individual member states of the European Union.
- Specific terms can be defined by the organization.

In the case of the part number example, it may be advisable to allow for variations such as BD 436678, BD-436678, #BD436678, and even 436678BD.

Entity extraction functions must work consistently and predictably. If there are four different formats for product numbers, perhaps due to the acquisition of a new company, then users will expect the application to work consistently across all products. They do not want to have to remember that when searching for product numbers with spaces (e.g., BD 436678) that the space needs to be omitted when entering the designation in the query box.

## Graph Search

There is a substantial amount of interest at the present time in using graph search technology, mainly through the adoption of graph search by Facebook. The underlying objective of using graph search is to be able to capitalize on relationships, the basis of the concept of six degrees of separation. Ironically, relational databases are not very good at managing relationships that have not been precoded into the database. Graph search, which has its origins in the work of Leonhard Euler, a 19th-century Swiss mathematician, enables a user to undertake what seems to be almost a random walk through a collection of relationships without having to use a search query to do so.

Early in 2014, Microsoft announced Delve, the commercialization of which had long been named Project Oslo, as a graph search application in which the nodes of the graph were auto-populated by events (e.g., a meeting or conference) taking place within Office 365. The entry point is a person's name. You might know that Sam is going to be at the next meeting of a project team. Using his name you will have access to documents or people that he is willing to share with you, in particular other people he is working with. These people can be displayed on the user interface and perhaps one of them is new to the organization you are working for. You can extend your graph to find out about them and what they are working on. In a sense, it is discovery by walking around the virtual equivalent of an office.

Although most of the very large-scale graph databases are internally developed, there are an emerging number of open source options, notably Neo4j.

## IDOL and Decisiv

There are two proprietary search technologies that each need a paragraph of their own. IDOL is an acronym for Intelligent Data Operating Layer, and was developed by Dr. Michael Lynch and his colleagues at the University of Cambridge in the early 1990s as a way of being able to use adaptive pattern recognition, together with Bayesian probability, to improve the relevance of search results. The principles of probabilistic-based retrieval date back to work at the RAND Corporation in the mid-1950s. It was commercialized in the Autonomy search application that was released in 1996. The IDOL technology was incorporated into a number of other software applications acquired by Autonomy. The company was then acquired by HP in 2011. In addition to the IDOL index, the application also required a Dynamic Reasoning Engine, a Classification Server, and a User Agent Server.

Another innovative approach is based on the concept of topic models. Topic modelling algorithms seek to identify the relationship of topics within a document. The mathematical core of this approach is the latent Dirichlet allocation (LDA). A similar model is referred to as latent semantic indexing (LSI), but despite the similarity of the names of these models, they are different in a number of ways. LSI is noteworthy because in 1999 Thomas Hofmann (University of California, Berkeley), working with Jan Puzicha (University of Bonn), proposed probabilistic latent semantic indexing (also referred to as probabilistic latent semantic analysis, or PLSA). PLSA is now commercialized by Recommind, which was set up by Jan Puzicha in 2000. Hofmann went to Google and then in early 2015 founded a new data analysis company called 1PlusX.

Both of these retrieval technologies are proprietary. It is unlikely that HP will allow any other vendor to license the Autonomy software. Although the Recommind implantation of PLSA is covered by a patent, there is scope for other vendors to use a range of other approaches to topic modeling. One of the leading academics in this area is David Blei, a Professor in the Statistics and Computing Science Departments of Columbia University, New York, and the inventor of LDA.

## Summary

To deliver high-performance search, it is important for a search manager to understand at least the basic principles of search technology. Although poor search performance may well be a result of poor quality content, a knowledge of search technology can be very valuable in identifying where in the chain—from crawl to index—the problems may lie. This knowledge will be especially important in working on the development of open source search applications (Chapter 6).

## Further Reading

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search, Second Edition* (Boston: Addison-Wesley, 2011).

Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines* (Cambridge, MA: MIT Press, 2010).

G. G. Chowdhury, *Introduction to Modern Information Retrieval* (London: Facet Publishing, 2010).

W. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice* (Boston: Addison-Wesley, 2010).

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, *Introduction to Information Retrieval* (Cambridge, UK: Cambridge University Press, 2006).

Cristian Moral, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez, "A Survey of Stemming Algorithms in Information Retrieval," *Information Research* 19:1 (2014).

Ian Robinson, Jim Webber, and Emil Eifrem, *Graph Databases* (Sebastopol, CA: O'Reilly, 2014).

CHAPTER 4

# Query and Results Management

Chapter 3 described the technology as far as the index being created. The next two steps in the search journey are to find the documents that match a query and then to provide a list of the documents ranked in descending order of relevance. This is where the search technology becomes more visible. At the risk of an inappropriate generalization, all search applications work along the lines set out in Chapter 3. It is in the management of queries and results, and in the display of results (described in Chapter 12), that differences become more obvious.

## Query Management

Query management presents some substantial processing challenges. In the case of indexing, there is an enormous amount of information that the indexing process can use to "understand" the nature of the information content. Users then type in a single word and expect the search engine to undertake a mind meld and work out what the query is really about. The query processing stage has to be able to undertake four processes very quickly:

- Check for obvious spelling mistakes and offer suggestions for correct spellings
- Use stemming and lemmatization to develop a range of potential query terms
- Identify entities or phrases that may need to be clarified or expanded
- Apply some semantic analysis to gain an insight into the likely nature of the query

A good example of query management in action can be seen on the public website of Microsoft. Typing the word *SharePoint* will cause a drop-down list of key variants to appear, such as *SharePoint 2007* and *SharePoint 2010*, as well as *shared view*. Sticking with *SharePoint* will produce a list of results that are entry-level publications on

SharePoint for people who have no previous knowledge of the application. For the query *SharePoint 2010 Disaster Recovery*, none of the entry-level results are anywhere to be seen, as the query processor has makes the assumption that anyone asking this question clearly knows about SharePoint technology.

For many years now, there has been a significant amount of interest in using natural language processing (NLP—not to be confused with neuro-linguistic programming!) for queries. The user types a sentence such as *Find all the projects we have carried out in India with a gross margin of more than 30%*. This is a well-formed instruction, but the information needs to have been indexed in order for the application to provide an answer. The margin information might be held in a finance system and there is no link to the project lists held in SharePoint 2010.

As with all aspects of search technology, it only matters that the particular approach to query management taken by a vendor works for the queries that your organization is going to generate. This is why it is important to undertake the user requirements analysis, come up with personas and use cases, and then work up some typical queries that can be used in the proof-of-concept tests.

## Exploratory Search

Enterprise search performance assessment is often fixated around the need to provide a specific document/information in response to a query, usually because the information architecture of the intranet or document management system is so broken that the document has effectively vanished. Little, if indeed any, attention is paid to exploratory search, where the user has little prior understanding of what query terms might be appropriate and how rich the collection might be in relevant documents. When users are presented with perhaps anything more than 200 relevant documents, they may well be overwhelmed with choice and then not know how to refine the search.

An engineer may be looking for ideas to reduce the failure of a power screwdriver in low temperatures. The immediate problem would be to define what is meant by "low temperature." If she defines this to be anything below 0° C, then this could be used in a range search. However, if her company has an engineering base in the United States, there could be a problem over temperature indices, as the test documents might be in Fahrenheit. This is a somewhat artificial example, but should be enough to show that just developing a starting point for a query is far from easy, even for people skilled in the subject area.

## Spellchecking

The quality of the search application's spellchecker makes a significant difference to user satisfaction, as it speeds up the search process by not wasting time looking for words that do not exist. In addition to spotting incorrect spelling, a good spellchecker offers suggestions, and this feature can be extended to auto-complete by presenting users with a list of words that match the query terms as they are being typed into the search box. Finding the balance between being helpful and getting in the way is not easy.

The suggestions will be made from a spelling dictionary that is generated from the index terms, but it should also be possible to add in special terms that are of value to the organization. This could include key members of staff with names that are difficult to spell correctly or office locations in places like *Rawalpindi* (the second *a* in the name is easily forgotten because it is not pronounced). The same goes for the *p* in *raspberry*.

## Retrieval Models

There are four different approaches (often referred to as *models*) to manage the process of matching the query against the index and delivering a set of results.

The first of these models is Boolean retrieval. It is named after George Boole (1815–1864), who as an English mathematician with a special interest in algebraic logic, in which logical propositions could be expressed in algebraic terms. Boole's work was taken up by Claude Shannon in the late 1930s as the basis for managing telephone circuits, and later, the circuits in digital computers. Boolean algebra is characterized by the use of the operators AND, NOT, and OR.

A query about the London 2012 Olympics could be represented as follows:

*London AND Olympics AND 2012*

If the user was interested in information about both the 2012 and 1948 Olympics, then the following query could be used:

*London AND Olympics AND (2012 OR 1948)*

The nested logic within the parentheses is familiar to anyone who has had to create formulae in an Excel spreadsheet.

This approach was taken by all the early search applications, but has the fundamental problem that the documents returned either meet or do not meet the query term. There is no room for fuzziness. Adding in more terms to try to be specific can result in relevant documents being excluded. It is not possible to rank the set of results as a list in descending order of relevance. This order is referred to as a ranked list.

To overcome this problem, Gerard Salton developed the vector space model in the late 1960s, although he did not publish the core papers on his work until the early 1970s. The mathematics of this model is very complex, but in principle it enables the computation of how similar a document is to the terms in the query. Many current search applications make use of the vector space model.

Search application vendors are usually unwilling to reveal exactly which model they are using in their products, and in any case, it is not just the retrieval model but how the results are ranked that is of importance to a customer. Each has strong proponents, but there is no one ideal model.

## Parametric Search

Parametric search enables the user to build complex Boolean queries from a set of facets or characteristics applicable to the content being searched. This is the approach that is widely used for Advanced Search on websites and intranets. Usually the parameters are selected from drop-down lists despite the challenges these present in terms of accessibility. One of the benefits of a parametric search is that it facilitates the use of the OR operator. Users can then search for information on projects in *Kuwait OR Dubai OR Oman*. Users of web search applications often overlook the implicit AND that is the default option, so that this query becomes *Kuwait AND Dubai AND Oman*, which only finds documents where all three locations are mentioned.

Although parametric search can be of significant benefit to users, the major problem is that a user does not know the extent to which a parameter may be having a major impact on the number of documents retrieved. In the days of remote access online database services such as Lockheed Dialog and SDC Orbit, it was possible to get a return from the search application of the number of documents that met each parameter and also the number that met all the parameters. This is rarely available in enterprise search applications, so the user has to wait for the results page to be presented before realizing that the selection of parameters is not optimal.

## Filters and Facets

The value of facets and filters can be very considerable but only when implemented with care. The two terms are often used interchangeably, but the differences between a filter and a facet are important. A filter reduces a set of results by whether or not they meet defined criteria, such as being published in a particular year or being related to a specific subsidiary or sales region. A filter will therefore remove some items from a result set so that, for example, only sales reports for the Nordic region are presented and reports for North America are excluded. Any further search refinement is then carried out on the reduced set of results.

Facets present a set of *characteristics* of the information in the repository, listing out elements such as year of publication, geographic region, size of project, and perhaps even the name of the author. Once the results of a query are listed out, the sets of facets, usually on the lefthand side of the results page, show the counts of the number of documents that contain the element. Computationally, this is quite a challenge.

The two main approaches are often referred to as *top down* and *bottom up*. In the top down approach, the number of hits per document is calculated from the inverted index. The bottom up approach works through the documents in the results set and then accumulates the number of occurrences. There are also some combined approaches, and exactly how the facet hit values are derived tends to be one of the nondisclosable elements of a commercial search application.

One result of the trade-offs that have to be made between the computational demands of accurate facet hit counting and being able to deliver results as expeditiously as possible is that the hit counts may be approximated. This is most obvious when the sum of the individual counts does not match the headline count.

For example, the Year facet may show there are 4,503 hits, but the individual year counts are 1,417, 734, 344, and 239, and continue to decrease by year. The user might be forgiven for wondering what happened to the other 2,000 or so results.

For filters and facets to work, there must be a very high standard of content quality and of metadata tagging. The user needs to be able to trust that if he selects 2013 sales reports, he does not also get some 2012 sales reports that have been tagged by a modified date and not the date of publication. A *false drop* like this on a set will cause the user to question the integrity of the search application.

# Ranking

The emphasis on ranking models in this chapter is because they are, to a very significant extent, the magic sauce that differentiates search applications. Although there is some scope to develop innovative technologies for stemming and tokenizing, these primarily affect performance and the ability to index complex documents. When it comes to ranking, the benefits could well be immediately obvious but perhaps only for certain categories of content where a particular algorithm works very well. The only way to find out which is the better ranking approach adopted by two different vendors is to assess them on defined, known collections of documents under controlled test conditions. That is not going to happen in the harsh procurement world of enterprise search, even if it should. Ad hoc tests will just confuse the situation.

There are a number of approaches to trying to give the user the documents she needs on the first page of the search results. Absolute query and relative query boosting are two examples of static ranking, and are based on business rules. For a number of queries, there could be one or more documents that are important to display either as the first or second result, or above the list of results. For example, any search for some HR-related terms such as *maternity leave* or *paternity leave* will always result in the user being presented with both the global HR policy document and the HR policy for the local unit. Both may be highly relevant because a manager located in India may want to check what the rules are in Sweden. This is sometimes referred to as a Best Bet, or absolute query boosting.

Under relative query boosting, for certain queries there could be one or more documents that a user should be made aware of, but which do not merit being placed at the beginning of a results list. Any search on *corporate performance* might have a rule that ensures that the latest quarterly report is always in the top 20 results, or possibly a PowerPoint presentation given to investors.

Ranking by decreasing relevance is just one possible sequence. In the world of enterprise search, date order can be important. A manager either wants to find out the most recent project reports listed in reverse chronological order (most recent first), or needs to find out the first time that a particular chemical was synthesized in the company's research laboratories (oldest first).

Without a doubt the most challenging task in search management is optimizing the ranking of search results. It requires a combination of knowledge, including the following:

- How to change the weights of each of the ranking parameters
- The content being searched
- The language (specialized terms) of the organization
- The business processes in the organization that might result in a requirement to search
- The expectations of the search user

This knowledge is very unlikely to be found in one person and is the main reason why a search support team is essential in achieving the highest possible levels of search satisfaction. Ideally, your team will have a combined knowledge of computational linguistics, information science, the mathematics of probability, and a sprinkling of computer science.

## Summarization

The entries in the results list will usually include a title, some additional data about the document (i.e., metadata), and then a summary of the document. There are many different ways of creating these summaries, including taking highly relevant sentences from the document and reproducing the text document, or displaying the search term within the context of a few words taken from the sentence in which it appears. This latter approach would seem to be an effective approach, but in a long document (a feature of enterprise requirements), this may just be a few sentences from a 200-page project report and may not be representative of the entire report.

A considerable amount of research continues to be undertaken into creating document summaries rather than using the rules-based approach, and it is very likely that there will be substantial enhancements to result summarization over the next few years. Although the algorithms are now relatively well developed, the challenge is how to endure that the processing time for summarization does not have a significant impact on the speed with which the results are presented.

Often little attention is paid to the presentation of summaries in search results, and yet there are often significant differences between search applications on how the summarization is achieved and presented. It is important to assess these in the specification and selection process, using some test collections to assess the differences. Another factor to take into account is the ease with which the summarization process can be optimized after installation.

## Document Thumbnails

Another approach is to display an HTML thumbnail of the document with the search terms highlighted, and with the facility to step through each occurrence of the term. Again, the more terms, the less successful this approach becomes, but it is especially useful for PowerPoint presentations when users are looking for a slide on which they remember there was an especially clever diagram that they would like to reuse. The extent to which thumbnails can be generated will depend on the file format of the document.

Another benefit of using document thumbnails is that it avoids the need to open up the document in another application just to view it for long enough to determine that the document is not relevant and close it down again. That puts quite a load on the hardware and network bandwidth. The quality of the thumbnails varies as far as being an accurate representation of the document.

There are some third-party suppliers of software to generate document thumbnails, including the Finnish company Documill.

## Query Auto-Completion

A popular feature of search applications is the listing of computer-generated queries that take the initial search term as a starting point. If I search for *ford* on Google, the suggestions returned are *Ford*, *Ford Focus*, *Ford UK*, and *Ford Mondeo*, all easily recognizable as frequent queries from a UK domain. Changing the search to *ford depth* (because I am interested in finding the deepest ford in the UK) then generates *Ford depth gauge* (which is close) but also *Deptford Goth* (a rock group) and *Eagle ford depth*, which is a rock formation in Texas.

As with so much of search technology, there is a lot going on behind query auto-completion (QAC). As the query is typed, the words are matched against a database of the frequency of query terms. This works well when there are very large numbers of queries logged by web search applications, but in an enterprise environment that is not the case. Instead, the suggestions are derived from the document index, though often only a partial index because of the need for a very quick response to the query. It is not unusual to find that this index is not security trimmed. Typing in *redundancy* might display *redundancy discussions London* and then mysteriously show no results because the documents relating to the discussions are locked down to a few designated managers.

Despite the apparent popularity of QAC, until recently, little research has been carried out on how users interact with the suggested queries. More research will undoubtedly now be carried out (it only takes one paper to catalyze a research program!), and over the next few years there are likely to be some significant improvements to QAC, to an important extent driven by the benefits in mobile search.

## Federated Search

In theory, the ideal strategy for search would be to have a query box that enabled a user to search all the information in the organization with a single query, no matter in which repository the information is stored. Several vendors are currently suggesting that information silos have to be broken down if the organization is to flourish.

There are two options:

*Option A: One big index*

In principle, it is possible to crawl and index any number of individual search applications, or business applications with a search component, and create a single index. That is not difficult. What is difficult is creating a ranking list of results that make any sort of sense to the user, including presenting them in a consistent way.

Given the number of applications and thus, the size of the index, the results lists are likely to be quite long. Delivering results with high precision is very difficult. Disaster recovery planning is very important in this option, as there are many points of failure. It is important to manage crawl schedules with care—one schedule will certainly not meet all requirements.

*Option B: Query federation*

The query is managed by one search application, which then sends out the query to other search applications. Results from all the applications are then either integrated, or more usually, presented in a number of different sections of the search results page. It is not sensible to produce a "ranked" list of results, as ranking cannot be calculated as an algorithm of the ranking in each of the individual applications. Technically, the presentation of results from query federation is known as the interleaving problem—how can the results be presented so that the user is able to appreciate how the items have been ranked against one another?

Both options require the use of connectors (see Chapter 3), which are challenging to write and maintain. A small change in configuration in one of the queried applications may end up disconnecting the connector. Commercial search vendors, such as BA Insight and Coveo, have libraries of connectors that they maintain, but they can also be obtained from systems integrators such as Search Technologies. When a connector between two search applications fails (though often they just fail to perform as expected), there is always an interesting discussion between the vendors concerned about which end of the connector has failed. Connectors will also manage security protocols, either through early or late binding. Almost inevitably, matching the security models in each application will introduce some latency into the delivery of results, and this needs to be carefully managed.

This approach works well when it is possible to query a search application (e.g., from a commercial publisher) but not crawl and index it.

There are quite a number of factors to take into account when considering a federated search approach:

- In both options, users are also able to query individual search applications. A lawyer looking for information from a matter database may not also wish to see apparently relevant results from the corporate intranet.
- It may not be possible to implement query suggestion and offer the same set of filters and facets to help manage long results lists, as these are driven by the metadata schemas used on the respective repositories and indexes.
- The user interface may become very difficult to use. If the results are divided up into small screen areas for each repository, then perhaps only a few results can be seen without scrolling. Google research indicates that users can scan a list of Google results in around nine seconds. This is a good benchmark against which to judge the time users take to review federated results
- Managing access permissions can be a significant concern, as the permissions may not map uniformly across all the repositories. This can slow down the presentation of a complete list of results.

Another challenge with federated search is making sense of the search logs, especially in the case of Option B. In the case of either option, adding a new (or even upgraded) search application to the list of searches or repositories can make such substantial changes to the ranking of results that users should be forgiven for thinking that search is broken. A final challenge is how to cope with cloud-based applications, such as searching both on-premise SharePoint and Google Apps in the cloud.

The devil is in the details, and there is no substitute for a prolonged period of both requirements gathering and proof-of-concept testing. An increasing number of commercial vendors offer some form of federated search, but do take the time to read the small print and at the end of each sentence write a short essay on "what the implications are for us." You can, of course, build a federated application in open source software, but at present this option is only for seriously brave and experienced search teams.

## Summary

The technology described in this chapter is the base technology for all search applications. Each vendor, or open source application, will have its own approach to exactly how each of these processes is delivered, but trying to differentiate between them on the basis of these core processes is not a good use of time. As with any software application, it is not how a process is carried out but whether the results are of value to the organization.

## Further Reading

Zhuowei Bao, Benny Kimelfeld, and Yunyao Li Automatic, "Suggestion of Query-Rewrite Rules for Enterprise Search," *SIGIR '12 Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, Oregon, 2012.

Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra, "Query Suggestions in the Absence of Query Logs," *SIGIR '11 Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, 2011.

Alexander Chucklin, Anne Schuth, Ke Zhou, and Maarten de Rijke, "A Comparative Analysis of Interleaving Methods for Aggregated Search," *ACM Transactions on Information Systems* 33:2 (2015).

Karen Sparck Jones, "Statistics and Retrieval: Past and Future," *International Conference in Computing: Theory and Applications (Platinum Jubilee Conference of the Indian Statistical Institute)*, Kolkata, IEEE, 2007.

Nigel Ford, *Introduction to Information Behaviour* (Facet Publishing, 2015).

H.J. Grierson, J.R. Corney, G.D. Hatcher, "Using Visual Representations for the Searching and Browsing of Large, Complex, Multimedia Data Sets," *International Journal of Information Management* 35 (2015): 244–252.

Peter Morville and Jeffery Callender, *Search Patterns* (Sebastopol, CA: O'Reilly, 2010).

Tuukka Ruotsalo, Giulio Jacucci, Petri Myllmaki, and Samuel Kaski, "Interactive Intent Modelling: Information Discovery Beyond Search," *Communications of the ACM* 58:1 (January 2015): 86–92.

# CHAPTER 5
# The Business of Search

Such has been the pace of change in the search business, that this chapter has had to be totally rewritten for this edition. All of the largest independent search vendors have now been acquired. There are still some that have found market niches, but given the rate of adoption of open source search, their business strategy is going to have to be agile. This chapter provides an introduction to the search business, with more details on open source search and on Microsoft SharePoint in Chapters 6 and 7, respectively.

Tracking developments in the business of search is eased substantially by the work of Stephen Arnold and his team of researchers who compile the Beyond Search blog, as well as publish reports on various sectors of the market. International Data Corporation and Gartner Group also track developments in this sector, but this research is only available to corporate subscribers. Gartner Group also prepares an annual Magic Quadrant report on this sector that is usually released publicly fairly quickly after publication by one or more of the vendors who have been given a strong endorsement in the review.

## The Acquisition Frenzy

Over the last decade, search vendors have come and (mostly) gone. Their technology was generally good, but until recently there was no compelling reason for organizations to invest in search technology. Moreover, large multinational companies required global support and most of the vendors had very limited marketing, sales, and technical support outside of the United States. There would perhaps have been substantial benefits in these vendors working together in some form of trade association to raise the visibility and value of search. However, each vendor considered themselves to be the world leader in technology and working together with their competitors was anathema. The end result was a lonely disappearance from the market.

Some did succeed, at least to a reasonable degree. Autonomy was the most visible and the only search software company to be publicly listed. However, by the end of 2012, most of these mid-range vendors had been acquired, as the following table shows.

| Company | Acquired by | Date |
|---|---|---|
| FAST | Microsoft | April 2009 |
| Exalead | Dassault Systems | June 2010 |
| Autonomy | HP | August 2011 |
| Oracle | Endeca | October 2011 |
| Isys | Lexmark | March 2012 |
| Vivisimo | IBM | April 2012 |

Subsequent to the purchase of Autonomy, HP took the view that Autonomy was overvalued, and HP then wrote down $8.8 billion of the $10 billion purchase price. Subsequent legal actions may have impacted the degree of confidence of current and prospective customers in the long-term future of Autonomy and its IDOL search software. FAST, Exalead, Endeca, Vivisimo, and Isys have all now been fully integrated into enterprise application suites.

As a result, IBM, Oracle, Microsoft, HP, and SAP are all able to offer search functionality in their enterprise suites. However, it can be difficult to ascertain exactly what functionality is offered and to be able to speak to search experts within pre-sales, sales, and support operations. The roadmap for the search functionality is also driven by the overall roadmap for the enterprise suite. As an example, some of Microsoft SP2013 functionality is currently only available in the cloud implementations.

## Independent Search Vendors

There are at least 60 independent search vendors whose main line of business is the development of search applications (see Appendix E for a list of these vendors). Most of these independent search vendors have revenues of less than $20 million, and many operate largely in a specific national market to reduce the costs of customer sales and support.

The challenge that these companies face is that they cannot afford to do much in the way of marketing and are virtually unknown to most IT managers. There is also a procurement issue in that procurement departments are always concerned about potential suppliers that have no published accounts. All the vendors will provide financial information under a nondisclosure agreement, but in many cases the profits will be minimal, as they are being ploughed back into the development of the software. Even then, the number of people who have a full understanding of the search software code base will be quite small.

## The Future of Commercial Search

There is no doubt that commercial search vendors have had a difficult time over the last few years. Many have vanished either through acquisition or a fundamental failure to attract and keep customers. Others, notably Coveo and Funnelback, seem to have been able to capitalize on this situation because there remains a good case for having a vendor deliver an integrated stack of applications and then maintain performance through technical upgrades and ongoing support. It is very difficult to get a reliable figure for the installed base of commercial search applications, but in trying to get a sense of the level of adoption, you need to take into account IBM and Oracle enterprise suite customers, any organization running SharePoint (especially SharePoint 2013), Google ESA installations, and, of course, cloud services such as Amazon AWS. The search capability of file-share applications is getting better, and most of the intranet products (e.g., Thoughtfarmer and Interact Intranet) offer good search applications. Sitecore now has Coveo as a partner—just one example of improved search performance in the CMS business. An important trend to note is the increasing use of Apache Lucene and Solr, with additional applications integrated into the suite and charged for on a commercial basis. This is, of course, the situation with IBM Omnifind, and other examples include Attivio, PolySpot, and IntraFind.

Clearly, open source development will continue to be a very important solution to a wide range of unstructured and structured content discovery requirements, but this approach does not suit every organization. An IT department may not have the resources in terms of skills and time to develop a set of search requirements, manage the development process, integrate it into the existing stack of enterprise applications, and then continue to manage the upgrade process as new requirements come along. The integration process is a particular challenge. As the benefits of search become more widely recognized, there will be a requirement to search across a wider range of repositories and applications.

This integration expertise could be provided by an internal or external development team, but software development companies often do not want to be in the business of systems integration. One of the justifications that is often used for open source search adoption is that the risk that the vendor goes out of business is eliminated. Provided that the application is built solely on open source components, then that would be the case, but as the industry continues to add in proprietary code to lock in organizations and to improve margins, that justification is increasingly of less value.

Much is made, rightly, of the benefits of having a development community that is constantly seeking to improve the code base. For example, the Elasticsearch support plans are all focused on development and production support. However, many organizations are looking for a different type of community, one that offers the ability to meet other users of the application and share implementation and development experience. Before the Microsoft acquisition of FAST Search and Transfer in 2008, the annual event in the search business was the FAST Forward user conference, providing an excellent opportunity for customers to meet together and share experiences, as well as helping FAST identify product development opportunities.

The need to provide their venture fund owners with a return on investment is inevitably going to focus the minds of the open source development community on revenue opportunities. My sense is that there is a slow but steadily increasing awareness of the importance of effective search solutions, thanks mainly to the results of surveys from AIIM, Findwise, and NetStrategyJMC, referred to in Chapter 1. The result is that the market for both new and replacement search applications is growing. That is good news for the venture funds behind companies such as Elasticsearch and Lucid Works. But it is also good news for startup companies that have a vision for a new generation of search applications built on open source platforms but targeted at IT departments (who still own or manage the majority of search applications) that feel more comfortable with purchasing a product with edges to it, and a traditional approach to product roadmap development and post-implementation support.

## Open Source Search Software

Over the last few years, open source search has moved center stage as a search solution. Although there are a number of open source search applications available, the dominant applications are Lucene, used in combination with Solr, and Elastic, which is based on Lucene. Both can be downloaded at no charge. The indications are that there have been over 20 million downloads of Elastic by early 2015. Chapter 6 looks in detail at the structure and future prospects for open source search.

## Intranet Search

There are a range of commercial and open source intranet application solutions, all of which include a search application. Most of these applications are custom designed to integrate very closely with the functionality (especially in searching for people) of the intranet. The issue here is the extent to which the search application can index other repositories—for example, a social networking application. In addition, consideration needs to be given to how the search application is going to be supported. It could be supported by the intranet team itself, except that this team is likely to be very small and have little expertise in search management.

## Search Appliances

A search appliance is a search application and disk storage ready installed in a standard rack casing. In principle, it can be installed and switched on in perhaps 30 minutes. The product concept has been made famous by Google with its Enterprise Search Appliance, but Google was not the first company to offer an appliance product. The search appliance was pioneered by the US company Thunderstone in 2003, though the company itself was founded in 1981. Other appliance vendors include Fabasoft Mindbreeze (Austria), MaxxCat**,** SearchBlox, Searchdiamon, and Teradata.

The Google innovation was the pricing policy, which is based on the number of documents to be indexed and searched. The search appliance license points begin at indexing 500,000 documents, and extend all the way up to 30 million documents or more. The Google Search Appliance is offered at two- or three-year license points, which include support, hardware replacement coverage, and software updates. When the contract period ends, a new contract has to be negotiated and a new appliance is provided.

This means that some careful calculations have to be made about the total cost of ownership over a five-year period that would be the minimum typical life span for a more conventional application. Most companies have no idea of how much information they need to index, much less the number of documents. Multiple versions of the same document quickly increase the number being indexed. Another factor to be considered is the cost of purchasing additional server licenses to provide for redundancy in the event of a server failure and also for development and test purposes.

In general, search appliances offer very good processing performance because the software and hardware are fully integrated by the vendor. However, it is usually difficult to tune appliances to improve relevancy, the range of connectors to other applications is limited, and customer support is often restricted to a local partner.

## Microsoft SharePoint

Microsoft SharePoint is probably the most widely installed of all search applications, with some organizations still working with SharePoint 2007 (aka MOSS07), SharePoint 2010, and now SharePoint 2013 either on premise or in the cloud. There is more about SharePoint search in Chapter 7.

There are two key points to take into consideration. The first is that the full functionality of SharePoint 2013 search is only available against content that is stored in SharePoint. The second is that specialized search expertise is needed to get the best out of SharePoint 2013. If you have been running FAST Search Server for SharePoint 2010 (FS4SP) on an Enterprise license, then the jump to SharePoint 2013 is not too difficult. By comparison, the jump from the Standard license search in SharePoint 2010 to SharePoint 2013 is considerable.

## Product Roadmaps

One of the most contentious issues around search applications is what the future roadmap for the application is going to be. The situation with search applications is no different from most other enterprise applications, with vendors being very reluctant to disclose more than perhaps a six-month release schedule. Two elements to pay particular attention to are any proposed changes to the server architecture and any proposed changes that might require a partial or complete re-index. Small changes in the user interface or the administration interface can usually be accommodated without too much effort.

## The Future of "Keyword" Search

Before considering the prospects for commercial search, the question of whether there is a future for "keyword" search is worth considering. In 1947, Winston Churchill remarked to the House of Commons that "No one pretends that democracy is perfect or all wise. Indeed it has been said that democracy is the worst form of government except all those other forms that have been tried from time to time." The same can be said for keyword search. The approach may not be perfect, though in the case of search, defining what is perfection is not possible. Other approaches have come along (probably the most visible being Autonomy IDOL) but none have become widely adopted.

The demands of the intelligence community, in particular, are stimulating the development of applications that can sift through very large quantities (often streaming) to find weak signals, but these applications are being used in organizations with highly skilled staff and very sophisticated technology platforms.

For the foreseeable future, most organizations are not able to make use of the capabilities of the wide range of keyword-based solutions that are available, and behind the scenes, there is a substantial amount of research being undertaken in how to extend the functionality and performance of these solutions.

## Specialized Search Components

Some of the software modules used in enterprise search applications are highly specialized. This is particularly the case with the management of languages. Two companies, Basis Technology and Teragram (SAS), are the market leaders in providing very sophisticated text analytics applications. Both companies have developed techniques for parsing and indexing Arabic and Asian languages that are widely used within the search industry. Another important sector is the development of document filters, which are available from companies such as Lexmark Enterprise Software and dtSearch. Oracle and HP/IDOL also have document filter modules, but the availability of these for use outside of their enterprise suites may be questionable.

## Cloud-Based Search

The scale of the cloud-based file-sharing industry is immense and many IT departments see cloud-based solutions as the default strategy for the organization. It also seems likely that hybrid search applications will emerge to take advantage of the scalability of cloud applications and yet maintain the security management and data privacy management features of on-premise applications.

It is vital to look very carefully at the small print to understand what is actually being crawled and indexed, what the costs are, and what the implications will be of migrating from one cloud provider to another in terms of transferring the index files. If these cannot be transferred then all the repositories will have to be re-indexed. There could also be differences between an on-premise solution and a cloud solution of the same product, with SharePoint 2013 being a good example.

Similar problems can arise with hosted extranet and collaboration applications, intranets, and enterprise social network applications. Currently, most of these applications provide no more than a fairly basic level of search functionality. The hosted search/on-premise search strategy needs to be carefully considered from a user perspective, as well as from an enterprise architecture perspective.

## OEM Applications

Some vendors will provide a version of their search application to companies in the document management, customer relationship management, and other enterprise applications. The version supplied to the customer may well have a reduced functionality compared to the current version of the product, and indeed may not be subject to the same upgrade roadmap as the standalone search product. In addition, it is highly unlikely that the search application can be extended to search other repositories.

## Systems Integrators

Smaller search vendors will often work directly with clients, especially where the software has been designed to work out of the box. There may be a need for a few days of support, mainly around the installation of the software on the server, sorting out disaster recovery options and testing them, and setting up the crawl routines.

There are now a number of systems integration companies that specialize in search implementation projects, offering a range of services, including defining the search requirements, managing the process of product selection, and then supporting the implementation. Most of these companies tend to focus their business around a selection of search software applications, but will have the skills and expertise to handle almost any search implementation project.

In some cases, vendors may feel that the implementation process is too complex for them to support, especially in countries where they may have little or no local office support, or where there are particular technical issues to be overcome, and will then partner with a local search systems integrator. This is usually a win-win situation for all concerned, though it is wise to make sure that the integration team is fully conversant with the version of the search software they are planning to implement.

Companies often outsource IT services, or use a systems integrator to provide support for the implementation of new applications. Search implementation usually only represents a very small revenue opportunity for systems integrators, and so there may not be many staff who can manage a search implementation. For this reason, systems integrators work with a small number of search vendors who can provide backup support to their consultants. It is therefore not surprising that a search integrator only works with a small number of search vendors.

## Summary

The market for independent commercial search application vendors is being eroded by the high level of adoption of Microsoft SharePoint, the rapidly increasing use of open source search solutions, and to some extent, the Google Enterprise Search Appliance. However, it is unclear whether the open source search business is sustainable given the requirement of investors for a return either through dividends or through a trade sale. This will not affect the availability of Lucene, Solr, and Elasticsource code for development purposes, but may result in a new generation of commercial vendors targeting at specific market niches through building specialized and charged-for applications on top of open source code.

## Further Reading

The search industry is tracked and analyzed by Stephen Arnold on his blog Beyond Search.

The Gartner Magic Quadrant report on Enterprise Search does not have a defined publication date each year. The report can be purchased from Gartner, but is often published (sometimes in a slightly edited format) by companies featured in the report. In the past, Forrester, Ovum, and the Real Story Group published reports on the enterprise search business, but these have been discontinued.

Steve Silberman, "The Quest for Meaning: The Story of Autonomy," *Wired*, February 2000.

CHAPTER 6

# Open Source Search

Over the last few years, open source search software has emerged as a very sound option for organizations looking for search applications for websites, specialist ecommerce sites, search-based applications, and enterprise search. At present, it is impossible to get any market information on the extent to which open source search platforms are being used. It seems likely that the majority of current implementations are for website search and for specialized service applications such as LinkedIn and Twitter.

However, organizations are increasingly considering using open source search software for internal intranet and other enterprise search requirements, and this is likely to accelerate quite rapidly now that most of the larger-scale commercial search software vendors have been acquired and integrated into enterprise suites (see Chapter 5). From a search strategy perspective, it is important that the strategy takes into account the potential introduction of open source applications, even if the incumbent application is SharePoint or a standalone commercial search application.

It is far too easy for an organization, and its IT team, to work on the basis that going open source will result in an effective solution in a very short period of time for very little expenditure. Under some circumstances, an open source solution could meet those requirements. However, if the requirements are not fully considered in advance and allowance made for changes even within the development project, let alone post-implementation, then the results will be disappointing and embarrassing.

## The Rise of Open Source Search

The initial step toward open source software was taken by Richard Stallman in 1983 with his GNU Manifesto. GNU, which stands for Gnu's Not Unix, is the name for the complete Unix-compatible software system that he wrote and wished to give away free to everyone who could use it. The concept of open source software dates from 1998 in the wake of the release by Netscape of the code for its browser. The Open Source Initiative was set up in the same year and drafted a definition of open source code that remains the standard.

The Apache Software Foundation (ASF), one of the major players in open source software, was set up in 1999. It is a membership-based, US-based, not-for-profit corporation with the objective of ensuring that the Apache projects continue to exist beyond the participation of individual volunteers. Membership is open to people who have demonstrated a commitment to collaborative open source software development, through sustained participation and contributions within the ASF's projects. An individual is awarded membership after nomination and approval by a majority of the existing ASF members. Individual Apache projects are, in turn, governed directly by Project Management Committees (PMCs), which are made up of individuals who have shown merit and leadership within those projects.

The Lucene open source search code was written by Doug Cutting in 1999. He had held a number of positions in IT companies working on search projects, including Apple and Xerox PARC. He went on to work for Yahoo! and is currently on the executive team of Cloudera. The name Lucene is Doug Cutting's wife's middle name. Cutting also developed Nutch as a web crawler and more recently developed Hadoop. He transferred the rights of Lucene and Nutch to the Apache Foundation in 2001, and it became a top-level Apache project in 2005.

Solr was created in 2004 by Yonik Seeley at CNET Networks as an in-house project to add search capability for the company website. It adds functionality such as hit highlighting, faceted search and filtering, caching, geospatial search, and a good web administration interface. Whereas Solr can be installed and used by non-programmers, installing Lucene requires programming expertise.

Yonik Seeley, along with Grant Ingersoll and Erik Hatcher, went on to launch Lucid Imagination, later renamed to LucidWorks, in 2008. Solr was given top-level project status by the Apache Foundation in 2007, and the Lucene and Solr projects were merged in 2010 and are now generally referred to as Apache Lucene/Solr. Yonik Seeley left Lucidworks in 2013 to set up Heliosearch, which offers a variant of Solr.

There is a common misconception that Lucene and Solr are "supported by a global team of developers." In some senses that is correct, but without good quality control and a sense of a development roadmap, the functionality of the code bases would quickly get out of control. The quality management is the responsibility of committers, who are the only developers permitted to make changes to the main code base. Noncommitters may also submit patches, but these will not become part of the main code base until a committer has reviewed them. The Apache Foundation provides information on the responsibilities of committers, which gives a sense of the way in which the process works. There is also a Lucene/Solr project website that lists Lucene/Solr committers, of which there are currently around 50.

Shay Banon released the first version of Elasticsearch in February 2010 based on work that he had been undertaking on search software development since 2004. During 2013 and 2014, Elasticsearch became an increasingly popular alternative to Lucene/Solr. It is not an Apache Foundation application, though it can be downloaded under the Apache 2 license and uses Lucene code. In 2015, the software was renamed Elastic.

The latest entrant into the market is Heliosearch, founded by Yonik Seeley, the original developer of Solr, together with Joel Bernstein and Erick Erickson, both formerly with LucidWorks.

Underpinning all three platforms is Java, a general-purpose computer programming language that was developed by Sun Microsystems to provide a platform-independent application development environment. In 2007, Sun moved Java to an open source license. Sun was acquired by Oracle a year later, and Oracle maintains the open source approach. The current version is SE8, which was released early in 2014.

## The Current Situation

The development of both Lucene and Solr has been very rapid over the last few years. Although many of these version releases have been associated with bug fixes and minor enhancements rather than fundamental upgrades to the software, the upgrade to version 4.0 was a major change and added SolrCloud, which allows easier scaling of Solr. In the same period of time, a commercial application would have perhaps been upgraded once or perhaps twice. The speed of version release indicates a responsiveness to user requirements, but it may be quite challenging for IT teams accustomed to the upgrade pattern of commercial enterprise applications. Version 5.0 of Solr was released in 2015.

Elasticsearch continues to be available as a free download, but in 2013, a commercial company was set up by Shay Banon and his colleagues very much along the lines of LucidWorks. In February 2013, the company, which is based in the Netherlands, gained a $24 million Series B round of funding from Index Ventures, Benchmark Capital, and SV Angel. The funds were to be used to improve its ability to meet and support the increase in customer demand for Elasticsearch. Elasticsearch is often referred to as the “ELK Stack” because the company also offers Logstash (a data pipeline management application) and Kibana (a visualization application) as a tightly integrated trio of applications. The company also offers a security management application (Shield) and a management dashboard application (Marvel). This is where commerce meets open source, as Marvel and Shield are available through a range of subscription-based support packages.

In March 2015, Elasticsearch was renamed as Elastic and acquired Found, a Norwegian company that offered hosted and managed search services based on Elasticsearch, putting the investment funding to good use. At that time, Elastic reported that there had been over 20 million downloads of Elasticsearch.

LucidWorks gained a $10M investment in 2010, and in October 2013, announced a further investment from In-Q-Tel, though there was no information about the scale of the investment. At that time, LucidWorks claimed that it was delivering 9,000 downloads a day and had 4,000 customers.

Both the leading commercial suppliers of support and development services are therefore being funded through venture capital investment. Venture capital firms want to see a return on their investment, and working out the value of these two businesses is going to be a challenge when the fundamental intellectual property is open source and free. In effect, both companies are providing consulting services to support the open source applications.

# The Open Source Search Ecosystem

The structure of the open source search business is quite complex. There are many options, at least for Lucene and Solr, that an organization could adopt to implement an application. One of the major advantages of selecting an open source approach is plugging into a large and growing community of developers and users and being able to take advantage of their experiences. There is a vast ecosystem of related projects (plug-ins, filters, administration systems, interfaces, etc.), many formal and informal events that may include local networking groups, and a huge amount of written material. Choosing open source may also make it easier to employ good developers, as often these people prefer open source as a development method. At present, LucidWorks, ElasticSearch, and Heliosearch seem to be struggling to develop a marketing proposition, unsure about whether they are targeting IT departments and the development community or business managers looking for search solutions.

## In-House Development

The Java code base of open source search applications is complex and cannot be just picked up from a book or a training course. Many IT departments have very limited experience of using Java as a development platform, and even in large organizations there may only be a few employees with an adequate level of expertise. Both Lucene/Solr and Elasticsearch provide rich and detailed APIs to allow for search applications to be built, and it is unlikely that developers will need to modify the core software itself. The issue is not so much one of finding software development expertise, but of also having the expertise in information retrieval to translate requirements into code, and then continue to support the evolution of the application. Search development is not a project but business as usual.

## Independent Developers

There are a number of small companies that specialize in developing open source search solutions. These companies have usually been working with open source solutions for some time and have gained a significant amount of practical experience along the way. Most of these companies will work with Lucene, Solr, and ElasticSearch, and some may employ committers to these projects.

## Search Implementation Companies

There are a small number of search implementation companies that are able to support a range of commercial and open source search applications. Examples in Europe include Flax, Search Technologies (which also has an extensive operation in the United States), Raytion, Findwise, and France Labs. Search implementation companies based in the United States include Open Source Connections, TNR Global, and Sematext. As with independent developers, some of these companies may employ committers.

## LucidWorks

LucidWorks (formerly LucidImagination) is probably best regarded as a special case of a search implementation company. It claims to have more committers for Lucene and Solr than any other company and over the last few years has worked hard to raise the profile of open source search. There is a substantial amount of information on the LucidWorks website but nothing at all about the pricing of the various support packages. The website also lists apps, such as connectors and language management applications, some of which are proprietary to LucidWorks, and others which come from specialist software companies such as Search Technologies, Raytion, and Raritan. This highlights the fact that open source applications do often need to incorporate additional software components.

## Applications Based on Open Source Search

A number of companies have gone one stage further than LucidWorks in that they do not offer open source development services but have developed proprietary applications that make extensive use of core open source applications. These include Attivio, IntraFind, and Polyspot. These applications fall somewhere between commercial search applications and open source applications. Perhaps the most striking example is IBM, which launched its Omnifind Yahoo! Edition in 2006 as a free download. At that time, Omnifind was a proprietary application. Although not apparent at the time, this was the start of IBM moving OmniFind onto an Apache Lucene code base.

## The Open Source Community

The open source search community practices what it preaches and works together in an open and cooperative way. There are now many Meetup groups, an increasing number of books and training courses, and hackathons to test out novel approaches to solutions and problems. This is in complete contrast with the commercial vendors, who tended to be very proprietary about not only the software architecture but even how to get the best out of the software.

Open source search applications are primarily being implemented on a standalone basis. The integration of these applications with other enterprise systems, especially when these systems use proprietary software, presents some challenges. The major IT vendors recognize that open source software is here to stay, but they also have to continue to deliver dividends to shareholders.

## Solr or Elasticsearch?

The question inevitably arises as to which is the "best" application, given that both are based on Lucene. As with so much else around search, the answer is that "it depends." For some time, Solr was the only game in town, but the arrival of Elasticsearch has resulted in some useful competition between protagonists on both sides to highlight strengths (less so weaknesses!) and to ensure that each remains a leading-edge search application.

There is no point in conducting a feature by feature comparison. First, there are too many features to compare, and second, it is how the development team makes use of the features that is important. For any given search requirement, both Solr and Elasticsearch development will be able to provide a very good solution. It comes down to having a very clear view of what the requirement is and then feeling reassured that the development team has the experience to deliver a robust solution that not only meets the current requirement but is also scalable and extensible for the likely roadmap of future requirements.

The current situation is along the following lines:

- If the organization has a Java-based application development environment, with a mid- to long-term strategy for continued investment in Java and with a good in-house team of Java developers who are accustomed to working with open source projects, then it can certainly lever the Lucene/Solr applications for the greatest effect.

- If Solr is already being used in an enterprise environment for text search, then a move to Elastic is not going to make much difference in terms of functionality and performance.
- Arguably, Solr is closer to the ideals of open source given the commercial structure of Elastic and the fact that at some time in the not-too-distant future the venture investors will need to make some money.
- If the requirement is to support text searching and analytical queries, then Elastic could be the best option.

Otis Gospodnetic, the founder and CEO of Sematext, explores the similarities and differences in some detail in his blog post "Solr vs. Elasticsearch—How to Decide?".

## Developing an Open Source Search Strategy

The first critical success factor for an open source search implementation is "strategy before specification." The danger with open source search development is that because there is no initial outlay on a license fee, the project slips under the strategy radar. If the organization is committing to spend perhaps £500K on a commercial product, then quite a number of departments and managers are going to be involved in the procurement and the sign-off. An open source search project can probably be funded out of the IT development budget, especially if the costs are being taken in two different financial reporting periods.

It is therefore all too easy to see a small-scale open source search application as a quick fix, perhaps for an intranet or a website, with a longer-term intention (but no plan) to widen the scale of the project at a later stage. There are two implications. The first is that it may be necessary to rewrite some of the initial code to cope with the change in requirements, or to add in new modules. The second is that the overall project cost may be more expensive than if the requirements had been developed in the initial stages of the development.

In addition, open source search applications need just the same level of care, attention, and support as any other search application, and so once the software is installed, the support costs are not likely to be any lower than for a commercial product.

Although an increasing number of organizations either have a search strategy or are in the process of developing one, the majority of organizations still do not have a search strategy. This is of especial importance in organizations that are planning to move to SharePoint 2013. Although it provides a very powerful and flexible search application for content managed within SharePoint itself, there are issues around the extent to which SharePoint 2013 can be more widely used to search non-SharePoint applications and repositories.

An important aspect of any search implementation is having a well-grounded security strategy. Extending an open source application that is being used for a website or for a web-based service where either all the information is in the public domain or access is limited to subscribers who gain access to a complete collection through a password is very different to managing the complex and often inconsistent content security issues common in organizations of all sizes. The comment that security access is "managed by Active Directory" is easy to state and much more difficult to implement than it might seem to the IT team, especially where there is a current or potential requirement to provide access to users who are not employees of the business and so do not have an AD identity.

The search strategy also has to take into account the IT strategy. An open source search development may be the first major open source project that the IT department has managed, and in addition it may need to develop through a different project management methodology. As the project proceeds, it is quite likely that the scope of the project may change as the functionality and ease of development open up opportunities that were not in the original specification. Open source development is best managed through an Agile project management methodology, with a good grasp of what a minimum viable solution might look like.

It is important to note the difference between fitness to specification and fitness to purpose. This is especially important with a search application because there are no workflow analyses that can be used as the basis for requirements. Implementing any search application without a search manager being in post is extremely risky. Furthermore, it is not until the application has indexed all the required content and is being used in a production situation that issues about search performance arise. Search applications need to be able to be modified easily should production experiences show that there are some issues that need to be addressed. This is a situation that is the same with commercial as well as open source search applications.

## Developer Evaluation

The challenges faced by the search team in evaluating potential suppliers of open source search development expertise are very similar to the challenges faced by web managers a few years ago when open source CMSs (e.g., Drupal and Joomla) started to appear on the market. With a commercial product, it is possible to meet several customers who are running the current version of the software and gain useful information not only about the functionality of the product but also the way in which the product was implemented and then supported.

Open source applications are usually custom built; even products such as LucidWorks may well include some additional apps to meet specific requirements. To a much greater extent, companies will need to evaluate the skills of the development team rather than the list of functions and features. For some companies, there may be little or no experience of selecting open source development teams, especially if most of the enterprise applications are from IBM, Microsoft, or Oracle.

Some of the factors that need to be taken into consideration in evaluating developers are listed in the following table. However, this list is not comprehensive and it needs the right set of skills on the purchaser side to understand the significance of the replies.

| | |
|---|---|
| Previous experience | What projects has the team undertaken that they feel are most similar to your project, and why?<br>What lessons have they learned from these projects that will be of relevance?<br>How close was the final cost to the initial budget? |
| Development skills | How do the developers keep up to date with developments in the software?<br>What training have they had either externally or internally?<br>What networks do they participate in and how far are they, in network distance terms, from committers?<br>What processes does the company have for assessing development quality?<br>If required, will the development team have the experience to integrate the search application with other enterprise applications? |
| Documentation | Ask to see examples of the code documentation and user manuals that have been provided to other clients. |
| Project methodology | Search development needs to be carried out on an Agile basis. Does the team have a best practices document on its development methodology?<br>How will this methodology integrate with the organization's own project management methodology (e.g., Agile and Waterfall do not work well together)?<br>What skills and support is the team expecting from you at various stages of the project?<br>Does the team use a project management application where all project documents, comments, schedules, and issues are located?<br>How are risks identified and managed? |
| Licenses | Although the core search applications may be downloaded under Apache Foundation licenses, there may be additional applications (e.g., document filters) that may be subject to a separate license. |
| Team vulnerability | Search development skills are in short supply. To what extent are the skills of members of the team backed by other team members? |

# Project Management

Many books have been written about software project management. This section highlights four issues that can make a difference to the outcome of a search project.

## Testing and Evaluation

It is very important to agree with all stakeholders what the test regime is going to be. There has to be a clear view on what tests are scalable and what tests (fully loaded files on a production environment) cannot be undertaken until quite late on in the project. Extrapolating production performance from small-scale tests of only part of the code base is extremely difficult to undertake even if the development team has worked on a very similar project in the past.

## Risk Register

A risk register is not a sign that the project will go off the rails but a methodology to determine when it is about to do so, and what actions need to be taken. It is usually reasonably easy to determine the risks of a project but much more difficult to set out what the early warning signs might be of the risk being about to emerge. Good risk management is about anticipating risks and reducing their impact by early diagnosis and action rather than waiting for the situation to arise and then having a solution ready.

## Project Manager

A search project is not an IT project. It brings no benefits to IT other than potentially reducing the IT budget. The benefits lie with the business and so the project manager has to manage the interests of the IT team, major business stakeholders, and the employees (or customers) who will actually use the application. Add in an Agile development environment, the still low installed base of open source search projects (and therefore project experience), and the difficulty of finding skilled project managers, and the task of finding and retaining the right project manager will be challenging.

## Project Documentation

In an Agile development environment, it is very easy to find that the documentation does not reflect the current status of the implementation. The documentation must be current with the project development, in addition to being written in a language that every member of the project team understands. One of the new words in Solr development is *shard* and this has nothing at all to do with a certain very tall office building in London. Many SharePoint project teams have a glossary of technical terms, so that everyone knows what a "list" and a "library" is in real terms. An open source search project will also benefit from a similar glossary.

# Post-Implementation Support

No search application development can be seen as a project. Even with the most capable of development teams, issues can arise very quickly after implementation as an unusual file format is discovered or a crawl needs further optimization. From the beginning of the engagement with the development team, there should be a very strong focus on how development support is going to be continued, including taking a decision on which upgrades should be implemented and when.

Certainly there is an increasing number of companies productizing open source search, but they are very small in IT vendor terms, often have limited support outside of their home territory, and are highly dependent on the skills of a small development team.

## Hadoop

Hadoop is another application that was developed by Doug Cutting when he was at Yahoo!, and is now available through the Apache Foundation. Hadoop provides a solution to managing very large data arrays and is not a search application for unstructured information. Hadoop is a distributed system for storing and processing information, and its ability to ingest data may outperform relational databases.

As a result, it has become widely used by businesses that want to collect and manage very large amounts of data, such as text from social media sites, sensor logs, and GPS-based location information, without some of the performance limitations of traditional solutions.

It is likely that many large organizations with extensive data collections will be experimenting or using Hadoop to manage access to these collections. Making the assumption that Hadoop is an enterprise search application is not a realistic one, though there are certainly text mining applications for which it would be well suited. Additionally, there are various integrations between search platforms such as Lucene/Solr and Elasticsearch and Hadoop, some sponsored by very well-funded companies such as Cloudera.

A detailed discussion about Hadoop falls outside of the scope of this chapter, and is included here just for the purposes of background information.

## Summary

Over the last few years, the Lucene-based Solr and Elasticsearch search applications have developed into very flexible and powerful solutions. So far, they have mainly been used either for website search or for specialized search applications where very high performance is called for, along with the ability of a development team to create customized applications. Their use in enterprise applications has been slower to develop.

There are many different ways to manage the development of these applications, ranging from in-house development to buying a packaged solution with perhaps some proprietary modules. Using downloaded versions of Lucene, Solr, and Elasticsearch will certainly result in savings in license fees, but the costs of development should not be underestimated, and need to be viewed over perhaps a three-year cost of development and support.

## Further Reading

Clinton Gormley and Zachary Tong, *Elasticsearch: The Definitive Guide* (Sebastopol, CA: O'Reilly, 2015).

Doug Turnbull and John Berryman, *Relevant Search* (Greenwich, CT: Manning Publications, 2015).

Packt Publishing has an extensive list of books on Apache Lucene and Solr.

# CHAPTER 7
# SharePoint Search

This is the only chapter in this book devoted to a single search application. The reason for this is the ubiquitous use of SharePoint in organizations of all sizes. By default, these organizations are able to capitalize on the search application that is a core element of the SharePoint 2013 architecture, and that makes it probably the most widely used of all search applications. However, even if the IT team has had experience in using FAST Search Server for SharePoint 2010, they may well be underprepared to take advantage of the much richer user experience in SharePoint 2013.

## A Short History of SharePoint Search

The search functionality of SharePoint 2003 and SharePoint 2007 was very limited, and Microsoft realized that without a significant enhancement of search, it could not position SharePoint 2010 as an enterprise-level application suite against IBM and Oracle. In 2008, Microsoft bought the Norwegian company FAST Search and Transfer, and rushed through the development of FAST Search Server for SharePoint 2010, often referred to as FS4SP. FAST Search and Transfer had developed FAST ESP as a very powerful enterprise search application that ran on both Linux and Windows servers. Microsoft continued to support the original FAST ESP application, though fairly quickly ceased to support the Linux version.

However, no further development was undertaken, and so from 2008 to 2011, the only version available was the 5.3 release from 2008. Full support for this application ceased in June 2013. The FAST ESP enterprise search application continued to be offered but support for this expired in June 2013. There is some limited assistance available until 2018. Despite the limited installed base of FAST ESP and also of FS4SP, FAST has achieved an almost mythical reputation for performance, and many SharePoint search managers refer to it in awe even though they have never used either

FAST ESP or FS4SP. As a result of the lack of familiarity, they also significantly underestimate the requirements for technical and search support.

## Search in SharePoint 2010

Microsoft offered two search applications for SharePoint 2010: SharePoint Search 2010 and FAST Search for SharePoint 2010. Because of the naming convention adopted by Microsoft, the impression was created that FAST Search for SharePoint 2010 (FS4SP) was identical to the FAST ESP application, and many search managers and IT managers were convinced that they had the full power of FAST ESP available to them. FS4SP is only available through an Enterprise CAL contract and so comes at a significant additional cost.

At the time of launch, FS4SP was positioned as a potential step toward a customer adopting FAST ESP as a broader-based search application by adopting many of the search management features of FAST ESP within SharePoint. This created the impression that there was going to be an option to upgrade to FAST ESP. Many organizations were surprised and disappointed both by the failure of Microsoft to enhance FAST ESP beyond Version 5.3 and then to announce that full support for FAST ESP would cease in 2013.

For customers accustomed to the comparatively weak feature set of the search application in SharePoint 2007, the migration to SharePoint 2010 Search needed care but was not a major leap in terms of search administration. FS4SP was a much more challenging prospect, especially if the organization had little if any search management expertise. SharePoint 2010 Search can be implemented almost out of the box, but this is certainly not the case with FSP4SP. The challenges are not just in the management of the backend servers but in the development of an effective search user interface.

## SharePoint 2013 Search

It is probably better to see SharePoint 2013 Search as a new product rather than an evolution, especially when compared to the Search Server application in SharePoint 2010. Microsoft has integrated search into all elements of the SharePoint 2013 platform rather than positioning it as one of the many elements within the overall application. Figure 7-1 is a schematic of the core modules of SharePoint 2013.

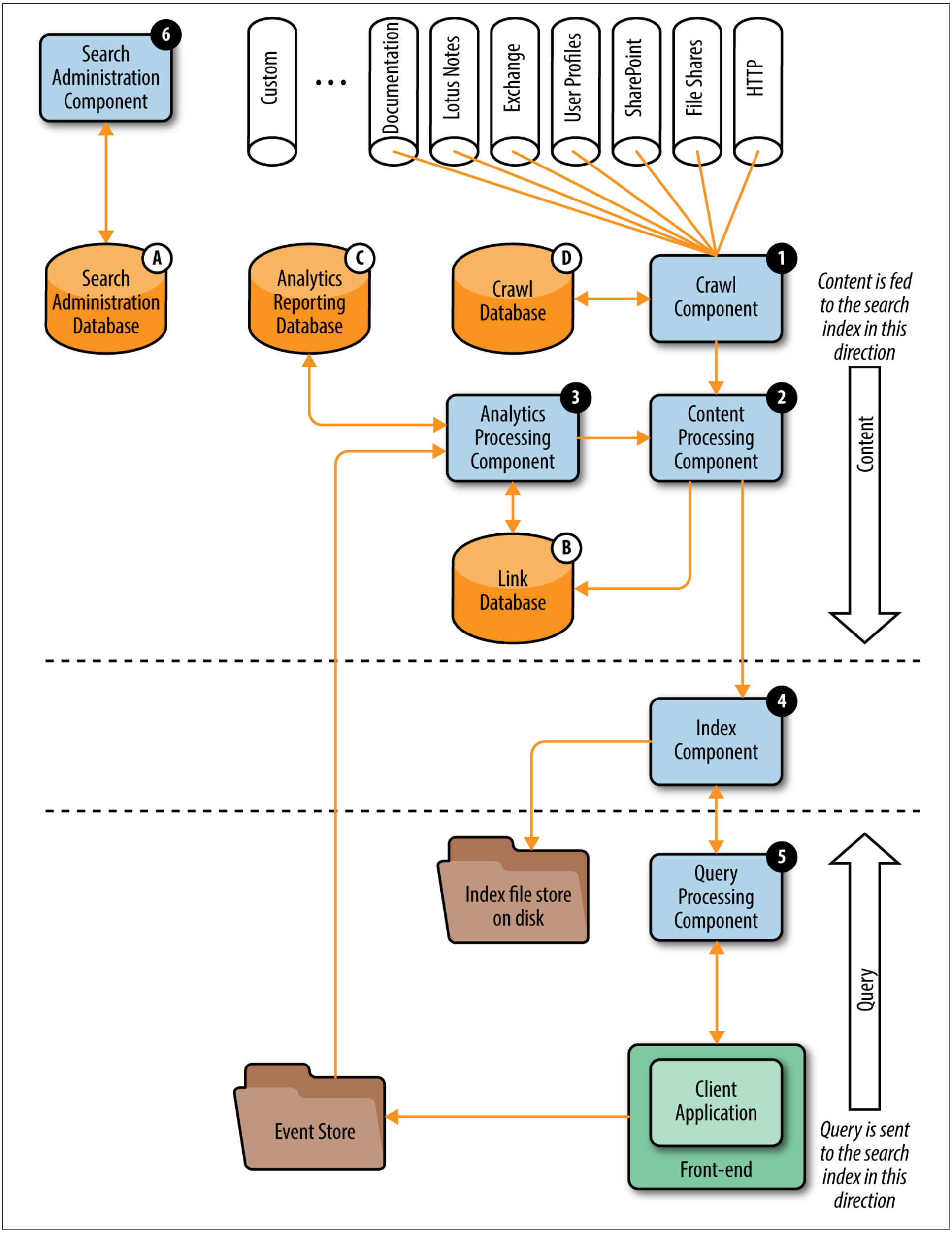

Search Administration Component
6
Custom
...
Documentation
Lotus Notes
Exchange
User Profiles
SharePoint
File Shares
HTTP
Search Administration Database
A
Analytics Reporting Database
C
Crawl Database
D
Crawl Component
1
Content is fed to the search index in this direction
Analytics Processing Component
3
Content Processing Component
2
Content
Link Database
B
Index Component
4
Index file store on disk
Query Processing Component
5
Query
Event Store
Client Application
Front-end
Query is sent to the search index in this direction

*Figure 7-1. Schematic diagram of the search architecture of Microsoft SharePoint 2013*

Some of the most important changes are:

- Complete integration of search within the SharePoint platform
- A simplification of the content processing pipeline
- Major changes to crawling and content processing
- The introduction of the Analytics Processing Component
- Substantial changes to the user interface
- Built to be a cloud-based application, offering a hybrid search of on-premise and Office 365
- More control at site collection and site level

The need to support a cloud-based architecture is one reason why certain changes have been engineered into SharePoint 2013.

Figure 7-2 illustrates the situation that will be faced by search managers familiar with either of the two SharePoint 2010 applications.

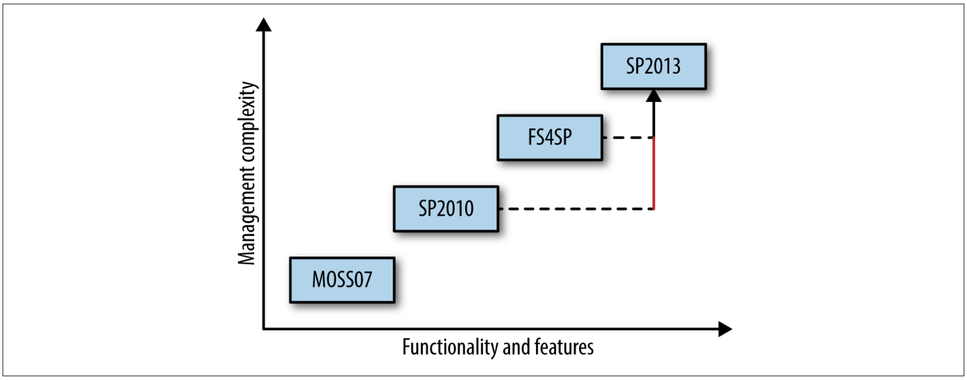


*Figure 7-2. With power comes management complexity*

Arguably, the base management of SP2013 is easier than with FS4SP because so many of the options have been tied down, but with more options on the user side, the overall expertise base of the support team needs to be wider and more integrated with the business.

Particular attention needs to be paid to the migration from the SharePoint 2010 Standard CAL search to SharePoint 2013 search. This requires very careful planning both with the IT team and with the business to understand the implications for migration and successful long-term search support. SharePoint users have long been familiar with the unique way that Microsoft uses terms such as List and Library. In moving from SharePoint 2010 to SharePoint 2013, there are some language changes that need to be fully understood. This is especially the case for Best Bets, Promotion and

Demotion of Results, and Synonyms and Scopes, all of which are now managed within Query Rules. In moving from SharePoint 2010 to SharePoint 2013, it is advisable not to assume that the same label achieves the same action.

## Installation, Administration, and Maintenance

From a technical perspective, the initial planning of SP2013 search deployment has to start with the hardware infrastructure. All the search components are tightly integrated into SharePoint 2013, so that each search component will be on one of the servers in the SharePoint farm. Both the crawling and querying components can be scaled out to more servers to improve the performance, but this needs good planning and a very good understanding of capacity planning for search applications.

The challenge lies in keeping a balance between content volume (item count) within SharePoint 2013, the query load in Queries per Second (QPS), and the crawl load in Documents per Second (DPS). Because of the range of crawl options, the crawl load element needs careful assessment.

From a software perspective, the installation of SharePoint 2013 Search infrastructure is relatively easy but again needs to be planned with care. It can be done either on the Central Administration UI or by repeatable scripts. Although this second approach needs more preparation, this is probably the best approach.

Search administration can be also done on both the UI and by scripting. Again, this second approach is getting more and more important as the search platform and environment become increasingly more complex. On one hand, we have many more features and opportunities in scripting than to be tied to the admin UI. On the other hand, we get a repeatable way of administration by scripting that is also critical.

It is important to appreciate that search administrators can delegate a lot of tasks to lower levels of the information architecture (site collections and sites). This means less overload on the central administration level, but also it needs more attention and governance as the environment can easily get to be a “search silo” without enforcing rules and policies.

## Crawl Management

Content freshness is one of the major measurements of search systems. In SharePoint 2013, the most important major improvement in content processing is the new type of crawling. Besides full and incremental crawls, there’s a new concept called “continuous crawl,” which runs every 15 minutes as a default schedule. It works on SharePoint content sources and enables changes (new or deleted items, changes in content or metadata) to be added to the index in minutes or even seconds. This new way of crawling is very agile. There can be multiple crawling processes running (to improve performance and content freshness), and these can be run in parallel with a full crawl on the same content source. However, a continuous crawl just ignores and logs any errors it finds.

Besides content freshness, this might be very important for large-scale implementations where it is not uncommon for a full crawl to take several weeks. In SharePoint 2010, no other crawl process could be run in parallel with a full crawl. Thus, the content, even if it had been crawled in the beginning of the full crawl, could not get refreshed until the full crawl finished.

With the new model of continuous crawl, it is possible to "refresh" the items already indexed even if the users modify them during the long full crawl process. As a result, there is always a current index.

## Analytics Management

A major change in SharePoint 2013 is the availability of an Analytics module. This module is not just a means of managing search logs but lies at the heart of many of the novel features of SharePoint 2013, such as the recommendation of content based on prior searches. The Analytics module also monitors documents that are opened by each user and makes the assumption that if the document is opened, it has importance to the user. This, and similar information, is used in ranking a document in a results list. This is carried out at a site collection level and enables SharePoint 2013 to deliver highly personalized search results based on the role of a particular user derived from which Site Collections are being used.

However, the Analytics module only tracks events occurring to content that is being managed within SharePoint 2013. External content that is being indexed by SharePoint 2013 will not be tracked by the Analytics module, and this could have implications for relevance ranking.

The Microsoft Developer Network site provides a good summary of how to develop customized ranking models for SharePoint 2013 search. The amount of applied mathematics in this post (e.g., around the BM25 rank feature model) illustrates why specialist expertise is required to support the power of SharePoint 2013 search.

## Working with Metadata

Metadata has always been the "glue" of search solutions, and is essential for search-based applications. In SharePoint 2013, the concept of metadata management is broadly the same as in SharePoint 2010. Crawled properties are automatically generated by the metadata of the content source, while managed properties are created and controlled by search administrators. These controlled managed properties can be mapped to the automatically created crawled properties in order to be able to use them in end-user scenarios. For example, they can be refiners (facets) that can be displayed with the results and can be used to sort or filter the results as well as query by them.

What has changed in SharePoint 2013 is the management and maintenance of the search schema. With the new delegated administration of search, managed properties can be handled not only on the global Central Administration level, but site collection administrators also have the privileges to create and manage their own site collection-level managed properties. This can be very useful as departments, projects, and so on. can have their own search metadata sets, without having any effect on others, but it still needs to be set within an overall governance approach to avoid issues with a coherent cross-site ranking.

## Working with Queries

Users who enter the queries and use the search system will have different backgrounds, knowledge, personas, and expectations. They might be also interested in different content: salespeople look for customer-facing presentations, developers need technical documentation, and finance need the proposals, invoices, and payment certificates. Their search "maturity level" also might be different.

Query Rules are the way to help searches responding to the *intent* of users, by creating conditions and corresponding actions. When a query meets the search system, it performs the actions specified in the rule to improve the relevance of the search results, such as by narrowing results, changing the order in which results are displayed, displaying additional Result Blocks, or using additional queries or modifying the current one.

The last improvement on query (and UI) to mention is called "Content by Search." This is a new way to provide dynamic content from the SharePoint search index, based on dynamic search queries. The query can be either entered by the user or generated automatically. Obviously, content freshness of items displayed depends on the latest crawl, which is why continuous crawl can be so critical in some scenarios, as discussed earlier.

## UI Customizations

An important addition to the "look and feel" of search in SharePoint 2013 is the Hover Panel. This is a side panel that gets displayed when users hover the mouse over a search result. It displays a preview of the document (out of the box for Office documents and web pages stored in SharePoint), the outline, the most important properties, and actions to take on the item. The Hover Panel varies by the type of result.

In SharePoint 2013, customization of these UI elements is also much easier to develop and update. There is now no need to create and modify long and complex XML configurations. Instead, Display Templates control which managed properties to use, how to use and display them, and also the available actions. These templates are used in the search user interface, in the result set, and in Content-by-Search. They can be configured to display specific managed properties.

On the Hover Panel, the most important aspects that can be customized are the properties to display and actions available by result type. Document previews also have some level of configuration. Finally, the way to display refiners can be also configured by Display Templates. With these options, we get an easy and powerful way to customize our search UI and are able to build up great search-based applications.

## Managed Navigation

Managed navigation itself is not a search-based concept, although it can be used with search in some very elegant ways. It is a dynamic, taxonomy-based navigation that creates SEO-friendly URLs that are derived from the managed navigation structure. It provides an alternative to the traditional SharePoint navigation feature, even with the opportunity to create a global, farm-level navigation experience without any custom development. Moreover, this navigation experience can be combined with search. The landing page can be a search result page, with Content by Search, refiners, and more. The query that drives the results is how and where the user navigates. It is dynamic: as soon as a new term is added into the navigation term set, it will get displayed as a new node and users will see the related items (results) immediately. This is highly intuitive, dynamic, and a good basis for creating a catalog-like experience or an interactive search-based application.

When SharePoint 2013 is used to index either non-Microsoft content (e.g., PDFs) or content that is not being managed within SharePoint 2013 (e.g., the corporate website), it is important to test out the extent to which these sources are being managed by SharePoint 2013. This process needs to take into account the role of the Office Web Apps server, which is the component of SharePoint 2013 that provides previews of documents in the search results. Just because a document can be previewed does not mean to say that the indexing and relevance rules are applied in the same way as other content.

## Search Governance

Effective governance is now recognized as an essential component of a SharePoint implementation. Search Server for SharePoint 2010 could operate out of the box and needed little in the way of search support. FAST Search Server for SharePoint 2010 needed careful management of the backend processing, but the options for the user interface and administration were more limited. In addition, FAST Search Server implementations were generally in large, multinational organizations which had at least some degree of internal expertise in search management.

This may well not be the case for SharePoint 2013, and the move to this version may well expose a lack of understanding within IT departments about the value and complexity of search. SharePoint 2013 needs to be implemented within a well-developed search strategy, especially as many organizations of all sizes will already be running other search applications.

For example, one must consider the types of content that must be searchable and from what content sources. Decisions have to be made about versions (whether to search in the latest public version or in each previous one as well), exclusions and inclusions, content freshness requirements and crawl schedule guidelines, and other factors. Out of the box, only the latest published versions can be indexed.

It is also very important to have policies for the metadata (i.e., you need to determine what should be searchable, establish standards for displaying this metadata, and specify under what circumstances it can be used). In SharePoint 2013, the governance of metadata has an important new dimension. Because of the multilevel search administration, policies will need to be developed for metadata on different levels (central administration, site collection), and we have to define what must be on which level. Enforcing these rules is the next challenge, though SharePoint 2013 does provide some tools for this. There is also a requirement to make the rules about who is responsible for which set of metadata and what actions should be taken to monitor the effectiveness of the metadata in searching.

With regard to relevance, decisions have to be made about the ranking models to be used and the basis on which these models need to be changing or new ones created.

Last, but not least, it is important to understand and make full use of the user interface. Search can be and is everywhere in SharePoint 2013, from the organization-wide Search Centers down to the single but complex Content by Search web parts. Governance is very critical to avoid having a poor search experience that takes time and effort to redress.

## Search Support Team

For organizations that have invested in FS4SP, the migration to SharePoint 2013 search is in many respects not going to offer significant challenges to the search development team, but for organizations using a Standard CAL SharePoint 2010, there will be a requirement for a wider range of skills to get the best out of SharePoint 2013 search. Certainly the out-of-the-box implementation will provide some benefits, but to get the best out of the application will require investment in a search support team.

With SharePoint 2013, there is no “easy” option. Although there may not be the same requirement for developer support, the rich user interface and the range of analytics all require a skilled team of specialists on an ongoing basis. The following table[1] summarizes the roles and responsibilities of a core support team for SharePoint 2013.

| | System Administration | Search Administration | Content Administration |
|---|---|---|---|
| **Task** | Capacity planning<br>Install<br>Backup<br>Monitoring | Crawls<br>Property Mapping<br>Result Sources<br>Query Rules | Metadata<br>Content Types<br>Search-Driven Publishing |
| **Working with** | Central Administration<br>SQL<br>PowerShell<br>Logs | Site Collection Administration<br>Search Reports | Site Administration<br>Term Store |
| **Needs to know** | SharePoint Architecture<br>Performance Testing<br>Security | Information Retrieval Concepts<br>Query Syntax<br>Query Rules | Information Architecture Concepts<br>Usability<br>Catalogues |

In an organization of any size, the Search Administration and Content Administration roles will be full time, and the System Administration work needs to be undertaken by someone who can make this work his top priority. In addition, there needs to be a Search Manager who monitors changes in business requirements and maintains a close relationship with business managers. This is moving toward a team of three or four people as a minimum to support SharePoint 2013 search. This team would need to be increased by at least one search administrator if the SharePoint 2013 search application is used to index other content repositories.

## The Challenges of SP2013

SharePoint 2013 is being positioned by Microsoft as an enterprise search application capable of federated search across multiple repositories out of the box. At the same time, the full benefits of the upgrades to the search technology can only be gained from managing the content totally within SharePoint 2013.

A major change to SharePoint 2013 is that there is no Pull API, which enabled content from other applications to be selectively moved to SharePoint for indexing. To some extent, this serious omission has been overcome by the availability of Continuous and Incremental crawls, but this is not a complete answer. The number of connectors available from Microsoft to interconnect SharePoint 2013 with other applications is currently quite limited. Other companies are offering a wider range of

[1] Source: Jeff Fried, CTO, BA Insight at the Enterprise Search Summit, New York, May 2013.

connectors, but these require careful implementation and support. The presentation of the search results from other repositories is not very elegant. Most of these can be eliminated by some additional configuration to provide transparent user experience, regardless of the source of the content, except the thumbnail document previews on the Hover Panel. For example, unless content is managed within SharePoint 2013, it is not possible to provide thumbnail previews of the content, though there are third-party apps.

Organizations currently using FAST Search Server for SharePoint 2010 would be advised to look at what elements of the application are not available in SharePoint 2013. In some cases, the SharePoint 2013 functionality is better, but that is not always the case. With development effort, some of the weaknesses in SharePoint 2013 search can be addressed, but this might be beyond the capabilities of a company that did not have a team of experienced search developers.

Another option would be to use the range of applications from BA Insight. Over the last decade, this company has specialized in offering solutions that build on top of SharePoint. It clearly has a close working relationship with Microsoft while being an independent supplier of solutions, including a strong collection of connectors and tools for auto-classification and people search.

## Migration from SharePoint 2010

The benefits and challenges of implementing SharePoint 2013 search have an important bearing on the migration routes from SharePoint 2010. Migrating from an Enterprise CAL SharePoint 2010 implementation using FS4SP will need a careful review of what features of FS4SP are now deprecated (i.e., not supported) in SharePoint 2013. In particular, this will affect any highly customized search-based applications, to the extent that companies may well choose not to migrate these applications but run them in SharePoint 2010. At least with an FS4SP to 2013 migration there will be the in-house skills to take advantage of the power of SharePoint 2013. There will, however, be changes in relevant rankings, and users may well find that content that may usually have appeared on the first page of results no longer does.

Migrating from a Standard CAL SharePoint 2010 implementation is not a trivial task. First of all, a team with the requisite skills needs to be allocated, or perhaps even recruited, ahead of the migration. What is emerging as a good migration path is to implement SharePoint 2013 and use it to index and search the SharePoint 2010 implementation. This will highlight areas where there seem to be changes to relevance rankings and give the development team an opportunity to learn not only the new functionality of SharePoint 2013 search but also the new terms used by Microsoft to describe many of the features.

Once this has been accomplished, then the migration of the other components can be undertaken. Although there have been many changes to elements such web content management, these are usually not visible to a user looking at a page of content. That is not the case for a user undertaking a search who may have some initial difficulty making use of the new features of the user interface.

However, the upgrade to SP2013 search is also usually being accompanied by a concurrent migration of content from SP2010 to SP2013. Ideally, the opportunity should also be taken to remove redundant, obsolete, and trivial (ROT) content and introduce a more rigorous and consistent approach to metadata. For any organization, this will be a very significant project where almost every task has a dependency on another task. There may be some opportunity to use software tools to support the operation, but almost inevitably there will be a need for someone to touch every content item. To give some indication of scale, if each content item takes in total 10 minutes to check in, review, enrich, and migrate, that works out to be around 200 content items a week once a few difficult ones have been resolved. Dividing the total number of content items by 200 can be a quite frightening calculation.

From a search perspective, testing the SP2013 search cannot effectively be carried out until all the content is migrated, and that can put a great deal of pressure on the search team to find and fix bugs and test out the user interfaces.

## Future Directions in SharePoint Search

Microsoft has committed to a future release of an on-premise version of SharePoint in 2015, with the future being on cloud-based applications such as Office 365. In 2014, Microsoft released Office Graph as a graph database platform. Microsoft has built in a way that APIs can be introduced, and in effect, Delve is one of these. For Microsoft, this is a new direction, and during 2015 it is likely that more such applications will be released. There are, of course, many benefits from adopting a cloud-based approach to enterprise applications, but these may have an impact on search performance. For large-scale applications, there could be less control over how the indexes are distributed and crawled, and there may also be some limits on the number of documents that can be indexed and searched. Of greater importance may be that Microsoft cloud solutions are optimized for Microsoft applications in the same way that arguably SharePoint 2013 makes it more difficult to manage the crawl and indexing of non-SharePoint content. Federating search across different cloud services is likely to be challenging, to say the least.

Microsoft offers three hybrid scenarios for joint SharePoint 2013 Server on-premise and SharePoint Online cloud implementations, each based on what Microsoft refers to as a hybrid topology:

*One-way outbound*

SharePoint Server 2013 Search services can query the SharePoint Online search index and return federated results to SharePoint Server 2013 Search.

*One-way inbound*

SharePoint Online Search services can query the SharePoint Server 2013 search index and return federated results to SharePoint Online Search.

*Two-way*

Both SharePoint Server 2013 and SharePoint Online Search services can query the search index in the other environment and return federated results.

The factors that need to be taken into account in deciding which scenario to adopt include the following:

- The requirement for users to search, find, and use on-premises content and data when they are away from their office and probably using mobile devices
- The extent to which there is a need to access secure data from SharePoint Server 2013
- The implications of data privacy legislation on the storage location of data and information
- The extent to which the SharePoint 2013 Server implementation uses custom code
- Search latency issues arising from distributed storage and/or network bandwidth availability
- Integration and analysis of search logs

In early 2015, Microsoft announced the acquisition of Equivio and Revolution. Equivio supports law firms with a set of tools that automatically generate relevance-based indexes from large streams of text, and Revolution is the leading commercial provider of software and services for R, a widely used programming language for statistical computing and predictive analytics. Although these are both small-scale acquisitions, they do indicate a commitment from Microsoft to respond to market requirements for innovative approaches to text and data analytics tools.

It will be essential to read all the Microsoft fine print over the next few years. This is not because Microsoft is hiding limitations in its applications, but because search is a very complex operation and is not something that can be bolted together at speed. This is especially the case if there is, or will be, a requirement to search across non-SharePoint applications. Connectors are difficult enough to implement in on-premises situations. Cloud applications are probably an order of magnitude more complex. Reading through the list of publicly available Microsoft Technical Research

Reports will give a sense of the scale and likely direction of Microsoft in the search sector, remembering that anything really dazzling will not be on display.

## Summary

Microsoft has progressively enhanced the functionality of the search application with SharePoint over the last decade. The main change from FS4SP in SharePoint 2010 to SP2013 has been the provision of a much richer user interface, but this also requires a different set of requirements to be gathered from the organization and a different set of skills within the SP2013 development and operations team. The migration from SP2010 to SP2013 search is usually undertaken within an overall migration project. These projects are very complex to manage and to forecast schedules and resource requirements with any degree of certainty.

## Further Reading

BA Insight publishes briefing papers on SharePoint 2013 search.

Mark Bennett, Jeff Fried, Miles Kehoe, and Natalya Voskresenskaya, *Professional Microsoft Search: FAST Search, SharePoint Search and Search Server* (Hoboken, NJ: 2010). This book is now out of print but provides not only a detailed account of SP2010 search but also advice on enterprise search management.

David Hobbs, *Web Site Migration Handbook Version 2*.

TechNet, Plan Search in SharePoint 2013 (and associated subsections).

# CHAPTER 8
# Search Governance

This chapter is arguably the most important in this book. No matter how good the search technology and how closely it meets user requirements, without an appropriate level of investment in the search support team, the chances of continuing to meet the requirements of the organization and the individual requirements of users are going to be close to zero. My main objective in writing this book was to get this message across as clearly as possible.

Implementing search should never be "a project." The work of ensuring that users continue to have high levels of search satisfaction will never come to a close. Each week, and perhaps even most days, there will be something that needs attention. The role of the search support team is not just to be reactive, but to anticipate when changes to the search application need to be made, or to identify a training requirement that will address an issue that is just starting to show up on the search logs and user satisfaction surveys.

## Who Should Own Search?

Answering the question "Who should own search?" is never easy. The first issue is defining what is meant by ownership. There will certainly be an important IT element in the budget, but the purpose of enterprise search is to enhance performance across the entire organization. All other enterprise applications that support business processes, such as finance, HR, asset management, and customer relationship management are owned by the department that would be unable to meet its objectives without specifying, implementing, and supporting the application on a business as usual basis, albeit with an appropriate level of support from IT.

Search is different as it not only delivers information to all employees but often in all locations globally. With the possible exception of Corporate HQ functions, all the departments referred to here are usually replicated in each country and/or each subsidiary. Search may also be delivering information in multiple languages and may be integrated with a number of other applications. No wonder organizations struggle to find an effective governance framework.

In its 2014 survey, AIIM asked a very important question, comparing who actually took responsibility for search and who respondents felt should take responsibility. The outcomes are shown in Figure 8-1.

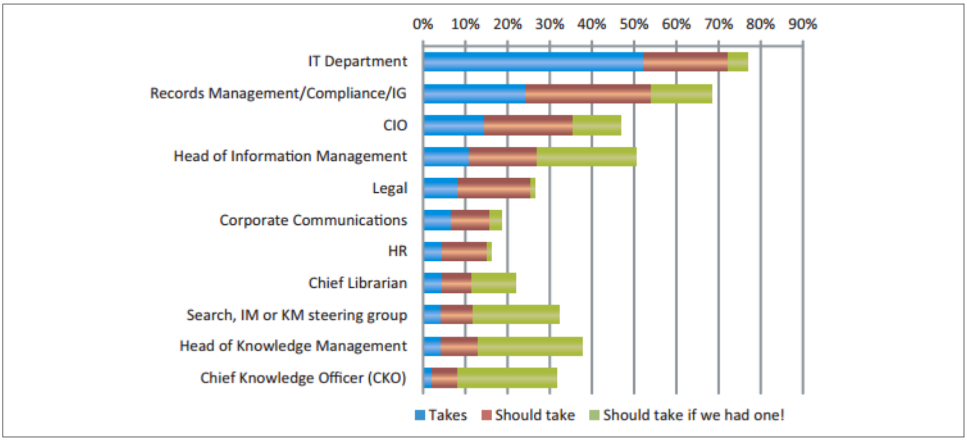


*Figure 8-1. Who would you say takes, and who do you feel should take, primary responsibility for search in your organizations? (N=308, multiple)*

For 52%, the IT department currently owns responsibility for search, but only half of the respondents were satisfied with this arrangement. On the other hand, the records management department is in charge in 24% of cases, but 54% of respondents would like to see it take charge. Most interestingly, 23% would like there to be a Head of Information Management, and 25% would like to have a Head of Knowledge Management, or even a Chief Knowledge Officer (CKO) at board level—albeit that almost no one has one of these already. The compromise is a search, IM, or KM steering group, in place in 4% of organizations, but suggested by 28%.

Looking again at budget, the major element of the cost is in staff and not in technology. In the case of the adoption of an open source search option, the balance is likely to be even more toward staff costs because there is no ongoing license fee. There will, of course, be ongoing development costs. Nothing is free in this world. In terms of the skills needed to support search, again AIIM provides important intelligence. Figure 8-2, from AIIM, gives a good indication of the level of support required.

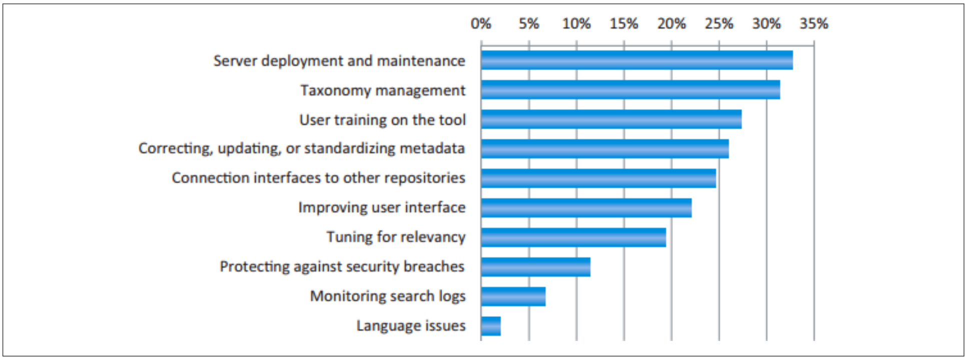


*Figure 8-2. What aspects of support have needed the most resources? (Max. two) (N=150, Excl. 33 Don't Know)*

Of these support aspects, server deployment, connections to other repositories, and protecting against security breaches require specialist IT skills. All the other aspects of support will need to come from people with a background in library and information science, or perhaps in computational linguistics. They will prefer to be reporting to someone who can relate to the content elements of search rather than the technology elements.

Figure 8-3 sets out the level of search expertise in the organizations that responded to the AIIM survey.

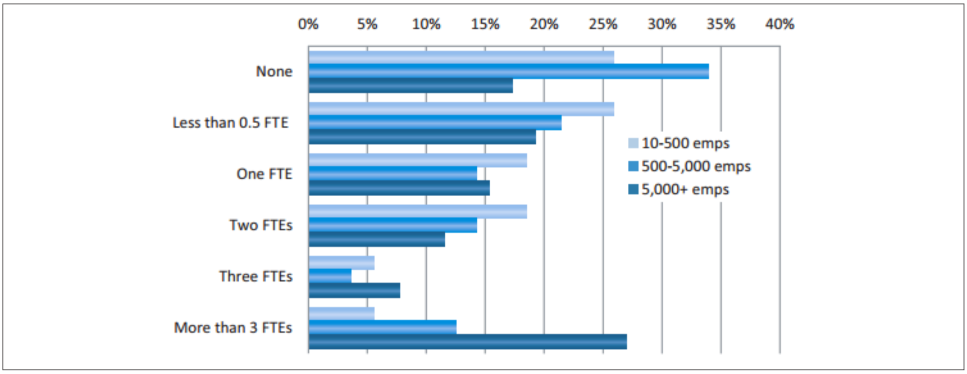


*Figure 8-3. How many dedicated (and trained) support staff do you have for your search application(s)? (N=192, Excl. 30 Don't Know)*

For a business-critical application like search, any number less than two means that there is no professional cover during the vacation period of the search manager, and if she leaves and an appointment cannot be made before she walks out the door, then the organization is exposed to a significant operational risk.

The reason for this risk elevation is that it is only by a diligent review of search logs within the context of the organization's operations that the reasons behind queries that are generating a low or zero number of results can be identified and addressed. It might be that the crawl has failed on a repository, a connector has failed because of an upgrade to the application, or a new area of business is generating novel inquiries. Whatever the reason, the organization is operating at an elevated level of risk.

In the end, the decision is likely to be political/organizational rather than pragmatic, and the downsides of the resulting decision need to be considered in detail and addressed.

Many organizations, especially those with multiple search or search-based applications, will undoubtedly have staff supporting these applications, even if only on a part-time rather than a full-time basis, adding to the complexity of the governance framework and spreading search expertise thinly across the organization.

In summary, this analysis indicates that there are five search team roles:

- Search Manager
- Search Technology Manager
- Search Analytics Manager
- Search Information Specialist
- Search Support Manager

At the specification and selection stage, not all of these are required full time. In principle, it might be thought that there is no requirement for the Search Analytics Manager at this stage, but given the importance of analytics, he needs to be involved in ensuring that the analytics requirements are fully specified and are tested during the proof-of-concept stage.

Even at the early stages of implementation, the team may be able to cope on a part-time basis, but the evidence is that this approach is not sustainable for very long. It is important to remember that search touches everyone in the organization who has access to a desktop, and any failure to locate business-critical information on a timely basis could have serious implications for the organization.

Enterprise search vendors tend not to be too explicit about the scale of support needed following installation of their software. There is a concern that if prospective customers are aware of how much support is needed, they may not proceed with the purchase. Even if the purchase of the software has been made some time in the past, there should be no reason why a search vendor should not be willing to share information about the size and roles of search teams of other customers.

## Search Manager

This role is not an IT role but instead requires a very good understanding of how information is used in the business, with a particular emphasis on unstructured information. The Search Manager might usefully have a background in information science or business intelligence applications, but the key success factor is that he understands the language of the business.

As well as managing the search team, the Search Manager needs to maintain a close working relationship with the search vendor, not only so that problems that arise with the search software are quickly addressed but also to gather and assess the experience of these partners from other installations. Sadly, it seems that many vendors are unwilling to bring their customers together to share experiences and good practice.

In most organizations, the most visible aspect of search is the intranet. As a result, probably the most important decision to be made is whether the Search Manager is a member of the intranet team or a member of the IT team.

## Search Technology Manager

This is an IT role, and the person concerned will be responsible for assessing server and network performance, crawling schedules, load balancing, and backup and disaster recovery. In a multinational company, this may require treading on the operations of national IT managers. Typically, an ERP or CRM application is country or at most regional specific, but enterprise search will be global from the outset and requires 24/7 availability. This may require an investment in hardware from the center, which cannot be justified by a national IT operation.

As a result, this can be a management role as much as a technical role, as the person concerned has to have the experience and the authority to ensure that things happen in operations over which she has no direct control. Just coming to an agreement on this can be a lengthy process of political negotiation, and needs to start right up front, and not when the software is about to be installed.

Another important responsibility of the Search Technology Manager is to manage information security, user authentication, and user permissions. It is usually not until an enterprise search application is implemented that all sorts of confidential information is found lurking on shared drives.

Finally, this role should take responsibility for API management and documentation. Effective enterprise search across multiple applications will require some complex APIs which have to be kept under review as the individual applications are upgraded or restructured. The scope of this role also includes tracking the performance of document filters and connectors, both of which can be susceptible to even small changes in application configuration.

In early 2015, the *New York Times* was seeking a Senior Search Technology Manager, and the responsibilities and expertise make interesting reading:

> Responsibilities include:
>
> - Building and mentoring a high performing team of search developers responsible for both the search platform and the metadata processes that support this platform.
>
> - Collaborating with stakeholders across the organization to balance priorities and support the company's strategic goals.
>
> - Participating as a strong voice in strategic technical discussions.
>
> - Practicing servant-leadership by mastering the systems you manage, so that you can roll up your sleeves and contribute code to even the trickiest tasks.
>
> - Advocating for a collaborative team culture that empowers individuals.
>
> - Leveraging the latest innovations in natural language processing, data science, machine learning, and distributed systems to build out our search systems.
>
> - Innovating solutions to the many search challenges unique to one of the web's most popular news sources.
>
> - Optimizing our systems for scalability, speed, high availability, minimal footprint.
>
> - Designing the infrastructure on which our systems run.
>
> - Collaborating with your colleagues across the company's technology, business, and newsroom departments.
>
> Requirements:
>
> - Previous experience as a technical manager
>
> - A passion for information retrieval
>
> - B.S. in Computer Science or equivalent experience
>
> - Unix / Linux proficiency
>
> - 3-5+ years experience programming in either Python, C++, or Java
>
> - Solid understanding of distributed, scalable web application architecture
>
> - Fundamentals of software design, coupled with a deep understanding of object-oriented software and design patterns
>
> - Experience with full product lifecycles, rapid prototyping, and iterative product development
>
> - Experience with Elasticsearch, Mongo, and Lucene a huge plus

This is a wide-ranging set of responsibilities and an equally wide range of technical expertise. It should not be taken as a template for a Search Technology Manager, but as an example of the extent of the expertise required to support the technology.

# Search Analytics Manager

One of the critical success factors for enterprise search is the quality interpretation of the search analytics. The volume of the search reports is very extensive. In one global consulting business, around 500,000 searches were being carried out each month. One of the most important tasks for the Search Analytics Manager is to work through the searches that resulted in zero hits being found. If the assumption is made that only 0.1% of searches failed to find anything, then this still represents a total of 500 searches a month, or around 2 each working day. Finding out why this search has failed may require some detective work, and certainly some feedback to the search user.

The Search Analytics Manager should also have responsibility for conducting user surveys to assess search satisfaction and search impact, and also for the analysis of help desk inquiries. All these sources of information need to be integrated with care.

# Search Information Specialist

Good search needs good consistent metadata, and yet metadata management is not given the priority it needs in an enterprise search implementation. As has been highlighted earlier, relevance ranking invariably places more weight on words and concepts in the title of the document. If the title is missing or is not well written, then the relevance of that document may be decreased, even if in fact the value to the user of the content of the document is high. The Information Specialist ideally needs to have a background in information science or in librarianship so that he has a fundamental training in metadata management and in the benefits and challenges of taxonomies.

A good taxonomy can be of considerable value in enhancing the search dialogue, but the development of taxonomies requires specialized skills, especially if a company is working in more than one language. Some search products (and Verity was a good example) offer customers support in the development of taxonomies, but it has to be realized that at present, and perhaps for some years to come, a totally computer-based approach to taxonomy development is not likely to be available. Of course, some search vendors decry taxonomies and say that their product does not require such an artifact from the world of library science. That may or may not be the case, so the Information Specialist will have the skills to determine the truth in this statement in terms of the particular collections that the company wishes to make searchable.

Another responsibility of the Search Information Specialist should be to conduct some standard test queries on topics that emerge from the search logs as popular searches. A lot can be learned from these queries, and they are a good basis for developing some best bets for common search queries.

## Search User Support Manager

This person acts as the user-facing member of the team, undertaking training and usability testing, and providing feedback from surveys on the performance of the application. Although in theory search applications claim to need only minimal training, the reality is that this is not the case, especially where federated searching is being carried out. Users may not fully appreciate the provenance of the various information repositories being searched and will need good guidance notes and suitable Help documentation on the search application.

Another important role for the User Support Manager is to develop and maintain good communications channels with users, perhaps using a section of the intranet, a wiki, or a series of blog posts to keep everyone informed about the ongoing development of the search application, highlight "tips and tricks," and report back on the solutions that have been found to the inevitable range of problems that have been identified.

## Search Help Desk

Because most employees will be using the search application, there are likely to be quite a number of calls with queries about the way in which the application seems to be working. A particular challenge of search is the technical complexity of the application, some aspects of which IT departments may not fully appreciate or be able to fix. Relevance tuning is just one example.

Most organizations have some form of IT Help Desk, and larger organizations will have a means of issuing tickets that log the query, track the progress of the resolution, and provide a database that can be analyzed for trends in the types of problems that have been encountered with hardware, software, and network components.

When it comes to search, there are a number of decisions that need to be made:

- Should the IT Help Desk be the first point of call for any search inquiry?
- Should the IT Help Desk be staffed with the skills needed to handle all search inquires?
- Should there be a separate Search Help Desk?
- If so, does the Search Help Desk become the first point of call for search inquiries or does it act as a support to the IT Help Desk?

The situation becomes more difficult with multinational operations as there could well be local IT Help Desks with only limited expertise in the main corporate applications. The use of these in some countries may be quite limited, but there could be many users of the search application, perhaps in a local language as well as in English.

There is no "best" model for Help Desk management, and much will depend on the skills and resources of the IT Help Desk(s). In the early stages of the implementation of an enterprise search application, many of the inquires may be about connectivity and technical performance, but as the application beds down, there will be more about content and relevance ranking.

## Team Skills and Training

It is not easy to find people with the skill sets needed to meet these roles and responsibilities. In the United States, iSchools seem to be paying more attention to teaching information retrieval and search technology than is the case in Europe, and currently there is no full-time undergraduate course in the world specifically on search and information retrieval. There are a wide range of master's courses, but these are focused on information retrieval research and not on the management of search in an enterprise setting. Computer science courses teach the basics of text retrieval but pay little attention to enterprise search management topics. This is a problem for organizations looking to recruit skilled staff and also for search vendors looking for developers.

The following is a summary of the basic skills that search teams need to have:

***Introduction to Information Retrieval***

- The historical development of information retrieval
- Distinctions between search and browse
- Relation between IR systems and databases
- A general architecture of IR systems

***Indexing***

- The various properties of text documents, such as structure, semantics, and metadata
- The concept of indexing, in particular full-text indexing
- The use of a pipeline for preprocessing text documents that includes tokenization, the use of stemming and morphological analysis, selecting subsets of terms based on term weighting, and the removal of noncontent-bearing words

***Retrieval and Ranking***

- Mechanisms for matching queries with documents, including the use of the inverted index
- Boolean and Best Match (e.g., vector space and probabilistic) modes of retrieval
- The notion of term weighting
- The use of query expansion (through relevance feedback) to maximize the success of matching queries and document representations
- The concept of an inverted index to improve search efficiency

***User Interaction and Interface Design***

- Understanding the users and their interactions with an IR system
- Supporting user interaction and search user interface design by effective design for interactive search

***Evaluation of IR Systems***

- The evaluation of information retrieval systems, including the different approaches for evaluating systems
- The use of test collections by system developers
- Measures of system effectiveness and user-oriented issues

***Web Search***

- The basic principles of web searching
- General-purpose versus specialized search engines
- The architecture of general purpose web search engines (e.g., Google)
- Crawling and ranking search results
- Search engine optimization

***Enterprise Search***

- The evaluation of enterprise search
- Architectures for enterprise search
- Commercially available systems
- Differences between enterprise and web search

Without a good, structured understanding of these topics within the search team, the team will be flying blind, relying on trial and error to improve search performance without understanding either the fundamental principles or the way in which they have been encoded in technology.

## Search Liaison Specialists

It is important to have excellent lines of communication deep inside the organization. It could be that just one business unit is having a substantial problem with searching across particular repositories and may not have the time or inclination to report back to the search support team.

This is where the appointment of search liaison specialists in as many business units as possible can be very valuable. They are the eyes and ears of the team, providing feedback on their own user experiences and listening for good news and bad news about the search experience coming from their colleagues. These liaison posts should be visible ones, so that users know who to go and talk to about their search experiences. The liaison specialists should be well trained in the use of the search applications so that they are in a position to provide on-the-spot assistance and to look at failed searches with an experienced eye. This liaison role needs to be included in the job description of the employee. Their manager should appreciate that if the employee moves from this current position, then the incoming employee may not be the best fit for the liaison role, and that someone else may need to be found.

## Supporting Global Enterprise Search

The support requirements are significantly greater when enterprise search is rolled out globally. There is likely to be a need for an Information Specialist for each major content language, especially in the case of German (where word length and complexity can raise some novel issues) and, of course, in ideographic languages such as Chinese, Japanese, and Korean. These and other languages (Finnish is the classic example) will need attention paid to stemming and lemmatization and to seemingly simple issues such as the way that organizational names (e.g., OECD) appear differently in French (OCDE). This may not be a full-time position, but certainly the expertise needs to be available to the search team.

For similar reasons there is a good case to be made for an analytics specialist for each business area in a highly diversified global corporation. The search terms used for one section or subsidiary of the business may well be very different from those in others. Investment banking and retail banking would be a good example. Staff in the Information Specialist and Search Analytics roles may not need to be located in the countries that they are supporting, but this is certainly not the case with the Search Support Managers.

However, there has to be a Search Support Manager in each major country, or at least each region (Europe, Asia/Pacific, North America), and language issues have to be born in mind. Although people may well speak several languages in business situations, they will prefer to search for content in the language in which they have the best command, so Spanish language search and support in South America is very important. Social media will also reflect national language.

As a result, the numbers can add up:

- One Information Specialist (IS) for each major language (x)
- One Search Analytics Manager (SAM) for each business area (y)
- One Search Support Manager (SSM) for each major country (z)
- Core search team of at least two for search management and vendor relationship management

Putting this all together:

Team size = x(Specialist) + y(Analytics) + z(User Support) + 2Core (Search Manager and Search IT Manager)

So for an organization operating in English, French, and German, with two main business areas, and with significant business operations in the United States, France, Germany, Dubai, New Dehli, Seoul, and Beijing, the numbers work out as follows:

Team size = 3(Specialist) + 2(Analytics) + 7 (User Support) + 2Core

That totals 14 members of staff, and for the purposes of this calculation, local IT support has been excluded.

This may seem quite a considerable team, but it can be interesting to find out how big the support teams are for enterprise applications such as an HR portal, an enterprise resource planning application, a business intelligence application, or a high-end document management application. For a large-scale enterprise resource planning implementation, the typical support level could be of the order of one team member for every 100 users.

## Search Center of Excellence

My experience is that the only solution is to establish a Search Center of Excellence (SCE) even if at the outset there is a team of one. A SCE will bring together all of the disciplines, skills, and experience needed into a single virtual team, and give search the visibility it needs as a core corporate resource. The team might not all be working on search full-time, but with a number of people working part-time, there will be the 24/5 or even 24/7 support that a global search application requires.

It is not possible to state how many people there should be in the SCE. In the case of an intranet, there are good indications that a ratio of one intranet manager for 3,000 employees is a reasonable starting point, but it is no more than that. In the case of search management, there are too many variables to take into account, including:

- The number of search applications or search-based applications
- The level of support provided by the vendor, systems integrator, or external development team
- The extent to which search is a business-critical application
- The number of different business activities, each of which is likely to have its own technical language and use cases
- The global distribution of search users
- The number of content languages
- The global distribution of indexed repositories
- The level of local technical support
- The extent to which IT Help Desks are able to provide support for search users
- The volume of searches, which will have an impact on the scale of the search logs
- The content quality of the repositories

Before setting out the roles and responsibilities of the members of a search support team, there are two important points to emphasize:

- In many cases, these roles may not be full-time, especially in smaller organizations. However they should always be priority roles for the people concerned, so that they can drop other responsibilities if there is a search challenge to cope with.
- As search begins to work better, the time required for support will increase because of the volume of searches and the increased expectations of users.

# Managing a Virtual SCE

Based on these and other case studies, some critical success factors are starting to emerge, though most are common to any virtual team. A virtual team is not necessarily one in which the members are at different locations. Team members could be widely separated physically and by department even when in the same office building. Although virtual teams are now commonplace, organizations rarely offer any training in virtual team management. This topic is outside the scope of this book, but there are references to virtual team management in the "Further Reading" list.

# SCE Case Study

A client of mine decided to set up an SCE to support the change from the current enterprise search application to SharePoint 2013. The organization is a global high-tech company with a European HQ but with substantial business interests in the United States and in Asia.

The search team was already set up as more of an informal community of practice than a formally constituted team. The five team members came from records management, quality management, IT, and two of the business units—one in Europe and the other in the United States. Individually, they were allocating around 30% of their time to search management. Two of the team members had a background in library and information science. Then, through the network of one of the team members, a sixth person, based in Korea, was identified who could only allocate perhaps 10% of his time but would be able to act as a local contact for the business units in Asia.

The business case for the team was built around the seamless transition from the current search application to SP2013. The company makes extensive use of search in research and in business development, and any interruption in service would be very damaging to ongoing business operations. The main responsibility for the team was to ensure that users throughout the company were fully aware of the switch in applications and the benefits that SP2013 would bring to searching a wide range of applications throughout the company.

A discussion about the tasks that needed to be accomplished in the 6 months prior to launch led to a list of over 40 such tasks. The team had the skills to complete the tasks but probably not enough time. It was decided to ask the company to set up a steering group for the search team which would be broadly representative of both the IT and business departments. The primary role of the steering group would be to help the search team prioritise their time and, if needed, take action to provide more resource on a short-term basis to ensure that the launch was successful.

The steering group had five members at first, which was then increased to six to bring in a senior manager from the United States. Only one of the group members worked in IT as the company recognized that the project had to be business-led and not just be a technology swop. Getting the time of senior managers for something that is perhaps tangential to their main business interests was eased by suggesting that there would only need to be a quarterly meeting of the group unless there was an urgent need to make a resources decision. One of the benefits of setting up the group was that it showed to staff and senior management that search was seen as an important application for the business.

In the project meetings, quite a substantial amount of time was spent discussing what the key performance indicators would be for a successful launch. As part of the launch planning, the team had identified quite a range of stakeholders across the company in order to develop a communications strategy, and it quickly became clear that working out what performance metrics to provide to each group of stakeholders was a core element of the communications strategy. A balance needed to be made between technical performance, query outcomes, and overall satisfaction with the search application. The company had not previously assessed search satisfaction, so the team decided to run a survey on the satisfaction with the current application; although they knew the results would show a low level of satisfaction, they were interested in establishing a benchmark for the SP2013 implementation.

Despite the experience of the team, or perhaps because of it, there was a concern about the time required to complete some of the tasks. We decided to build in quite a considerable amount of "float" time to allow for the unknown unknowns rather than look smart with a timetable for launch down to the nearest day. It is much easier to reallocate spare time than find additional resources at short notice.

## Security and Compliance

Search implementations tend to bring up some difficult issues around security management and regulatory compliance. Managing information security is a substantial business and IT task to ensure that documents are only able to be searched and opened by users with the appropriate level of security clearance. It is quite possible that the existing security access protocols will need to be revised to take account of the power of the search application to find information that it is not supposed to find, or more correctly, to find it and yet not deliver it unless there is a business/security case for doing so.

Potentially even more challenging is meeting the requirements of data privacy legislation, especially if the organization is subject to EU legislation, which extends to anyone from any country that is working in the EU. The problems of conformance are not just related to intranets. For example, sending details of an employee's CV to the United States from the United Kingdom as the result of a search being carried out in the United States without the consent of the employee could be in breach of the legislation. There is a view by some companies that if information is only sent to other sites of the company then the legislation does not apply. This is not the case, and full consent needs to be obtained. This is because there is also a very important distinction between personal information and sensitive personal information in EU legislation.

Sensitive personal information includes the following for a data subject:

- The racial or ethnic origin
- Political opinions
- Religious beliefs or other beliefs of a similar nature
- Whether the person is a member of a trade union
- Physical or mental health or condition
- Sexual life
- The commission or alleged commission of any offense

One of the key issues is that employees must give their informed consent for this sensitive personal information to be held in a database. Some intranets have an internal staff newsletter. In the interests of good communication, there might be a news story about how a member of staff had been ill, but was now coming back to work for a few days a week. This could be regarded as sensitive personal data, as it related to the health of the person, and this information should not be able to be disclosed to anyone outside of the EU.

Many consulting projects, especially in human resources and change management, may require the consultants to check on personal information about employees. Using a corporate intranet from a single site to gain access to this information is likely to be forbidden, and, of course, if this information is to be held by a third party such as a consulting company or an outplacement agency, then the employee's permission needs to be sought in advance. The employee also has the right to ensure that the information being held is correct, and this will require companies to implement intranet systems so that the employee can only see his own record, and not that of others. For employees that have left the company, this right will extend as long as their files are maintained, which also gives rise to a range of problems, such as the time that a company should reasonably maintain those files.

In reviewing search logs, there could be searches on voluntary redundancy, sexual harassment or discrimination, or for the addresses of senior staff. All these might be taken as an indication that the person carrying out these searches was planning to take redundancy, sue the organization for sexual harassment or discrimination, or send the addresses of senior managers to an animal rights activist group. The extent to which search logs might be construed to contain personal information has not yet been tested in the courts.

Data privacy compliance is especially important to take account of in the use of photographs in staff databases. Because a photograph will almost certainly contain information that enables a person's racial or ethnic identity to be inferred (even if incorrectly), staff photographs fall under the provisions of sensitive personal information, and specific permission needs to be sought from all members of staff before the photograph is added. This has to be informed consent, so the member of staff needs to understand the implications and cannot be penalized for not giving consent. The fact that the photograph is on a staff badge does not mean that the photograph can be used for a staff directory. The photograph on a staff badge is there to enable security staff to ensure that the badge is being used by the designated badge holder.

On January 25, 2012, the European Commission published a proposal for a new regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data. There are some substantial changes proposed to data privacy legislation, in particular the move from a Directive to a Regulation as a means of gaining greater harmony over Member State data privacy implementation. The risks of noncompliance under the Draft Regulation are substantially greater than under the current legal framework and, for the most serious breaches, a national data privacy authority may impose a fine of up to a maximum of 2% of a company's annual worldwide turnover. The new regulatory regime will come into force in 2015/2016, but the work needs to start soon on identifying any potential areas of noncompliance.

It is essential that the advice of lawyers specializing in data privacy is obtained. It is likely that in-house legal teams will not have any substantial expertise in this complex area, especially when an intranet needs to be compliant with a number of different national legislations. Currently around 40 countries have some form of data privacy legislation.

Another area where legal advice is important is the management of a situation when a court requires the company to disclose information for a court case or a regulatory compliance check. Here, the processes in the United States are somewhat different to those in most other countries, and are set out in Chapter 26 of the Federal Rules of

Civil Procedure. The implications are complex and potentially costly and should not be put on the list of "This can never happen to us."

## Training and Support

Training is not an activity that starts and stops in the implementation phase. It is not uncommon for 10% of employees to leave and arrive in the course of a year, and new employees will certainly be stress-testing the search application within minutes of sitting at their desk or switching on their smartphone. The communities that need to be trained and supported include the IT department and managers of applications that are being indexed.

## Establishing Good Communications

A good communications strategy is very important in achieving a high degree of user satisfaction. Some of the elements of this strategy might include:

- A search advice section on the corporate intranet, perhaps positioned close to the search box
- A blog from the search support team with success stories, tips for good searching, and information about upgrades and enhancements to the search application
- Presentations and workshops given by members of the search support team
- Best bets owners talking about how valuable users have found the best bets information to be
- A wiki that can be used to record issues that have arisen and the progress toward fixing them, and also to post advice on how to get the best from the search application
- Establishing search communities of practice in business units
- Publishing log information on a regular basis
- Finding and sharing some search success stories, perhaps by talking to people who were critical of the existing application and are now (we hope) very enthusiastic about the new search application

Communications is as much about listening as publishing, and so users should be encouraged to talk to members of the search support team.

A common problem with search applications is that progress is difficult to spot. Changes may have been made to one small aspect of the implementation with benefits to a particular group of users, but the search home page looks the same. My recommendation is to adopt the dot release model used by the IT industry. Define the current search application as Release 2.0 and place this prominently on the search home page. As enhancements are made move from 2.0 to 2.1, and onward. Then progress in developing the search application is made more visible, and the communications plan can then highlight and emphasis this progress. If there is more than one search application, then referring to all of them initially as 2.0 looks like a fix. A more creative approach to numbering is called for.

## Summary

Managing search is achieved by skilled people—who understand the technology and capabilities of the search application—working closely with users at all grades, roles, and responsibilities, and in all locations. That is a tough call and a significant investment, but without the skilled people, the organization is putting its business activities, objectives, and reputation at risk. There is a lot of work to do every single day, and with rare exceptions, in smaller organizations, being a member of the search support team is a full-time position.

## Further Reading

Bonnie Ranvild Frisendahl, "Virtual Information Training for a Global Business," *Business Information Review* 31:4 (2014): 237–242.

Martin White, "The Management of Virtual Teams and Virtual Meetings," *Business Information Review* 31:2 (2014): 111–117.

# CHAPTER 9
# Making a Business Case

The preceding chapters have provided a background to search management with a focus on search technology. Now the emphasis shifts toward the management of search. This chapter sets out the due diligence that needs to be undertaken before considering any significant upgrade of an existing application or the implementation of a new application. The 2014 Findwise Enterprise Search and Findability Survey asked respondents if they had a strategy for search. The outcome was as follows:

- IT focused strategy (11%)
- Business focused strategy (9%)
- Combined business/IT strategy (17%)
- Not yet but planned (29%)
- No (28%)
- Did not know (6%)

The AIIM survey in 2014, where the majority of respondents were from North America, indicated that only just over 12% had a search strategy. Because almost every employee in the organization will use search applications on a reasonably regular basis, this lack of strategy and therefore of overall planning, is a concern.

The search engine will need to interface with other applications, and there are some legal and compliance issues. In the future, the boundaries between search, business intelligence, and content analytics are going to become increasingly blurred, and delivering access to enterprise search through mobile devices is going to be essential within a year or so.

In this chapter, some of the core elements for a search business case are considered. Appendix A then sets out these and other elements in an A–Z list of topics. The reasons for considering business case development in this way is that the core elements will be important in appreciating the remaining chapters of the book, and then

Appendix A brings all the topics together in a list with references to where the topics have been considered in the book.

## Return on Investment

When I am initially asked to help an organization make a return of investment (ROI) case for investing in enterprise search, my initial request is to see business cases for other enterprise applications where the investment has been justified using an ROI calculation. This is ostensibly to find out the basis for calculating overhead costs and, in particular, the basis on which IT support costs are allocated.

Usually no such business case exists, or where it does, the calculation is based on vague assumptions, and yet approval is given for the investment. The justification is usually based on the proposition that without the investment, the organization will not be able to function, supported by the signature of someone on the board of the organization. Sadly, at present, no one on the board wishes to be the sponsor for enterprise search and knows little about the technology itself or the value of the technology to the organization. The request for an ROI seems to be a defensive measure in case things go wrong down the line.

Search is a high-touch application. It will be used personally by a substantial proportion of all employees in a service business, and even in a manufacturing business, enterprise search will support decisions being made which affect even staff on a production line. This cannot be said for a finance system, a customer relationship management system, or a treasury management system, as just three examples.

The main reason why an ROI cannot be calculated is that there are no standard processes involved in the way that for a finance application, the time for entering an invoice might be measured.

Making assumptions about the time that could be saved by investing in a good search application will not be founded on any sort of reality. It may well be the case that users spend very little time on the current search application because of its poor quality. It could well be that with a new application, users spend more time discovering valuable information, perhaps in repositories that had not been crawled in the legacy system. As a result, any business case that is based on time saved goes right out of the window.

Another problem that you will face with ROI is that it is very difficult to get pricing information for a new search application (see Chapter 12). In addition, the costs of the software and hardware are only a small proportion of the total investment, most of which is an investment in people who probably already work for the organization.

The reality is that if an organization needs a financial ROI to make an investment in search, then it fails to appreciate the value of information as an asset, and in addition thinks that there is just a single metric on which to judge the business case for search. In fact, there are multiple business cases, probably the same as the number of use cases set out in Chapter 3. Search has to be seen within the overall context of the business and its objectives, and that is why a full business plan is needed for search.

## Invest in Skills Before Software

Search does need to be planned. It is technically challenging, users have both high expectations and a high dependency on the success of search, and there will need to be a substantial investment in personnel for the search support team. As you read through this book, you will find there is just one single theme. I call it White's Rule of Search Investment:

> The impact of search on business performance depends more on the level of investment in a skilled team of people to support search than it does on the level of investment in search technology.

There is a corollary:

> Without an investment in a skilled team of people to support search, no matter how great the investment is in search technology, there will be no impact on business performance.

Over the last couple of years, much of the investment in collaboration applications has been justified on the basis that the organization is not working together as effectively as it should, and implementing a collaboration application will transform the situation. There is usually only anecdotal evidence about poor collaboration, and in due course, only anecdotal evidence about improvement. When it comes to search, the situation is the same. Someone (usually senior) has complained that they cannot find anything with the current search application and the organization needs to get something better.

A common feature of both collaboration and search is that it is very difficult to find out who owns the planning process. IT may hold the budget, but will usually have no information about how the application is being used on a day-to-day basis.

The bottom line is that for an organization in which search is business-critical, then whatever the size, there should be two people supporting the search operation. For smaller organizations, these two people may perhaps not be full-time employees, but there must be enough overlap to ensure that every working day there is someone on call to support people needing to rely on the search application.

It would not be too much of an overstatement to say that if you cannot find the people who will form the search support team there is really no point in making any investment in an enterprise search application.

To summarize, there are five search team roles:

- Search Manager, taking management responsibility for search delivery
- Search Technology Manager, looking after the IT elements
- Search Analytics Manager, running and analyzing search logs and search surveys
- Search Information Specialist, with responsibility for search quality
- Search Support Manager, providing training and user support

In the initial stages of an enterprise search project, these roles could be undertaken alongside other work, but once the implementation begins, these roles need to be filled on a full-time basis. There simply is no option. There is much more on search team management in Chapter 11.

## Stakeholder Analysis

The term *stakeholders* is common parlance in organizations, but usually little is done to analyze them in any formal sense. The matrix shown in Figure 9-1 can be useful in this respect.

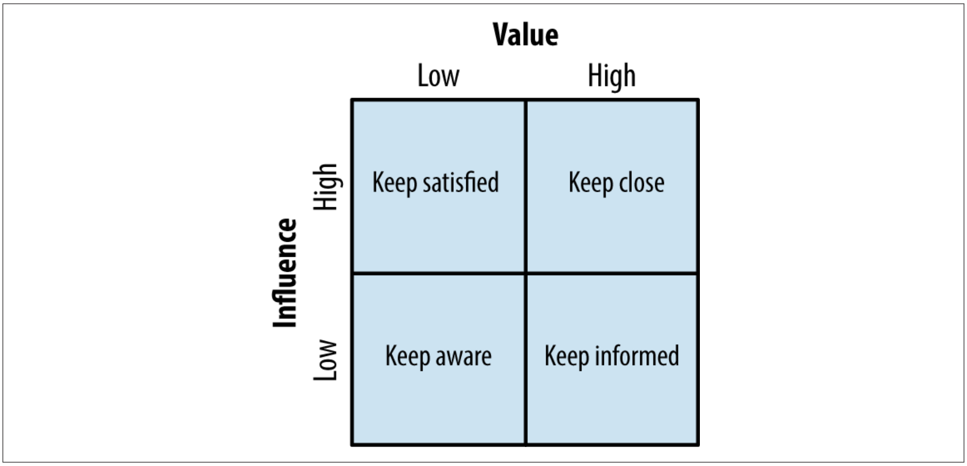


*Figure 9-1. The value and influence matrix*

The first step in creating this matrix is to brainstorm a list of potential stakeholders, and for each group, or individual, define the following elements:

- Name and position
- Potential positive and negative impacts on the project (Influence)
- What would the project expect the stakeholder to contribute? (Influence)
- What is the stakeholder’s expectation of the outcome of the project? (Value)

Once the stakeholders have been plotted onto the matrix, the stakeholder strategy starts to take shape. For the High Influence/High Value quadrant (perhaps a General Manager), a member of the project team should build and maintain a one-to-one relationship. The High Value/Low Influence quadrant would have intranet management, document management, and records management teams. These teams need to be kept informed and their views brought into the discussions of the project team.

The classic example in the High Influence/Low Value quadrant would be the entire IT department. Finally, in the Low Influence/Low Value quadrant would come the managers of most corporate departments who already have a lead application (finance, marketing, etc). The "keep aware" strategy is a blend of keeping them informed on a regular basis and being aware if they start to have concerns about the direction of the search project.

As the project proceeds, some of the stakeholders may need to be moved to a different quadrant, but the maximum effort needs to be expended on identifying the stakeholders in the High Value quadrants and delivering on their expectations.

## Business Impact

It is important to set out the objectives of the organization, as these could have a major influence on the way that search develops. The acquisition strategy is especially important, as this could require the search application to index substantial new repositories at short notice and also result in the negotiation of new license deals with a number of different vendors. There could be challenging divergences in metadata values and consistency.

One way of identifying ways in which the search application could support the business is to review the risks that are almost always published in the annual report, and work up a mitigating approach to each risk that involves the search application. In 2011, Hoffmann-La Roche, one of the world's leading pharmaceutical companies, identified five simple but challenging questions that employees were probably asking themselves and colleagues on a regular basis:

- Can I handle this?
- What is the implication?
- Can we find out sooner?
- Will it work?
- Have I chosen wisely?

If enterprise search can provide answers to those questions then the business case can be made on just a few pieces of paper.

As well as this top-down approach, the techniques set out in Chapter 8 will provide all the evidence that is needed as to how search can have an impact on business operations and the achievement of objectives.

Even though search is used very widely in the organization, it is virtually impossible to make a convincing business case across all, or at least, most employees. It is advisable to build the overall business case for investment on a number of individual business cases that resonate with senior managers. These might include improving customer service, shortening the time to prepare business proposals, reducing the time to develop a new product, or being more responsive to the actions of competitors. The metrics that will illustrate success will be different for each of these business cases, but will be grounded in business processes.

Nothing carries more weight than a story from a respected manager about how he failed to find information that could have made a positive impact on the organization. It's a trick used by many management authors, who use call-out stories to make an impact in an otherwise mundane book on some aspect of business operations. Beginning the business case document with a really strong search success or search failure (or both!) is a guarantee that readers of the business plan will already be predisposed to agree to the investment that is requested at the end of the document.

Some of the reasons given for investing in search include:

- Accelerate retrieval of information from known information sources
- Improving the reuse of information and knowledge
- Increasing the extent of collaboration through finding people with relevant expertise
- Eliminating information silos and the risk that important information was not being found and used
- Accelerating the speed of finding people both by name and expertise
- Raising the awareness of what was already known
- Eliminate duplication of work because relevant information could not be found
- Improving the consistency and quality of response to queries from customers and partners
- Creating a more personalized intranet solution
- Providing support for compliance management

Unfortunately, all of these are "soft" outcomes and are very difficult to quantify. One way of bringing the issues to life is to include some examples of poor searches. It often strikes me as odd that search business plans rarely provide screen shots of the current system along with a commentary in (a) the impact of search failure on business performance, and (b) the potential benefits of additional investment. That is why search needs to be owned by a senior manager who fully understands the potential impact of search on the business and also the risks of poor search performance.

## Search Owner

If building a search team is difficult, finding someone who will take business responsibility can be even harder. This is probably because there are no business and compliance-critical workflow processes that are supported by enterprise search. Look around at the main enterprise systems and they are owned by the manager responsible for the workflow: Sales Director, Operations Director, HR Director, and so on. In around 70% of organizations (based on the Findwise survey), the decision on a search application and the management of the application are the responsibility of the IT department. In many organizations, search is owned by Corporate Communications, almost certainly when the same search application is being used for the website and for internal enterprise search.

Here are two questions for you. In your organization, what percentage of the total amount of content being indexed is owned by either Corporate Communications or IT, and what percentage of the total number of employees work in IT and Corporate Communications? The answers will be small numbers, almost certainly less than 10%. IT should be delivering support services and certainly have an important role in search subsystem performance management. When the day comes that IT people regularly attend meetings with business units with supplier and customer-facing staff, then IT can own search. But not until that day.

In an ideal world, search should probably report to the senior manager whose performance bonus is based on meeting customer requirements, either through product development and delivery or service development and delivery. This could be a General Manager, or Director of Manufacturing. All that the search owner really needs to do is fight for a sensible capital and operating budget.

## Content

A search engine needs to be instructed about the content that needs to be indexed. The place to start is a content audit based around the repositories of information that the organization holds. There is no need to do a document-by-document audit, but just to look at the issues that will be important in ensuring that search works well.

The work involved in undertaking this content audit should not be underestimated. In the process of converting this list into an Excel database, any cell that is not completed could mean that the content is not indexed, or not indexed properly, and so becomes invisible.

## Owner

The first, and often most difficult, step is to find out who owns the repository that needs to be searched. It may have been set up some time ago, and the initial owner might even have left the company. If there is a current owner, chances are that the original intention of the repository has long since been overridden. This is often the case with a departmental repository where the department has been merged or fragmented over time.

## Scope

A brief description of the content should be prepared, along with a description of the categories of users who contribute and use the information in the repository. The total file size and total number of documents are important to know when it comes to sizing the search application. For the same reason, the rate of addition of documents by time will give an indication of how frequently the repository needs to be re-indexed, or whether the documents are such that they need to be indexed as soon as they are added to the repository.

## Document Size and File Formats

All the file extensions should be identified and listed out, and at a minimum the maximum file size of the collection should be ascertained. This is especially important in the case of videos, images, and other specialized files such as CAD drawings. Some file formats, among them Microsoft Visio and Microsoft Project, may not be easy to search. This might not seem to be a major handicap until you appreciate that identifying people working on a specific project or legal matter could be of assistance in identifying expertise.

## Language

Making the assumption that all the content is in a single language and that the language is English is only a safe assumption in a very few countries of the world. It could be that the French version of a document has been stored alongside the English version in what otherwise looks like a totally English-language collection.

Another aspect of language is the use of highly technical terms, product name conventions, and acronyms. Some years ago, I discovered that there were such things as nutrunners. These are constructed terms, and the way that they are deconstructed in the indexing process needs to be taken into account. Lawyers refer to *matters* as a noun and not as a verb, a further example of where novel uses of language need to be identified at the earliest opportunity.

## Technology

An important component of the technology section is to provide a list of current applications that already have search functionality, possibly as an embedded application. Examples might include document management and records management systems and enterprise resource planning systems. There are often more of these embedded search applications than most managers appreciate. These are often optimized for a specific application, and probably an enterprise application could not provide the same level of search performance and satisfaction. Having a list of these applications enables decisions to be made about whether there would be a benefit in providing a federated search environment.

There are a number of technology issues that need to be surfaced in a search business plan, and these include:

- The use of open source software
- Mobile access to enterprise information assets
- The adoption of cloud/software-as-a-service applications
- In-house versus external development and maintenance
- Relationship of the corporate website search, which may well be “independently” owned by the marketing department.

In most organizations, there will already be a number of search applications in use. It is easy to suggest that having just one powerful engine will solve all current search problems at a stroke. It could easily add to them. An important section of any business plan should set out whether or not a single search application is going to meet current and anticipated requirements. This is not just a technology issue, but has to be approached both from an IT and a user perspective. It may be useful to set out some scenarios under which the adoption of a single search application across the organization would be an appropriate solution. The reason for doing this is that at some time in the future, someone will raise the issue about why the organization seems to have so many disparate search applications. At that time, the analysis can be reviewed to see if indeed there is a case to be made.

As with any software application, a search vendor will release versions of the software to either address bugs or to provide additional functionality. This section of the business plan should set out the basis for considering whether to implement a new version of the software, bearing in mind that there could be risks with connectors to other applications.

## Infrastructure

Enterprise search can have some challenging infrastructure requirements as far as storage, in particular, is concerned. The topology of a large-scale enterprise search application with good disaster recovery and the minimum latency on queries will need careful planning. The issues will not just be about the size of the index relative to the size of the repository but also the write speeds of the disk arrays. Many capacity planning specialists will be in novel territory when it comes to planning search capacity.

Almost certainly there will be a need for test, development, and production servers. For large-scale enterprise search applications, there will need to be multiple production servers with distributed indexes.

Network bandwidth to distant but still important offices can also present issues that need careful review. Substantial files could be downloaded very rapidly for perhaps 10–15 minutes as a user works her way through the top 50 results looking for a specific piece of information. This can be a particular issue with PowerPoint files, and a number of vendors offer a feature to render the document, or PowerPoint file, as a small HTML thumbnail image.

## Disaster Recovery

It is tempting to put search at the bottom of the disaster recovery priority list, but arguably, it should be right at the top. It may enable the organization to keep going while other applications are brought back to life. After all, the search index will contain a copy of most, if not all, of the information that the organization possesses, and if the application can provide users with an HTML thumbnail of a document, that could be more than enough for business as usual to continue.

A disaster recovery plan usually sets outs a recovery time objective (RTO) defining the maximum application downtime and a recovery point objective (RPO) noting an acceptable restore point. For an enterprise search application, these need to be considered from basics rather than the blind adoption of objectives from other enterprise applications. It is not just a case of getting the application up and running from a user perspective but understanding and accounting for content that may not have been completely crawled or an index that has not been correctly updated. The index itself may have been distributed around the world, and with disaster recovery will come the need to re-synchronize the indexing process.

Sometimes changes to the search application will require either a partial or complete re-index and it is very important to know how the system will respond to one or more of the index servers crashing during the process. It is not uncommon for a re-index to take days rather than hours.

## Access Permissions

Organizations are rightly very concerned about the risks from employees finding information that they are not entitled to see. Even if there is a corporate security policy, you must question whether it is granular enough and implemented rigorously enough to ensure that ACLs can be created and maintained. The potential impact on staff and the reputation of the organization from a failure of the security policy could be very dramatic.

This is especially the case if the document being created is being indexed as soon as it is saved to a repository. At that moment in time, there is in effect a duplicate of the content accessible to anyone with the correct security permissions. The index will almost certainly be backed up on a second server for disaster recovery purposes. Removing the document from the repository will almost certainly not remove the content from the index. If the document was a list of senior executive salaries, then a search for this information might well disclose the amounts even if the document has been deleted.

## Metrics

The performance analytics that can be generated from a search application can be overwhelming. The challenge is in deciding what would be key performance indicators. Chapter 15 sets out five categories of metrics:

- Technical, mainly IT metrics such as up-time
- Query, the analysis of query terms and the time spent on reviewing search results
- Discovery, including the extent to which highly relevant results are posted on the first few pages
- Satisfaction, which at a base level is an assessment of the degree of confidence in the outcomes of a search
- Impact, which reflects how the search application has resulted in improvements in operational performance

For planning purposes, a discussion around a selection of these metrics that can be reported up the management line can reveal differences between the expectations of stakeholders that need to be addressed at the outset of the project.

## Metadata and Taxonomies

A few years ago at a search conference, there was a presentation about a new search implementation. The search manager reported that one of the tests that had been run during the implementation was to find members of staff called Jane. To everyone's surprise, most of the high relevance results were to male employees. It turned out that they had all written and submitted their CVs on a template owned by someone called

Jane, and the search engine was placing more value on this metadata item than on the name field.

The problem with metadata is that the content contributor has to add it, and does so either with reluctance, or without due care, or a combination of both conditions. This section of the search strategy needs to highlight the importance of metadata and how it will be generated, either automatically (e.g., the name of the content contributor from the system login information) or through manual addition. Entity extraction is a halfway house.

If the content has been contributed through a content management or document management application then there will probably be good metadata tagging. If it is just a shared file server then even basic folder metadata could be inconsistent.

All the evidence points to the benefit of a taxonomy and metadata enhancing search performance, and especially in presenting highly relevant information. However, taxonomies are time-consuming to compile and to maintain. As with so many search-related issues, a balance needs to be established between the value of the taxonomy and the benefit to users, taking into account that the users of the information may not be the people who have to add the taxonomy metadata in the course of saving the document. The Findwise 2014 survey indicated that search performance was significantly better by at least 10 percentage points in organizations where there is a managed taxonomy and a metadata schema.

The development of a thesaurus is covered by ISO Standard 25964, of which Part 1 is specifically about the thesaurus development for information retrieval. Developing organizational vocabularies is a skilled task and may well need to be outsourced to specialist consultants. However, elements of even the best constructed vocabulary may change on an unpredictable basis as the organization expands its operations, and it is important to plan for how these will be accommodated in the vocabulary.

An associated element of metadata management is whether there should be an investment in software tools that can create classifications from a textual analysis of content. Examples would include SmartLogic and Concept Search, and some search applications (e.g., from BA Insight and Recommind) include classification applications.

## Help Desk

A search application needs its own help desk, even if it is a virtual one, and there needs to be a service-level agreement both ways between the IT and Search Help Desks because it may take quite a bit of effort to work out what is causing the problem and what actions should be taken to remedy the problem.

## Usability and Accessibility

Despite the high-profile efforts of usability experts such as Jakob Nielsen, few organizations seem to take usability seriously. Search usability testing is especially important because of the complexity of many search user interfaces with a profusion of filters, facets, annotations to results, and perhaps even graphical representations of clusters of search results. The dichotomy is that everyone wants a highly usable search application but is unwilling to fund either internal or external resources to undertake usability testing on a regular basis.

## Training and Support

The view is sometimes taken that search should be so intuitive that there should be no need to providing training and support. This view is often based on the "simplicity" of the Google search box, a view that ignores that books have been written on the very wide range of hacks that are available to users of Google search.

The same is true of an enterprise search application. Certainly there should be as few barriers as possible to carrying out a basic search, but in an enterprise context, there are probably very few basic searches as finding most, and ideally all, of the relevant information is very important.

## Risks

It is always advisable to have a risk management strategy for enterprise search, and this probably needs to account for the following risks:

- Lack of resources in the search team leading to poor search performance
- Search manager leaves and there is no internal candidate
- Search vendor is acquired or goes out of business
- Search vendor is unable to provide an adequate level of support
- Lack of expertise in specifying open source search applications
- Key development skills in the open source contractor are not available
- Inadequate security management leads to a breach of access permissions
- Disaster recovery procedures prove to be inadequate
- Enterprise networks are giving rise to significant performance problems
- Poor performance of connectors and APIs
- Best bets are no longer best bets

## Summary

Writing an enterprise search strategy or a business plan is not an exercise that can be completed in a few days sitting at a desk. Of the list of topics covered in this chapter, the most time consuming will be the content audit. I find it difficult to understand why organizations do not document their search strategy, and then maintain it on at least an annual basis. As with any major IT project, the more work undertaken in the planning stages, the lower the risk and the greater the benefits post-implementation.

## Further Reading

Marti A. Hearst, *Search User Interfaces* (Cambridge, UK: Cambridge University Press, 2009).

Peter Morville and Jeffery Callender, *Search Patterns* (Sebastopol, CA: O'Reilly, 2010).

Tony Russell-Rose and Tyler Tate, *Designing the Search Experience* (Burlington, MA: Morgan Kaufmann, 2012).

Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, and Susan Dumais, "Slow Search: Seeking to Enrich the Search Experience by Allowing for Extra Time and Alternate Resources," *Communications of the ACM* 57:8 (August 2014): 36–38.

# CHAPTER 10
# Defining User Requirements

All effective systems are based on a good understanding of user requirements. “We want it to work like Google” is an aspiration and not a user requirement. In this chapter, a range of approaches is suggested to help define user requirements. There is no single approach that is better than the others and usually a blend of several is required. However, a balance needs to be kept. At one end of the spectrum is the Google approach, in which innovations are tested out on customers and if there is a positive reaction then the innovation becomes a Google product. Apple is at the other end of the spectrum. The late Steve Jobs commented that Apple needed to provide customers with what they wanted even though they didn’t know what this was.

The general lack of support for search invariably means that little attention is paid to defining user requirements, and all too often changes to either a user interface or the implementation of a new search application are largely based on anecdote and hearsay.

The value of user research is not just in defining the requirements for technology but also in setting a benchmark that can then be used in the future to prioritize search enhancement activities.

In this chapter, some of the techniques that can be used to define user requirements are presented. These may help define perhaps 80% of what is required. The remaining 20% will only be discovered over time, and some proportion of the 80% will be found not to be of value. This is because:

- The organization itself will change over time, giving rise to new requirements and making others less important.
- As users become competent in using the search application, they will start to push the boundaries of what is on offer.

- Software upgrades will offer new search functionality.
- As new content sources are indexed, additional functionality may be required to optimize search performance.

Fortunately, search applications are well suited to being modified and enhanced to meet emerging requirements, unlike many other enterprise applications where a change in business practice may require substantial and costly changes to be made.

Determining user requirements for any enterprise application takes time and effort, but search requirements are an order of magnitude more difficult to elucidate. The primary reason is that most enterprise applications are based around workflows and procedures (e.g., authorizing payment of an invoice) that are well understood and well documented. Business analysts will document every process and prepare a business requirements document. This will then form the basis of a functional requirements document. The software application will be built on the basis of the functional requirements, and tested against them post-build. There may be some limited user testing but rarely does a good user experience feature in the business requirements.

This is not the case with search, which is probably being used because other workflow-based applications are not providing the information being sought. To extend this example, a manager may be looking for the policy on the levels of authorization authority, perhaps in an office in a different country where a project is being managed by one of her direct reports. Rarely is this information in the finance system! Looking at the navigation options in the intranet may also present some challenges. Will the policy be under Finance, Policies, or Germany, as just three examples. Trying to understand why and how people search is uncomfortably close to mind reading!

The second aspect of user requirements is understanding what is fit for purpose rather than fit for specification. Experienced managers may already have the organization's guidance note on authority levels, and may simply want to make sure it is the latest version. They are able to query on the title and will have a very good expectation of what they anticipate being able to find. Indeed, not finding a more recent version would be a successful outcome. For less-experienced managers, who may perhaps be new to the organization, the query is more complex, as they may want to look at a range of policy documents before making a decision so that they have a context for their next action.

Defining user requirements is often overlooked in situations where search is just one element of an application, such as the implementation of a new CMS or a packaged intranet solution, or a migration from SharePoint 2010 to SharePoint 2013. The focus is often about how existing content will fit in a new information architecture and not until the last minute (if at all!) is any attention paid to how user requirements for search are going to be determined, incorporated, tested, and delivered.

All work on defining user requirements should always take into account not just functionality and features but also the content that users want to search. It may well be the case that because a particular repository has not been crawled and indexed, it is invisible to users.

## Why User Requirements Are Important

An initial reaction might be that there is only one user functional requirement, and that is to have a search box in which to enter a query. After all, that is all that the Google interface does! The reality is that it does far more than offer a search box. On the top righthand side of the screen are options to search for Images, together with a rather unusual 3x3 matrix that offers more search options. These include Scholar and Video Search. Google has clearly worked out that second to a text search, its users will search for Images, and that is why that option is a direct click rather than the drop-down menu from the matrix.

Because search is a dialogue, the user interface should provide a way for the dialogue to be conducted. Normally this is through filters and facets, and this is where user requirements need to be determined with some care. There may only be space on the screen to display three or four facets and filters. A common filter is to offer a searcher the option to select results by file type (e.g., Word, PDF, PowerPoint, Excel, etc.). In some cases, this may be a useful option, but in others, those options are of little value. One of the most important outcomes of user research is to define what would be helpful filters and facets. This is not only to support the design of the user interface in due course, but because the content may need to be tagged specifically for the search application to be able to generate the facets.

Another common approach is to enable the search to be restricted by date, with options such as This Week, This Month, and Last 3 Months. The benefit of offering these options is substantially increased if there is also an option to sort results in reverse chronological order. The number of new items added may be too few to make a This Week search useful, especially if the index is only updated over a weekend so that in effect only the previous week's content is being displayed.

Once these user requirements have been assembled and incorporated into the design of the application, the requirements also form the basis of the user testing that is so important in delivering a high-quality search experience.

## Information-Seeking Models

For well over 30 years, there has been a great deal of research into trying to understand how users go about seeking information. It is beyond the scope of this book to try and summarize all these models. Some of them have intriguing titles such as berry-picking, information foraging, information scent, and orienteering. There is a good summary of these by Marti Hearst in her book, *Search User Interfaces* (Cambridge University Press), and Peter Morville and Jeffery Callendar take a fresh and pragmatic view of information seeking in their book, *Search Patterns* (O'Reilly). What you will gain from reading about these information-seeking models is that what is being attemped is the reduction of complex cognitive processes to a single process that can be evaluated in practice.

An especially interesting approach has been taken by Tony Russell-Rose and Tyler Tate, and is described in detail in their book, *Designing the Search Experience* (Morgan Kaufmann). They have taken three top-level categories of information discovery proposed by Garry Marchinioni and added to them a set of what they describe as modes:

*Lookup*

1. Locate: to find a specific (possibly known) item
2. Verify: to confirm that an item meets some specific objective or criterion
3. Monitor: to maintain awareness of the status of an item for the purposes of management or control

*Learn*

1. Compare: to examine two or more items to identify similarities and differences
2. Comprehend: to generate independent insight by interpreting patterns within a data set
3. Explore: to investigate an item or data set for the purpose of knowledge discovery

*Investigate*

1. Analyze: to examine an item or data set to identify patterns and relationships
2. Evaluate: to use judgement to determine the value of an item with respect to a specific goal
3. Synthesize: to create a novel or composite artifact from diverse inputs.

The authors suggest that these search modes do not occur randomly, but cluster to form distinct chains or patterns. In the case of enterprise search, these mode chains might be:

- Comparison-driven search (analyze-compare-evaluate): "Replace a problematic part with an equivalent part without compromising quality or cost"
- Exploration-driven search (explore-analyze-evaluate): "Identify opportunities to optimize the use of tooling capacity for my commodity/parts"
- Strategic insight (analyze-comprehend-evaluate): "Understand a lead's underlying positions so that I can assess the quality of the investment opportunity"
- Strategic oversight (monitor-analyze-evaluate): "Monitor and assess commodity status against strategy/target"

- Comparison-driven synthesis (analyze-compare-synthesize): "Analyze and understand market trends to inform brand strategy and communications plan"

It is important not to take these as a "checklist" of use cases, but to read the book and understand the research and analysis that went into their development.

What these modes do illustrate is the range and complexity of information seeking, especially in the enterprise, and therefore why a significant amount of resource should be allocated to understanding user requirements.

My contribution to the discussion about information seeking is rather simplistic. I refer to it as the Eureka! Triangle (see Figure 10-1).

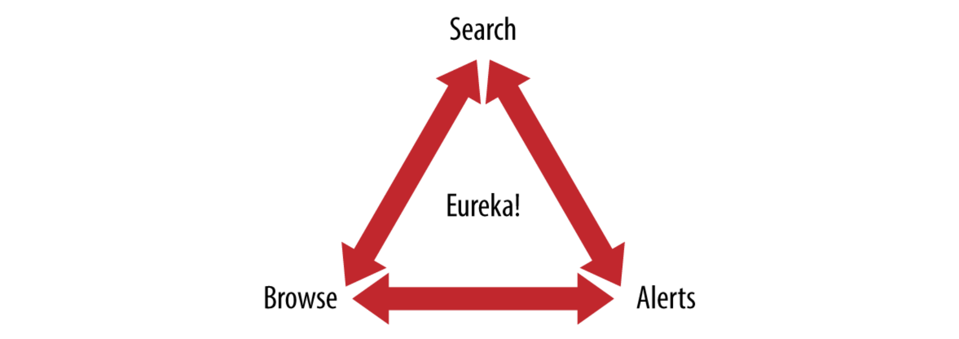


*Figure 10-1. The Eureka! Triangle*

When looking for information, we will use the processes of browsing through the navigation of an intranet or folder structure of a document management system, of searching, and of being able to set up alerts either from RSS feeds or search profiles running in the background. These need to be kept in balance. In the case of intranets, there can be such a focus on information architecture that when in usability tests someone uses search as an option, the intranet team may feel a sense of failure. The same applies to document management systems. Step outside to the web world, and organizations invest substantial amounts of money in designing home pages only to find that a significant number of site visitors arrive via Google and Bing and start deep inside the website and possibly never see the wonderful carousel on the home page.

## Spreading the Message

Before any user requirement work is undertaken, it is essential to have a good communications strategy that keeps everyone informed about the progress of the project. It could be that after a lot of user research the outcome is that there is no clear business case for an investment in a new search engine. As well as managing the expectations of all the stakeholders, a news item on the intranet should make a point of inviting employees new to the organization to come forward and talk to the project team. The reasons are twofold. The induction period is always stressful, and it is likely that newcomers have stress-tested the current applications. The second reason is that they may have experience with other search engines and come with a different set of expectations about a good search experience.

In many countries, notably in Europe, internal surveys need to be managed with a high degree of sensitivity to data privacy. This is especially the case in organizations that have Workers Councils, in particular Germany, and any corporate-wide surveys should be discussed with representatives of the Workers Council at the earliest opportunity. The Council may well have views on how the survey should be communicated to employees.

# User Requirements and User Satisfaction

The work carried out on defining user requirements is also of significant value in assessing search performance. From the outset, the choice of a user research approach and the way that it is carried out should also take into account the potential use of the approach in search evaluation. If a survey of requirements is going to be conducted, then the questions should be chosen so that at least some of them provide benchmarks for performance assessment in due course.

The following subsections describe ways in which user requirements can be determined, listed in alphabetical order.

## Climate Surveys

Many organizations carry out what are often referred to as "climate surveys" to assess the attitudes of staff toward culture, management approach, and operational issues. These surveys are usually carried out annually. In view of the importance of information in making decisions, there should ideally be two prompts about information management that users can rank on a Likert scale.

> "I feel confident that I can find the information I need to meet the objectives I have been set."

> "I always feel certain that the information I find is the best and most current available."

Prompts such as these are very helpful in assessing the post-implementation success of the search engine. If the current level of satisfaction is 60%, there is certainly going to be room for improvement. However, it may take a lot of negotiation with the "owner" of the survey (usually HR or Internal Communications) to add in one or two questions about information management.

## Diaries

Asking people to maintain a diary of their search experiences can provide valuable information, but the design of the diary sheet needs to be developed with care, and with some pilot trials. Expecting people to complete a diary on a daily basis for a period of time is not realistic. This is at best a dipstick test to see if there are any outlier search requirements that have not been identified using other techniques.

The information that could be collected in diary entry would include:

- The reason for the search ("Needed to find the latest version of the security policy")
- The search terms used (*security policy*)
- How many results were returned?
- Were you successful in finding the information, and how long did it take you?
- If you were unsuccessful, what did you do next?

The best way to get useful outcomes is to agree with volunteers that perhaps just two days in a specific week they are going to use the diary, perhaps when they are planning an internal presentation or preparing a project report. A quick telephone call during the course of the day to be supportive will be welcomed by the volunteers as will a public acknowledgment of the role that they have played. These volunteers in particular would be a good set of participants in later proof-of-concept or implementation tests. It is very important to ensure that participants do not see this as an investigation of their productivity.

## Decisions

It can be useful to understand the decisions that need to be made on a regular or ad hoc basis where search is likely to play an important role. One example is in setting up a project. Employees with specific experience may need to be identified, work that has already been undertaken by the organization reviewed, and the latest versions of project management policies and procedures completed. Discussions with some senior project managers can shed light on the information that needs to be identified.

## Focus Groups

It can be very tempting to run focus groups. The logic is that getting together a group of people who make extensive use of search would be a good way to start to develop a set of requirements. However, it is highly likely that these people would be able to use almost any search engine and get the best out of it. Providing a good solution to people who find the current search application untrustworthy or difficult to use is just as important, but it can be very difficult to find potential participants.

There is usually pressure from senior managers to set up some focus groups. These rarely have the desired effect, as the participants may be unwilling to highlight problems that they find in obtaining and using information lest the other participants mark them down as incompetent. Running a focus group also requires two people, one to facilitate and one to record the comments, so some of the potential gains in interviewer time are already at risk. Then there is the challenge of making sure that all the participants turn up, so that the group is representative of a group of employees. Having someone miss the meeting and then insist on having an individual interview again wastes time and delays the conclusion of the project.

It is probably better to use focus groups later in the requirements-gathering process to validate some initial outcomes than to use them as an initial source of requirements.

## Help Desk Calls

A review of help desk calls is a very important part of the user requirements gathering, even if there has not been a specific search help desk in the past. The help desk tickets may reveal many points of failure, even if rarely points of success. It is also important to bear in mind that reducing calls to help desks is important in terms of employee satisfaction and help desk productivity.

## Interruptions

Many organizations encourage employees who are trying to locate information to make contact with others in the organization who may be able to help. However, there is a downside in that when we are interrupted by someone seeking our expertise, we rarely continue on the same task but sit back and re-prioritize our list of pending items. As this book was being written, I was interviewing a senior lawyer in a global law firm. His comment was to the effect that he was embarrassed by the number of times he had interrupted the work of his colleagues to track down information that the search application was not able to find.

As a result, there can be a productivity hit for the organization. Asking experienced staff about whether there are any patterns in the requests they receive for information and advice may reveal a systemic problem in employees finding certain types of information.

## Microsoft Product Development Cards

In 2002, Microsoft user experience researchers Joey Benedek and Trish Miner developed a set of 118 adjectives that could be used to define usability in test situations. These adjectives are often used in the initial stages of an intranet or website implementation but are just as relevant in the early stages of defining search requirements.

Some of the adjectives in the list are directly relevant to search, including:

- Comprehensive
- Convenient
- Customizable
- Easy to use
- Fast
- Secure

The approach is especially useful when trying to understand the good and bad points about a current search implementation. There are various ways of using these terms in the process of starting to define user requirements. Ideally, each word should be written on a card, and a set of cards given to small groups of users. The number in each group should be no more than five, because the objective is to get a discussion going about the terms that best describe the current search application, and the terms that should define the re-launched search. Initially, each group should be asked to select eight cards for the current search application, and then in a second run for the new application. Once eight have been selected, then the groups again might be asked to bring the total down to five.

This approach is highly qualitative and its value is more in starting to gain the involvement of users than in developing a checklist of requirements based on the final outcomes of the card sorting tests.

It is possible to carry out this process remotely, just asking people to highlight the descriptions they have selected, but the best results are gained from a number of groups working together, presenting their results, and then having a short discussion about the similarities and differences among the group results.

It is important to position this as a “fun” process that is just one input into defining the overall user requirements.

## Mobile Workers

Despite the widespread use of mobile users, the particular requirements for searching enterprise information from a tablet or other mobile device are often overlooked. Employees using mobile devices for this purpose may well have found workarounds for inadequate support from the enterprise. They may be reluctant to respond to a survey asking for their requirements in case this discloses to their managers processes that may not be in line with organization policies. Face-to-face interviews may be the only option.

In the era of responsive design, it is very tempting for the design team to create an intranet design that can be reproduced on a mobile device. It is arguable that in every case, the entire intranet should be delivered over mobile devices. In the case of search, it is essential that the user research process looks very carefully at what content users need to search for, how they will go about entering the query, and how they want the results presented.

For good reason, filters and facets can be useful ways of managing long result lists, but the end result is that the user interface extends across three columns on a desktop display. Compressing this interface onto even a tablet may cause problems because the user will have to invoke the virtual keyboard that may then obscure perhaps half the available screen space. Users will also be reluctant to open up a series of documents that have been presented as relevant results and scroll through them to find the information they need.

Search vendors seem just to be ignoring the problems. Search for *mobile* on the website of any of the vendors and you will be very disappointed with the response. That is no excuse for understanding what users want from a search application on their mobile device, nor is it an excuse for providing the desktop search experience on a mobile device. User expectations are sure to increase as Siri and Cortana provide greater functionality. They will then start asking, "Why can't our search be like Siri?"

## Personas

A widely used technique in the design and development of websites and intranets is the use of personas. A persona is a fictional person who represents characteristics of a group of people with similar requirements for information to undertake tasks.

Personas bring many overall user-focus benefits, including:

- Users' goals and needs become a common point of focus for the team.
- The team can concentrate on designing for a manageable set of personas, knowing that they represent the needs of many users.
- By always asking, "Would Anne use this?" the team can avoid the trap of building what users ask for rather than what they will actually use.
- Design efforts can be prioritized based on the personas, and so design and project creep can be managed.
- Disagreements over implementation decisions can be sorted out by referring back to the personas.
- Implementations can be constantly evaluated against the personas, where appropriate, using business end users who were involved in the development of each of the personas.

The usability consultant Donald Norman sums it up well:

> Do Personas have to be accurate? Do they require a large body of research? Not always, I conclude. The Personas must indeed reflect the target group for the design team, but for some purposes, that is sufficient. A Persona allows designers to bring their own lifelong experience to bear on the problem, and because each Persona is a realistic individual person, the designers can focus upon features, behaviors, and expectations appropriate for this individual, allowing the designer to screen off from consideration all those other wonderful ideas they may have. If the other ideas are as useful and valuable as they might seem, the designer's challenge is to either create a scenario for the existing Persona where they make sense, or to invent a new Persona where it is appropriate and then to justify inclusion of this new Persona by making the business case argument that the new Persona does indeed represent an important target population for the product.

However, be aware that intranet personas may not be appropriate to the requirements of enterprise search, and it is advisable to develop a set of search personas that drill down into search requirements in more detail.

Figure 10-2 shows an approach I have used to segmenting user requirements into four broad categories, each of which could be represented by one or two personas.

The term *current domain* is used both in an organizational sense (my current business unit) and in an expertise sense (I am a chemist). A *novel domain* could be someone moving to a new business unit, or taking on different responsibilities, such as a research chemist taking on a business planning role. Precision and recall should not be taken as absolutes but as indicating either a requirement for a few specific documents or for a much larger group of relevant documents.

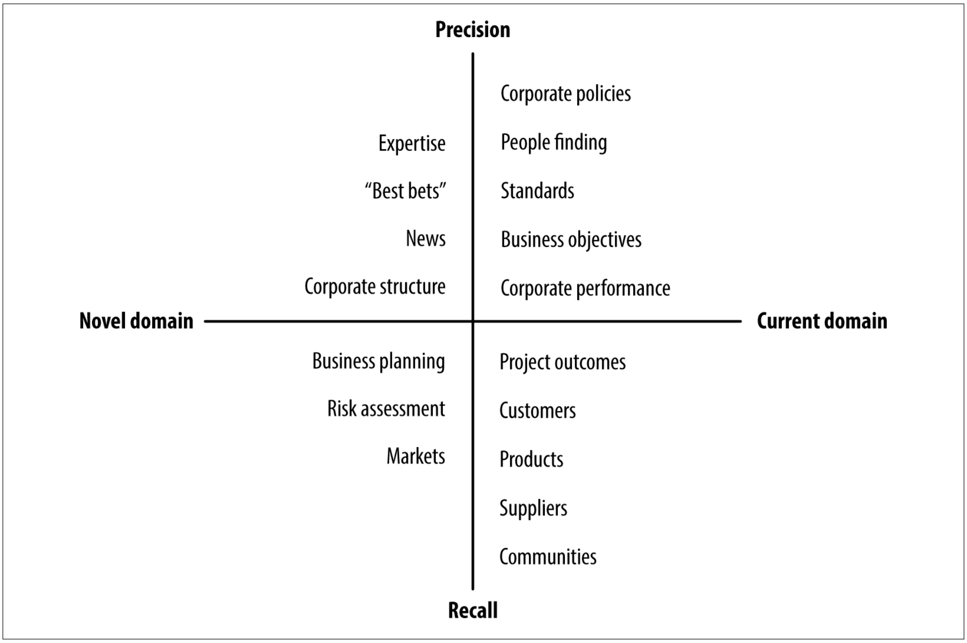


*Figure 10-2. Four categories of user requirements*

## Team Meetings

One of the critical success factors in search is gaining an understanding of the user context. Search logs may disclose what search terms have been used, but not why they were used.

Every organization has team meetings, though increasingly these are virtual team meetings, which require substantially more planning. Teams tend to have regular tasks, such as providing monthly status reports on new projects, revising corporate policies, and tracking the activities of competitors. Sitting in on these meetings can help identify the types of searches that are carried out and what would be the desirable outcomes of the search process. The benefit of teams over focus groups is that team members will feel comfortable with one another and have a collective focus on certain corporate objectives that may well determine career development opportunities or compensation awards.

However, there is no point in just turning up at the meeting and asking for input on search requirements in the Any Other Business section of the meeting. The program of attendances at the team meetings needs to be highlighted on the intranet. It is also important to have the discussion about search fairly high up on the agenda, so that it is positioned as an important topic. Having the discussion on the agenda also (hopefully!) ensures that attendees come prepared.

As mentioned before, teams increasingly work and meet on a virtual basis, and this requires more preparation, as the attention span of participants may well be lower when taking part in a meeting which may have been scheduled at a time that is not totally convenient for them. On the positive side, as the attendees will be participating through a networked computer, it may be possible for them to demonstrate some of the aspects of the current search application that they would like to see enhanced.

Always offer members of the team the option to talk privately about their search experience and requirements. They may not wish to disclose to their colleagues that they are having difficulty with the search applications.

### Usability Tests

Sadly, in many organizations the resources to carry out usability studies are very limited, and often there are no corporate usability specialists. Work on the usability of the corporate website may well have been outsourced. Using external expertise is not ideal for internal applications because a good understanding of the business is needed in both establishing the tasks and interpreting the results.

There is a lot of debate about how many participants should be used for each test. Jakob Nielsen suggests that five participants will highlight most of the main issues with the search application, and for the purposes of gaining an indication of user requirements for the specification of a new search application, that is probably a good number to aim for.

## Establishing Use Cases

A use case is defined as a list of steps defining interactions between a user (sometimes referred to as the “actor”) and a system to achieve an objective. There is no “correct” way to present a use case, and the use cases set out here are very informal ones. However, they can be useful in starting to translate user requirements into a specification, something that is more difficult to do with personas. Any given employee may display many use cases.

The 10 use cases set out here are very pragmatic, based on my observations of people at work in organizations. They are deliberately set out in alphabetical order as there is no single or set of use cases that are more common or more important than any of the others. The use cases have titles that should be recognizable in most organizations.

The objective of listing out these use cases is to help shape the questions that need to be answered out of the user research.

## Analysis

It is quite common in organizations to look for trends in performance, which could be financial, or measured in more complex key performance indicators (KPIs). To undertake this analysis, a user may want to find a defined set of reports, and some or all of these may contain a substantial element of numeric data. This is the area of content analytics and data/text mining, and on the edges of business intelligence.

## Compliance

In this use case, there is a requirement for high recall to verify that all the critical information has been identified. Although this is typical in a compliance situation, it can also occur when there is a need to locate all the project reports on a defined project, or all the products that use a specific chemical over which there is a concern about poor quality standards.

## Expertise

The need to locate people, and in particular people with expertise, is often overlooked in designing search. All too often there are two search boxes, one for *search* and one for *people*, which is unhelpful when the user is trying to find out about who knows someone, or even who knows which are the relevant documents. Many searches are carried out in an effort to find people with relevant expertise, and not just for the document itself.

## Induction

In many organizations, it is not unusual to have a staff turnover of more than 10% per annum, and there is sometimes a specific area of an intranet that supports early-stage induction into the organization. In addition, there are many employees who will take on new roles and responsibilities during the year, perhaps in a different office or even in a different country. An important issue here is whether the search application will be able to provide some form of either a best bet so that the results of a search can be placed in context, and/or some tagging from other users which rates a document of being of particular value.

## Item

The user's search will only be satisfied by finding a specific document, perhaps a presentation to a team or a project wrap-up document.

## Learning

A feature of the learning persona is that the user is not at all sure about the best way to frame a search query. She may be seeking information on the work that the organization has undertaken to reduce its carbon footprint, and this could be covered by a very wide range of terms, from corporate social responsibility to green engineering.

## Mobile

The most basic element of the mobile persona is that the user will be using a screen format that is smaller than the average desktop. The more difficult elements are the authentication that may be required, the inability to print out the results of a search, single tasking resulting in the need to open a different application to read the item listed in the search results, and the way in which the query is formulated. This formulation could be heavily dependent on location if GPS is used as a background search criterion, something that may not be apparent to the user, or even useful if the implicit criteria is not relevant to the search.

## Monitor

The main characteristic of this persona is that the search requirements are fairly consistent over a period of time, and the ability to be alerted to new information as soon as it has been indexed is usually very valuable.

## Product

When a user is searching for information on a particular product or service, either as a basis for internal review or to meet the requirements of a supplier or customer, then a near miss is not good enough. If product code AC34-345-12 does not appear on the first page of search results, then the user has a problem on his hands.

## Task

Supporting standard tasks should be an important role for a search application, but few companies have any firm idea of what a task involves if it is not embodied in a workflow process. Understanding the information content of a task is going to be increasingly important in speeding the decision-making process, and many organizations and search vendors are looking with considerable interest at search-based applications.

One-on-one interviews with employees can often uncover surprisingly complex tasks that depend on accessing multiple information sources. An example might be to set up a project team. This may require finding information on:

- The procedures for setting up a project
- Finding out if a project of this type, or for this client, has been carried out previously
- What forms need to be completed and forwarded to other departments

- Who the members of the project team should be
- The current availability of the prospective team members
- Internal guidelines on this particular type of project
- The project progress reporting procedures

## User Interviews

It is very easy to spend time interviewing users and end up with little relevant information. This is because the interview can easily move away from the core subject of the interview and get into specifics of design and content that are then difficult to scale up to a set of user requirements.

In setting up user interviews, it is easy to think in terms of departments or roles, but in specifying search requirements, some lateral thinking is called for.

Some important categories of users that are often overlooked in the interview program include:

- Personal assistants to directors and senior managers
- Employees who have recently joined the organization, not just because they will be coping with the usual induction issues but also because they may have experience of how search is delivered in their previous organization
- Employees with a background in the sciences, law, and medicine, who will be familiar with large-scale information systems from their time at college, and during the course of their careers

In conducting interviews, I have found the diagram shown in Figure 10-3 to be of value in getting the discussion going.

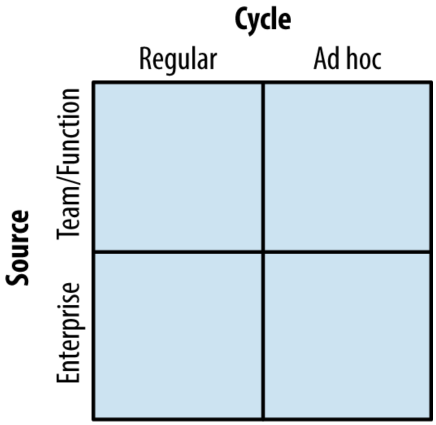


*Figure 10-3. Cycle and source diagram*

The objective is to gain an understanding of information gathering that is carried out on a regular basis (and could be supported by search alerts) and ad hoc requirements that are almost always carried out under time pressures. This diagram also distinguishes information that has been collected and is under the management of a team or department and the need to discover information that may be anywhere in the enterprise.

I encourage interviewees to write on the diagram, and I collect these as I go along. In many cases, the interviews have to be carried out by telephone and sending this diagram in advance with a brief description of its purpose enables me to get quickly into the interview without wasting time. It is possible to let a face-to-face interview extend to 50 minutes, but a telephone conversation needs to be limited to 30 minutes.

Before starting any program of interviews, reading Steve Portigal's book *Interviewing Users: How to Uncover Compelling Insights* (Rosenfeld Media) is an essential first step.

## User Surveys

Conducting user surveys with web-based survey tools has transformed the effort required to carry out large-scale surveys and have the results available in a short period of time. There are some important guidelines that should be taken into account in designing the search survey:

- Start out with no more than 10 questions, which will probably take a user around 10 minutes (or a cup of coffee) to complete. Anything longer will need very careful design.
- The questions should be intuitive, so that respondents gain an immediate understanding of why the question is being asked.
- Ideally provide an indication of how far through the survey a respondent has reached.
- Don't ask questions that rely on feats of memory about what the respondent did over a past period of time. "Do you use search now more than you did a year ago?" has no value at all.
- Don't expect respondents to write essays in a text box. Invite respondents to contact you if they would like to talk through issues in more detail.
- Recognize that it may be better to send out different surveys to specific user groups than try to accommodate the views of the entire workforce with a single set of questions.
- If using Likert or Likert-like surveys, do not average out the scores. Use the median.
- Commit to summarizing the outcomes by a given date, and invite respondents to comment on the results.
- Test the survey, and then test it again.

For more guidance, turn to *Surveys that Work* by Caroline Jarrett. As with user interviews, there is a substantial body of good practice about the conduct of surveys. You are only going to do it once, so it is advisable to do it properly. The future of the organization could depend on the outcomes.

## Search Benchmarking

If the aim of an enterprise search project is to improve search performance, it is important to benchmark the current application. Great care is required to ensure that the test searches that are carried out are directly comparable with those undertaken initially in the proof-of-concept tests (Chapter 13) and then after the implementation (Chapter 14). The search queries need to be "real" queries, not just queries dreamt up over a cup of coffee by the project team. The content scope should also be defined; perhaps all documents associated with a particular project or product launch. This collection is sometimes referred to as the Gold Collection or Golden Collection, as it will be used on a regular basis. Not only is this collection of value in benchmarking the current application against the new application but also in assessing the impact of changes that are made to the ranking parameters.

Search benchmarking is especially important in the case of website search, as here the competition is certainly going to be Google. Trying to implement a search application that is "better" than Google is a waste of time unless you are prepared to invest the $10 billion that Google currently spends annually on research and development. In many organizations, such as universities, the website is a core information resource but the queries that might be posted from academic and research staff are likely to be very different to those from prospective students.

### Search Logs

Search logs are an invaluable source of user requirements. Their role in managing the performance of search is discussed in Chapter 15 but they are equally important in the assessment of user requirements.

### Stories

Stories about search successes and failures can be very powerful in supporting a business case but not in defining the functionality of the search application. Extrapolating from even a number of stories some specific features that are required is not sensible.

### User Feedback

All search applications should encourage users to provide feedback on their search experience, be it good or bad. A simple form on the search home page that gives users an opportunity to write a brief comment is all that is needed. The form should automatically capture the query terms. Asking users to fill out a detailed questionnaire never works. Calling them personally to discuss the search outcomes always pays dividends.

## Writing the User Requirements Report

Almost certainly what will emerge from this work is a classic 80/20 set of requirements—good agreement on the core requirements and quite a number of outliers. It is important to make sure that the reasons for these outlier requirements are fully understood. It is essential that the draft user requirements report is circulated widely, and certainly to anyone who was involved in any way with the user research. It may not be until these employees read the report that it becomes evident that one particular group feels they did not present their case clearly enough. Other readers, seeing the results, may be able to contribute additional insights, and perhaps a story that can be used for emphasis.

During the course of 2015, I worked on a major intranet and search assessment project with Sam Marshall (Clearbox Consulting). Sam introduced me to the use of

Trello as a way of gradually working toward a definitive set of user requirements. It is certainly an approach I can recommend.

All this research and analysis takes time. The overall schedule might go as follows:

*Month 1*
: Plan out the user research project and brief all those who will be involved, and stakeholders, as appropriate, about the objectives and scope of the research.

*Months 2 and 3*
: Allow two months as a minimum for the user research. Setting up meetings with individual teams can often be a critical step in the timing as these may only happen on a monthly basis.

*Month 4*
: Summarize the outcomes and check any anomalies before preparing the draft requirements report.

*Month 5*
: Allow several weeks for a review by participants before concluding the user requirements work and writing the final report.

This suggests that work on the user requirements research probably needs to start six months before the process of writing the requirements for a new search application or for an enhancement to the current search application. This may seem quite an extended period of time, but this is an application that could make a significant difference to the performance of everyone in the organization and the performance of the organization itself.

## Summary

Your employees will search in many different ways. There could be one small user group to whom a search engine with a particular feature could have a significant impact on operational performance. The user experience with a search engine starts at the point that the user realizes that he needs to find a piece of information and ends with the successful use of that piece of information to make a good decision. The range of use cases will mean that a range of different techniques are going to have to be employed, with consequences for the research schedule and for the resources needed. As much as possible, use techniques that can be used to measure the success of the implementation. Above all, remember the adage that if it can't be measured, then it can't be managed.

## Further Reading

Steve Portigal, *Interviewing Users* (Brooklyn, NY: Rosenfeld Media, 2013).

Steve Mulder and Ziv Yaar, *The User Is Always Right* (Berkeley, CA: New Riders, 2007).

HR-Survey, "What Is an Employee Climate Survey?"

In addition to these resources, the entire Rosenfeld Media catalog is concerned with user experience and user requirements research.

## CHAPTER 11

# Searching for People and Expertise

One of the most important uses of a search application is to be able to find people working for the organization or to identify people with specific skills and expertise. The importance of employees being able to find other employees by name or by expertise is of crucial importance in taking full advantage of the investment that the organization has made in its workforce over a period of many years. Much is made of the benefits of collaborative working, but this style of working has to start with creating the best possible team.

The US IT company Autodesk created a visualization of how the organizational structure of the company changed between May 2007 and June 2011. In each of those 1,498 days, the entire hierarchy of the company was constructed as a tree with each employee represented by a circle, and a line connecting each employee with his or her manager. Larger circles represented managers with more employees working under them. The tree was then laid out using a force-directed layout algorithm.

From day to day, there are three types of changes that are possible:

- Employees join the company
- Employees leave the company
- Employees change managers

The point that this visualization makes is that networks are constantly forming and (no matter how hard people try) breaking as an organization develops. Relying on personal contacts is not good enough to be sure that the best available people are working together on a project or task.

In the quest for "relevance," most of the limited effort available to optimize a search experience is directed at document search, and little attention is paid to people search. It is important to appreciate that a document search may be carried out in order to find people with specific expertise. In such a situation, the most relevant document may not be the one with the most useful information but the document that identifies one or more people to turn to for assistance.

## Name Search

Name search is very easy for users to evaluate. All they have to do is search for someone they know. From the moment they find that the search application does not find this person, they are unlikely to trust the search engine again. Almost certainly they will also use their own name as a search term and then be either very surprised or very concerned about the amount of information they find!

Even if the nominal business language of an organization is English, the issues of language quickly appear in name searching, as many employees will have family or given names that reflect the culture and heritage of their family and not the language that they may speak or their office location. Indeed, even working out which is the family name and which is the given name can be very difficult. In Chinese, as in many other Eastern languages, the family name precedes the given name, but it could be that the person concerned has inverted the structure for use in a Western culture.

Another common challenge is when a name requires the use of an extended alphabet. The Swedish given name *Åsa* is not the same as *Asa*, and in a Swedish alphabetical list, comes after Z. However, it may well have been transcribed differently, especially if the HR database cannot cope with extended character sets.

HR database issues often lie at the heart of an effective people search. The HR database provides a legal record of employees, and will use a legal name as would be set out in a passport. It will usually have a range of fixed fields and rarely will it contain any information about the role of the employee or her experience. For certain it will contain information that is subject to data privacy legislation, and many HR managers will be concerned that this information does not leak out, enabling the HR database to be searchable. In many organizations, there may be contractors working on short- or long-term contracts and employees who are working part-time, perhaps on a job share basis. These people may not be held in the master HR database. It can also be useful to maintain a record of past employees. A document written by one of these employees might be found in a search, and being able to track back who the employee worked for and which department he was in might help to track down his replacement or someone who was familiar with his work.

To gain an appreciation of the complexity of searching for a name, it is advisable to read the range of briefing papers from Basis Technology. The text of the section that follows is based on a Basis Technology white paper, “The Name Matching You Need – A Comparison of Name Matching Technologies,” published on February 29, 2012.

Typographical errors

A slip of the finger at the keyboard causes transposition on of characters, missed characters or other similar errors. (e.g., “Htomas” or “Elizbeth”)

Phonetic spelling variations

Some names simply sound alike, but are spelled differently (e.g., “Christian” and “Kristian” and “Anna” and “Ana”)

Neglecting to confirm spelling produces errors. (e.g., “Cairns” vs. “Kearns” vs. “Kerns”; or “Smith” vs. “Smyth”)

Transliteration spelling differences

Multiple transliteration standards or “approximate” transliterations from a non-Latin script to English lead to multiple spelling variations. In the case of Arabic to English, Arabic has many consonant sounds which might be written with the same English letter, or Arabic vowels may be expressed more than one way in English, giving rise tomany spelling variations. (e.g., “Abdul Rasheed” vs. “Abd-al-Rasheed” vs. “Abd Ar-Rashid”)

Initials

Sometimes all name components are spelled out, other times initials are used. (e.g., “Mary A. Hall” vs. ” Mary Alice Hall” vs. “M.A. Hall”)

Nicknames

In some cultures, nicknames are numerous and may be often used in place of a person’s formal name (e.g., “Elizabeth”, “Beth”, “Liz”, and “Lisbeth”)

Re-ordered name components

The order of family name and given name may appear swapped due to database format or ignorance of cultural naming convention. (e.g., “JohnHenry” vs. “Henry, John”; or “Tanaka Kentaro” vs. “Kentaro Tanaka”)

Missing name components

Sometimes a middle name or patronymic (personal name derived from ancestor’s name—e.g., Olafsson = “son of Olaf”) may be absent. (e.g., “Abdullah Al-Ashqar” vs. “Abdullah Bin Hassan Al-Ashqar”; or “Philip Charles Carr” vs. “Philip Carr”

Missing spaces

Some names are commonly written with spaces in different places, both in common English names (e.g., “Mary Ellen”, “Maryellen”, and “Mary-Ellen”) and those less common in English (e.g., “Zhang Jing Quan” and “Zhang Jingquan”).

Names in different languages

> Names from languages using different writing systems can be notoriously difficult to match against English representations of the names. Here is just one name spelled in English, Russian, simplified Chinese, and traditional Chinese, respectively:
>
> “Mao Zedong”, “Мао Цзэдун”, “毛泽东”, or “毛澤東”).
>
> © Basis Technology 2012

Chinese, Japanese, and Korean names present very substantial challenges. For example, Korean names have a single-syllable given name and a two-syllable family name. Western names are uniformly spaced between given name, middle name, and surname. By comparison, the three syllables of a Korean name can be written as all attached or spaced. Inconsistencies in separating the two syllables of the given name then leads to difficulties in Anglicized name identification.

A further complication arises from the fact that in the case of Korea (and the situation is similar in China), there are comparatively few family names. In the United Kingdom there are thousands of family names, but there are only 286 Korean family names listed in the 2010 South Korean census.

A recent challenge in language management has been the arrival of Arabizi, which is also referred to as Arabish or Araby. This is a form of Arabic used for text messaging, blogs, and microblogs, and for communicating in Arabic when only a QWERTY keyboard is available on a smartphone or tablet.

Wikipedia provides good coverage on issues around naming conventions, including entries in the following languages:

Arabic

Chinese

German

Japanese

Portuguese

Spanish

## Name Matching Technologies

Four types of methods are most frequently used to score name similarity, with Basis Technologies recommending a hybrid approach, as no single method will solve all the potential problems.

> Common key
>
> These methods, such as Soundex, reduce names to a key or code based on their English pronunciation, such that similar sounding names share the same key. Common key methods are fast and produce high recall (finds most of the correct answers) but have

generally low precision (i.e., contain many false hits). Precision is yet lower when matching non-Latin script names, which first must be transliterated to Latin characters to use this method.

List-method

This method attempts to list all possible spelling variations of each name component and then uses the name variation lists to look for matches against the target name. The result can be slow performance if very large lists must be searched. Furthermore, this method will not match name variations not appearing in its lists.

Edit distance

This approach looks at edit distance, that is, how many character changes it takes to get from one name to another. For example, “Catherine” and “Katherine” have an edit distance of 1, as the “C” is substituted for “K.” Edit distance methods work for Latin-to-Latin name comparisons, but precision suffers as each edit is weighted similarly, so a replacement of“c” for “k” is considered equal to a replacement of “z” for “t.”

Statistical similarity

A statistical approach trains a model to recognize what two “similar names” look like so that the model can take two names and assign a probability that the two names match or not. This method produces high precision results, but may be slower than the common key method.

## Job Titles, Roles, and Responsibilities

A requirement that is often overlooked is how to provide a way for users to find out who is responsible for a specific task. The person responsible for quality management may not have the title of Quality Management. That person may be the Director of Product Engineering or the General Manager–Bristol. This requirement is important in organizations with a number of country offices, especially in those cases where the office is quite small and many different areas are being covered by the same person.

In the course of writing this chapter, I found myself working for a law firm that provides a very comprehensive directory of lawyers, often with biographic information and case experience running to several thousand words. Yet the only information on an equal number of business support staff was a name, a title, and contact information. In effect, half the employees did not exist in terms of the skills and experience they brought to the firm.

The solution to this requirement depends on the quality of information in the HR database, and this may vary by country. It may be useful to work through the roles that are more likely to be cross-country or cross-subsidiary and focus in on providing job titles, roles, and responsibilities for a core group of staff.

## One Box or Two?

There is more discussion about whether there should be a people search box as well as a search query box than probably any other user interface topic. If there is no standalone people search application, then having two search boxes makes no sense. It does make sense to consider presenting the directory-type information at the top of the search results page, even if there are other results that mention the person by name.

If there is a standalone people search application, then a second search box may be appropriate, but it needs to come with a help or scope note so that users know the limitations of the application such as only presenting the HR-approved name. If the people directory does not have basic stemming and wildcard capabilities (e.g., to cope with *Ana* and *Anna*) then there would have to be a question about whether searching the directory had any real value. To be able to search for *Ana* and *Anna* with *A?na* assumes that the user knows that *Ana* is a Hebrew name that is widely used in Spain and Portugal.

The best way to come to the correct decision is to assume at the outset that there will only be one search box and then try to make a user-centric case for providing a second box for people search.

## Evaluating People Search

It is important to pay attention to the use being made of people search options. If the people search is being run across an SQL-type database from HR, it may be quite difficult to track down failed searches. If it is the main search application that is being used, then a list of the searches carried out on staff names, including failed searches, should be a component of the monthly metrics program.

It may also be an interesting exercise to ask a group of employees to search for themselves (including their expertise) and see if they are satisfied with the result set.

## Expertise Search

A common justification for a search application is that it will enable employees to find colleagues with the expertise they require to solve a problem. This is a laudable objective, but in reality it is very difficult to achieve. A good place to start in explaining why this should be the case is to consider some of the principles of rendering knowledge set out in 2008 by Dave Snowden, one of the leading knowledge management practitioners. Four of the seven principles are set out here with a synopsis of Dave Snowden's commentary.

***Knowledge can only be volunteered. It cannot be conscripted.***

You can't require people to share their knowledge, because you can never measure if they have.

***We only know what we need to know when we need to know it.***

Human knowledge is deeply contextual and requires stimulus for recall.

***The way we know things is not the way we report we know things.***

When people are asked to describe how they made a decision after the fact, they will tend to provide a more structured process-oriented approach that does not match reality.

***We can always know more than we can say and we will always say more than we can write down*.**

This is probably the most important. The process of taking things from our heads to our mouths (speaking it) or to our hands (writing it down) involves loss of content and context.

The implications of these four principles is that no matter how much effort is put into persuading employees to write down what they know and what their skills are, it will only be a very partial and probably biased commentary. Searching through these profiles is not a complete solution to identifying expertise and knowledge.

Then comes the challenge of keeping these profiles current. If there is no overall policy on profile management, supported by managers, then there is no incentive for people to spend time on this process.

Another problem with expertise search is that there is a danger that the same people come to the top of the expertise search and then are interrupted on a far too frequent basis by people who may not be seeking the knowledge of the expert but hoping that they can tell them where a report or other document can be found.

Searching any knowledge repository is only searching the tacit knowledge of the organization (i.e., the knowledge that can be written down). Of course, the minute it is written down, it has an uncertain value lifetime. It could be years or could be a few seconds. Another approach is to search through as many repositories as possible and collate together a knowledge profile based on the following:

- Participation in projects, teams, and groups
- Comments in blogs, microblogs, and discussion forums
- External papers and reports (e.g., papers in peer-reviewed journals and patents)
- Internal papers and reports
- Being cited in internal papers and reports (i.e., not just as the author)
- Internal and external presentations
- Structured CVs

- Membership of professional and trade associations

This approach has been commercialized by Sinequa. The challenge here lies in ranking the value of this heterogeneous collection of content. One of the downsides of this approach is that it can be biased against new employees who will not have written many reports or given presentations and who may not wish to reveal the details of projects they worked on for a former employer.

In summary, using search to locate expertise is not going to produce a definitive list of "people who know." It may provide a starting list, but the moment a user finds a reference to someone whose skills they have little confidence in, the level of trust in the application will drop to zero and will not bounce back.

Trying to develop expertise search without a knowledge management strategy and a knowledge manager is not to be recommended. Using search as a tool to reduce or even avoid investment in good knowledge management merely adds to the level of operational risk that the organization is working under.

There is an interesting paper from Ido Guy and his colleagues at IBM that describes a large-scale study of the use of the IBM employee directory. The use made of expertise search was so low that it was included in the 6% of "Other" requirements.

## Summary

The ability to find out about people in an organization is a very important element of an enterprise search strategy and requires an appropriate level of user research and testing. Because of the spelling complexity of names, which even in a small office could reflect many different nationalities and spelling conventions, matching a name to a specific query is not a trivial operation.

Using search as a means of locating expertise and knowledge is even more challenging, as it is impossible for anyone to write down all that they know about a subject, and difficult to infer from searching through reports and other documents they may have written. To have any chance of success information, search and knowledge management strategies must be aligned.

## Futher Reading

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si, "Expertise Retrieval," *Foundations and Trends in Information Retrieval* 6:2–3 (2012): 127–256.

Omer Barkol, Ruth Bergman, Kas Kasravi, Shahar Golan, and Marie Risov, "Enterprise Collective: Connecting People via Content," HP Laboratories Technical Report HPL-2012-102R1.

Basis Technology, “The Name Matching You Need: A Comparison of Name Matching Technologies,” 2012.

Ido Guy, Sigalit Ur, Inbal Ronen, Sara Weber, and Tolga Oral, “Best Faces Forward: A Large-Scale Study of People Search in the Enterprise,” *CHI ’12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, May 5–10, 2012.

Olli-Pekka Kauppila, Risto Rajala, and Annukka Jyrämä, “Knowledge Sharing Through Virtual Teams Across Borders and Boundaries,” *Management Learning*, 42:4 (2011): 395–418.

Inbal Tadeski, Omer Barkol, and Ruth Bergam, “A Study on Subject Matter Expertise,” HP Laboratories Technical Report HPL-2012-239.

Kush R. Varshney, Vijil Chenthamarakshan, Scott W. Fancher, Jun Wang, Dongping Fang, and Aleksandra Mojsilovi, “Predicting Employee Expertise for Talent Management in the Enterprise,” *KDD ’14 Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, August 24–27, 2014.

# CHAPTER 12
# Search User Interface Design

The purpose of this chapter is to highlight just a few of the interface design issues that need to be taken into account. There are no screenshots because of the immense difficulty in gaining permission from organizations to publish what are almost always depressing examples of user interface design for search.

The user interface for most enterprise applications offers little opportunity for creative thinking. The displays are usually driven by the need to present items of process-related information to a relatively small number of users who will probably be using the application throughout their working day. IT departments therefore have little requirement for user interface design skills, especially as the website search design was probably taken out of their hands by marketing or communications.

The result is that most search user interfaces seem to be whatever came out of the box when the application was installed. A commercial vendor may have developed some initial screens, but there is no ongoing usability testing and development. In an open source search application situation, there is usually very considerable flexibility for the user interface, but it needs to be specified at the outset, and that requires a very good understanding of user requirements and the potential capabilities of the software. Several books have been written on user interface design and these are listed at the end of the chapter in "Further Reading" on page 166. *Designing the Search Experience*, by Tony Russell-Rose and Tyler Tate (Morgan Kaufmann), is the definitive book on the topic at present, because the authors blend research with practical experience.

## Supporting a Dialogue

The overall aim of a search results page is to support a dialogue between the user and the search application. It could be that the user strikes lucky. There is one result that absolutely meets all her requirements and with a click she is away. That may happen in a website search situation, but is much less common in an enterprise situation. As Gerry McGovern has pointed out for many years, enterprise authors have no incentive to be found, whereas the entire purpose of a website is to ensure that users can find the information they need. As a result, care needs to be taken that the apparent rules for website search design are not transferred to an enterprise application without due care and attention.

The reality is that there are no rules. Even the placement of the search box at the top righthand side of the page is now being challenged by the requirement to support responsive design for mobile application. For all other elements, it depends on so many different factors that rules are irrelevant. Just one of these factors is that the search application may not be able to deliver the search interface that users would like to use because at no point in the selection process were user interface requirements taken into account. This can happen just as much with an open source search application as a commercial application.

There has been some debate over the years about the value of search against the value of browse. Do users start with browsing and then search, or search and then browse an area of the site that they have identified? It does not matter. At some point in time, probably every day, users will search.

## Testing Out Dialogues

It is fairly easy to test out what makes a sensible dialogue between a user and the search engine, and can be accomplished by card sorting. This is a well-known and widely used technique to determine the structure of an intranet or website, but seems not to have been used to any visible extent in deciding on facets and filters. As with information architecture development, the work starts with understanding potential routes to the information required, working up from the user requirement what some of the frequent search queries are likely to be.

In the case of engineers in a major construction company, they will always be looking for past expertise they can use in developing well-argued and costed proposals for a new client. Some of the options that might be of value are:

*Date*

Because they may be aware that the company has only been working in this sector for the last three years

*Clients*

Because seeing a list of other projects by client may highlight work that they were unaware the company had undertaken

*Scale*

They may only want to look at projects in excess of $100 million because anything smaller will not have required the particular solution they are considering

*Project team leader*

Because this might be a way to initially tap into the expertise that the company has in this area

The card sorting may well reveal that offering location as a filter option is of no value because the company prides itself in being able to work in any geographic area in the world. Although there are a number of proprietary card-sorting applications developed for web and intranet architecture development, using the Trello application is also worth assessing.

## Usability Testing

There is no substitute for usability tests for search interfaces. All too often the overall intranet, for example, is subject to usability testing but little attention is paid to the specific act of search. Indeed web designers see users reaching for the search box as a criticism of the information architecture and not as a core element of the overall discovery process. There are two reasons for this. The first is that in the development of the personas and use cases, search is relegated to being a side show, with comments like "Needs sophisticated search" or "Wants an advanced search feature". Doesn't everyone?

The second reason is that there is a lack of understanding about the different use cases of search (see Chapter 2), which means that at the most basic level, no difference is made between a search where precision down to a specific report is required and a search where every possible relevant document needs to be listed out.

## Accessibility

Taking account of accessibility issues in the design of search results pages is a significant challenge. In "Accessibility in Google Search", Google notes that it has "taken some deliberate steps to further improve the accessibility and tools that are commonly used by people with disabilities such as blindness, visual impairment, color deficiency, deafness, hearing loss, and limited dexterity."

The missing disability is dyslexia, which is a condition in which there is a wide spectrum of challenges to the user, and in many cases they make no mention of their condition because they have been taught or have discovered ways of coping with understanding text. Search results displays often make life very difficult for people with dyslexia. Many with the condition are able to recognize words by their shape. The indications are that 10% of the UK population and perhaps 15% of the US population has some degree of dyslexia. This means that in an organization with 1,000 employees, 100 will need support.

In the case of search results pages, the following design considerations should be taken into account:

- Do not underline text
- Do not use capital letters
- Use high contrast color palettes, but not black on white
- Use a line space of 1.5 or 2
- Reduce the amount of information presented on a single page
- Keep line lengths short

## Search Box

A guideline verging on a standard is that the search box is at the top righthand side of every screen. I often feel that the size and position of the box is a reflection of the value that search is given by the design team. This is the start of the dialogue, and a well-positioned and framed search box will give some comfort to the users as they frame the query.

Certainly the search box needs to be large enough to cope with longer search queries than might be used for a website. This is especially the case in industries such as chemistry, engineering, and pharmaceuticals. A search for *risk assessment polytetrafluoroethylene* is not unreasonable, but it is important for the user to be able to see all 40 characters and spaces of the query in the box.

Placing the box at the righthand side may cause a problem if the intranet supports responsive design, as this may result in the search box disappearing on smartphone devices.

The decision on whether to provide two search boxes, one for topic search and one for people search, is considered in Chapter 11. The question of whether to provide an "advanced search" option also needs careful consideration. This is where search personas are so important. There could well be a requirement for a specific advanced search option for a particular user group. A good example would be lawyers where the term *matter* is used for a client engagement, and the parameters of a client matter are very clearly defined in the process of dealing with the matter. The same would be the case for a chemical substance search. There is a good example of this at *www.chemspider.com* where the search box is almost the entire width of the page and a substantial amount of assistance is provided in framing the search.

In many cases, the search box may be a search portal box, providing a single point of access either to a federated search application or to a number of individual applications. There is a trend toward providing users with a drop-down list of application options from the search box. This has no value at all. It forces a user to make a choice without knowing how rich a source of information a particular application will be. It also means that a user is not able to select (in a federated search application) more than one application. There is also the issue of whether drop-down lists are acceptable from an accessibility perspective.

In situations where the user is being offered a choice of applications, any a priori attempt at a user interface solution without undertaking usability tests of the options will almost certainly result in low user satisfaction. It certainly fails the Google test; if it was such a good idea, why doesn't Google offer this option rather than an initial box of options and then, if required, a web page screen of all the Google search applications, each with a brief description? Among the options worth considering is what might be termed an Advanced Search option where the user can select which applications to include on a web page, with links to more information about each application, as required.

## The First Response

The initial response from the search application must provide the following:

- Confirmation of the search term used
- The number of results found
- Options for the next steps in the process of search

These might seem basic, but they are often overlooked. First impressions are everything in search. As highlighted in Chapter 2, providing thousands of results is not a successful search unless there are very good ways in which the number can be reduced very quickly indeed. In my view, query suggestion as a way of reducing the number of results in the initial presentation may not be optimal. Although there has been some research into query suggestion on websites, there seems to have been none in an enterprise setting. I would suggest that it is more important to offer suggestions when the user has had an initial opportunity to look at the first one or two pages of results, as then he will have some appreciation of the scale of the result refinement problem and some additional query terms that he could use.

## Result Review Speed

Over the last few years, the pattern of use of Google has shifted from looking predominately at the upper-left quadrant of the screen to looking all the way down the lefthand side. This is a result not only of the various search tools that Google offers on this side of the screen, but also because users are realizing that even for Google, placing the most relevant results as the first two or three returned is getting increasingly difficult. Users are therefore looking at all the listed results, and taking just 8–9 seconds to do so.

This is a good metric for enterprise search, though difficult to measure. The point is that search results need to be formatted for speed of review. This means:

- Good quality titles
- Good result formatting, with no underlining and an effective use of typography and color contrast
- Removing irrelevant icons (e.g., PDF logos and images) from the results
- An indication of the date of the content

## Titles

Search results are invariably presented as a list of information sources (usually 10, though sometimes more), with a title and some related information that should be of assistance to the user in assessing the relevance, credibility, and value of a source of information. The title associated with the document is of great importance, as users carry out a first scan of the results to see if they have found at least a good starting place on their journey toward information paradise.

If the titles are unhelpful, then the review process is significantly slower, and there is a real danger that a highly relevant result is either missed totally or judged to be of little value because of poor title quality. This is why it is so important to set out clear guidelines on title construction. There is some scope for creative communications in this respect. Listing out a selection of really obscure or misleading titles in the search team blog is a good place to start.

## Result Metadata

Every element of the metadata presented in the search result has to have some value. A table needs to be constructed with each of the elements and a commentary not only about how an item of metadata would enhance the process of discovery but what the impact would be of not having the item. A good example might be file size, which on an enterprise network is not going to be a challenge. In principle, it might present challenges for mobile applications, but with the development of 4G networks and access to WiFi networks, even on board my local bus, limitations on access to larger file sizes is fast becoming irrelevant.

File format can be of some value. An HTML page will probably contain links that could lead to a further step along the discovery road. A PowerPoint document is likely to be used for a presentation, but many consulting companies also publish their reports in a PowerPoint format.

## Summarization

This is a good place to consider the complexities of summarization. In Chapter 4, the technologies of summarization were considered, and now they have to be put into practice. It is accepted good practice to highlight the query term, ideally in the context of enough text to give the user some indication of the relative importance of the document. As with so many aspects of search, consistency is important. Although I am not aware of any research on the subject, my sense is that for results lacking a text summary, users may well not even look at the title as a measure of relevance value, perhaps making the judgment that if there is no summary, the source document must be too poorly structured for the summarization process to work.

## Facets and Filters

The technical issues around facets and filters were considered in Chapter 4. In this chapter, the consideration is about how to present the filter and facet options. The two approaches are either along the top bar of the search, or down the lefthand side of the page. Using the righthand side may well mean that the filter and facet options are overlooked. If there are a small number of filters, perhaps no more than six, then presenting them as tabs on a horizontal bar at the top of the results page is certainly an option. In the web space, this approach is common with university search results pages where it is (in principle!) possible to create well-defined buckets for people, departments, research, and courses. The downside of this approach is that future expansion is not easy because the tabs then become quite narrow, and in general, looking at a horizontal list is not easy, especially where there is no logical (typically alphabetical) order to the tabs.

It is this question of order that is one of the most difficult to resolve definitively. Should the facet elements be presented in the order of the number of hit occurrences or by the list of facets in alphabetical order? In my view, the default should always be to list in alphabetical order (or chronological order) unless there is a business/user case for a different sequence.

The second issue to be considered is how many facet options to provide. My advice here is to start small and only extend the number of facets when the user research indicates that this would offer a significantly improved user experience. This does become a challenge when the nature of the business interests of the organization are so diverse that the facet options seem almost endless. As a rule of thumb, offering more than four facet options is likely to confuse rather than assist the users as they try to work out (often based on hit counts, which seem odd) what the next step in the discovery process should be.

## Best Bets

One justification I often hear for the introduction of Best Bets is that users cannot find the information they need on the intranet because the information architecture is fast becoming a total mess. Search, browse, and monitor are three different and related ways of finding information, and it is important that there is a balance between all three. Search will never offset poor IA.

A much-quoted defense of Best Bets is that users click on them. That is not surprising, as that was the idea of adding them to the results page. The more important question is what users do after clicking on a Best Bet. In an ideal world, they need not click on any other result in the session, but a check of the search logs will show that is not the case. In some circumstances, the Best Bet will provide useful information, but there will always be a concern that it is not the most current or comprehensive document, so users will always want to look at other results just to make sure.

In many organizations, the initial list of Best Bet query candidates is driven from the top 100 (or more) search queries. The item that has the highest number of clicks is then promoted to be a Best Bet. It is time for a true story. Some years ago, I was working on the intranet of an Anglo-French company that had extensive interests in China. A search for *china* resulted in a set of results where the sixth most relevant item was about the use of china crockery in Royal Air Maroc Business Class. A check on the search logs showed that this was one of the most clicked results on the intranet because everyone searching for *china* found it and it would seem then showed it to colleagues, who repeated the exercise. It is very dangerous to assume that a high level of clicks for a document is an indicator of the value of a document.

The search manager then has to find people who will know what the definitive document is to match the term (100% precision!) and then not only add the document to the top of the results page but continue to ensure that the Best Bet remains the Best Bet.

The first question to ask is whether the search logs can be believed. Often the terms are listed as single words (e.g., *bearing*) even though the query may have been *track runner bearings*. Assuming that the search logs are being used effectively, the second question is why the information is not more readily available on the intranet or similar application. If information on bearings is organized hierarchically, then users will be led through a route that will progressively refine their requirements, and when they reach a possible document, the landing page will provide context and related links. A Best Bet will not have this context, and rarely will it have a URL that enables a user to work back through related content.

The big challenges of Best Bets are deciding which search queries deserve a Best Bet, identifying the document and an owner, and then making sure it remains current. These all take time and effort of what is always an under-resourced search team. There are so many better things to do with the time that will be spent (and I would argue wasted) in managing Best Bets. Things get much more interesting in intranets with multiple language content. If a German engineer is searching for *Rollenlager* (roller bearing), should the German language "Best Bet" be presented or the "agreed" English language Best Bet? This is not a question I can answer, but here again, the time and effort that will be expended in discussions around the optimal solution will be considerable.

There could well be a query situation where there is a document that may be relevant to a search but not directly related to the query. An example could be in a chemical company where any search on a range of defined hazardous chemicals also brings up relevant health and safety information. A query for a particular country market might list as a Best Bet a document that highlights the company's policies on ethical business conduct. I call these "second order" Best Bets. They will still take time to develop and maintain, but at least they are not trying to turn the search application into a 100% precision performance application, and it could be that a single Best Bet (on hazardous chemicals) could match a wide range of queries.

The final decision has to be one that is based on solid user research, and there is often a lack of this in most organizations because of underinvestment in the search team. Without a good understanding of what, why, and how users query, any Best Bets initiative will be seen as a very visible sign that the organization fails to understand the value of search.

Despite the widespread use of Best Bets, there is very little published research on whether they work, and the only reasonably detailed commentary on the topic that I have been able to locate comes from Lee Romero.

## SYOS and FYOR

This section is my attempt to add some more acronyms to the technology of search. Search your own site (SYOS) is discussed in Chapter 15, which covers search evaluation. All search managers should start their day carrying out a number of searches on topics agreed with the search team, and asking if the outcome would be in line with a user's reasonable expectations. This is a very ad hoc piece of research, but has the benefit of reinforcing the raison d'être of search to deliver relevant information to users with the minimum of effort.

Find your own report (FYOR) is along similar lines. Somewhere in the system will be a report that you have written. Try finding it without using the words in the title, as other users may not know what the title is. Even looking for the same document on a regular (say monthly) basis can be illuminating, as its usual position on a list of search results may suddenly be promoted or demoted. Neither is a poor outcome so long as there is a logical basis for understanding the change in rank order.

## Summary

Enterprise search user interfaces can easily become overcomplicated with options driven by the functionality provided by the technology. In developing user interfaces, every single element and its positioning need to be considered both individually and collectively, with each element having to justify its inclusion on the basis of enhancing the quality of the dialogue between the user and the application. Because of the importance of this dialogue, there is no substitute for a carefully constructed program of usability tests prior to the release of a new user interface, or a small change to a current interface. Using search user interfaces that have been adopted by other organizations should always be the subject of a discussion with the organization about why the interface was developed and what the user reaction has been.

## Further Reading

Marti A. Hearst, *Search User Interfaces* (Cambridge, UK: Cambridge University Press, 2009).

Robert Hof, "How Do You Google? New Eye Tracking Study Reveals Huge Changes," *Forbes*, March 3, 2015.

Peter Morville and Jefferey Callender, *Search Patterns* (Sebastopol, CA: O'Reilly, 2010).

Tony Russell-Rose and Tyler Tate, *Designing the Search Experience* (Burlington, MA: Morgan Kaufmann, 2012).

Daniel Tunkelang, *Faceted Search* (San Rafael, CA: Morgan and Claypool, 2010).

Max Wilson, *Search User Interface Design* (San Rafael, CA: Morgan and Claypool, 2012).

# CHAPTER 13
# Specification and Selection

Now that all the work has been done on identifying user requirements and reviewing the performance of the current search application, perhaps the time has arrived to specify, select, and install a new enterprise search application. In many respects, the processes for selecting an enterprise search application is just the same as for other enterprise applications, but there are some differences.

These include:

- The company will not have previously undertaken the procurement of an enterprise search application, so there is no prior experience to go on.
- Even if there are existing search applications, the level of knowledge inside the IT department about how they work and how to evaluate an enterprise application is likely to be low.
- There is probably no single business owner of search, and yet it is because of the poor performance of existing search applications that the company is now undertaking the selection of a new application.
- Most of the companies in the business are totally unknown to either the IT department or to procurement.
- It is not easy for a company to understand the differences between the applications offered by the vendors and by open source developers.
- If the procurement doesn't deliver significantly better search performance, the problem will be immediately apparent to most employees on the day it is launched.

This chapter provides advice on how to undertake the specification and selection of an enterprise search application. There are a lot of common elements between choosing a commercial application and an open source search application. However, there are some considerations that are unique to open source search because of the development options available. These are considered in Chapter 8.

## How Much Will It Cost?

This is a very good question, but search vendors are extremely reluctant to disclose any pricing information. The distinguished exception is dtSearch, which publishes a price list that is based on the number of seats/user IDs that will be purchased. The following are among the pricing approaches that vendors adopt:

- Server based, though often working out how many servers are needed is not easy to accomplish
- Document based, which is used by Google for its search appliances (however, it no longer publishes a price list or what definition of a document applies); the document definition can be a challenge in the case of Excel spreadsheets and anything that looks like a database
- Seat based, as in the case of dtSearch
- Subscription based, often used by open source search vendors for their management packages
- Module based, with again initial uncertainty about which modules might be required

To make things even more complicated, vendors will mix and match these options. In the case of Google, it is also important to take into account that the license is for a fixed period of years and for an upper limit of documents. All these variations and the unwillingness of vendors to disclose even representative pricing makes it very difficult indeed to develop a business case that is based on reliable and comparable license fee information.

Because most organizations have not purchased a commercial solution for some years, if at all, and have no understanding of the likely development costs for open source applications, there is no prior knowledge of the costs that will be involved for installation and implementation, nor for a total cost of ownership over a five-year period.

## The Project Team

There are four stages to the selection and implementation process. Implementation is the subject of Chapter 14, but this is a good place to take an overview of the stages:

- The project to write the specification for the enterprise search application
- The project to select the vendor and/or integrator/developer
- The project to install the application

- The project to implement and then develop the application

It is far too easy to see the enterprise search project as coming to a halt when the software is installed. One of the key messages of this book is that the hard work really starts once the software has been installed.

The reason for setting out the stages is that the project team needs to have the appropriate skills at each stage. There will need to be a substantial involvement of IT with the installation, but less so in writing the specification and undertaking the selection process. It may be that there are some gaps for which currently there are not people with the appropriate skills and experience, or perhaps the time to work on the project. These will need to be identified and filled ahead of the appropriate stage of the project. The gap that cannot exist is that of the search manager, the person who will ultimately have the responsibility for making sure that the search application meets the needs of the business and every individual employee who uses it. This position is so important that if a search manager cannot be identified and appointed, then it would be advisable to delay the start of even writing the specification until the position is filled.

At the outset, it is advisable to work out how the project team is going to transition into the search management team or be members of a Search Center of Excellence.

## RACI Responsibility Matrix

A very useful way of managing the responsibilities for a search project is to adopt what is often referred to as a RACI responsibility matrix:

*Responsible*

Project members who will be directly involved in the specification, selection, and implementation

*Accountable*

The person who has the overall responsibility for the successful delivery of the search application (this person holds the budget for the project and can make executive decisions on extending the project schedule, budget, or scope as required; in most organizations this will be a senior IT manager)

*Consulted*

Managers who have a direct interest in ensuring that the specification, selection, and implementation takes their business requirements into account

*Informed*

Managers who have an indirect interest and only need to be kept aware of the progress of the project

There are a number of variations on this matrix, but the objective is to manage all the many different interests in the outcome of the project with the least amount of effort.

## Specification Project Team

This team will prepare the business case and the functional specification. There needs to be good representation of the business units that have the most to gain from the new search application. In addition, there will ideally be an enterprise architecture specialist, the intranet manager (who will not only know about search but will be in a position to support the communications plan for the project), and someone with a good knowledge of the network architecture of the company. After adding in a project manager, the search manager, and the project sponsor, the result is a specification project team of perhaps 8–10 people.

## Selection Project Team

If the specification team have done its job well, the selection team can have more of an IT focus. Security management is a very important element of a successful implementation, and the project team needs to have expertise in identity management and/or systems security. At this stage, obviously the procurement department needs to be closely involved. Again, the search manager and project sponsor will be on the team along with the project manager.

## Installation Project Team

The membership of this team is discussed in more detail in Chapter 14, but is included here because this team needs to be set up from the outset of the project.

## The Global Dimension

The majority of enterprise search implementations will be across more than one country, and probably more than one language. Right from the outset, the implications of a transnational implementation need to be taken into account in setting up the project team that will take responsibility for preparing the specification, undertaking the selection, and then managing the implementation. It might be the case that the initial implementation is in the United States and then the application will be rolled out more widely across the company. Even if this wider implementation is not going to be carried out for a year or more, all the business and IT units involved need to have some degree of representation on the project team at one of the three levels just discussed.

## Risk Management

Whatever the project management approach being used, a risk management section is very important. Risks are commonly scored by the impact on the project and the probability of the risk event arising. In the case of implementing an enterprise search application, there will be no prior experience in the company, and using probabilities from other enterprise application implementations is not a sensible way to proceed.

The risks that tend to arise in search projects include the following:

- A change in the project sponsor
- An inability to manage secure access to confidential information
- Meeting architecture requirements to support the required server performance
- The loss of the search manager
- Lack of internal IT resources because other enterprise implementations have been given a higher priority
- Lack of continuity of representation from IT
- At the proof-of-concept stage, none of the vendors meet the core requirements
- Poor quality content and/or metadata means that there is little perceived improvement in search quality
- Inadequate time allowed not just for performance and user testing but to carry out redevelopment based on the outcomes of these activities
- Substantial business change, such as a merger or acquisition

## Project Schedule

Selecting and implementing an enterprise search application can take some time to bring to completion. The typical steps and the time that should be allocated to each one are listed in Table 13-1.

*Table 13-1. Typical steps*

| Step | Duration |
|---|---|
| Determine and document user requirements | |
| Develop a short list of vendors, developers, and implementation partners | |
| Visit other companies that have recently implemented an enterprise search application | 2 months |
| Prepare the request for proposal | 1 month |
| Circulate to the short list of vendors and allow time for the response | 1 month |
| Assess the proposals | 1 month |
| Set up the proof of concept | |
| Invite vendors to participate | 1 month |
| Undertake proof-of-concept evaluation and select a preferred vendor | 2 months |
| Negotiate a contract | 1 month |
| Prepare for installation | 1 month |
| Implementation and acceptance testing | 1 month |
| Initial rollout and assessment of user experience | 2 months |
| Total time | 13 months |

Adding in a request for information round will extend this project schedule by a month, and there could be additional time needed for the proposal preparation and evaluation stages if there are some regulatory requirements for public and government procurement schedules.

It is sometimes possible to complete one or more of these stages earlier than expected, but it is still wise to allot at least a year to complete the entire process, from the time the decision is taken to implement a new search application to the time that users begin to use the fully configured application. During this time, the business may have changed its objectives and/or there could have been changes to other enterprise applications, and it is advisable to have a major project review before the proposals are assessed to ensure that the specification is still valid.

Of course it will be different for your organization; it always is! You will look at this schedule and change the word “month” into “week” for many of the project stages. There are three possible outcomes:

- You will reduce the elapsed project time, but the search application will not deliver to requirements and expectations
- Costs and resources will be based on a shorter project time, but if the schedule cannot be met, then there will be cost over-runs and members of the project team may have to go off to other projects

- You will deliver excellent long-term search satisfaction to budget and to the shortened project time, and then be able to make a very good career as a search consultant

The choice is yours.

# Writing the Specification

The first decision to be made is whether to go straight to a request for proposal (RFP), also known as an invitation to tender (ITT), or to prepare a request for information (RFI) as a means of reducing the number of proposals received. The decision depends on the company's level of expertise, and also whether there is an intention to go down an open source route. In the case of open source applications, there are an increasing number of development companies, and a preliminary shortlisting can be valuable. The route taken will also depend on the procurement policy of the organization, especially in the European Union where there are some EU-wide public sector procurement rules.

Most companies will have a preferred way of writing a specification. In this section, the information that vendors will expect to see in the specification is set out. The order in which it appears in the specification is unimportant. With every one of these sections, it is important to present not only the current state of affairs but what the expectations are going to be over the three years of operation after implementation.

## The Story so Far

Some background on the decision to go out to tender for an enterprise search application should be included. Vendors appreciate a high level of honesty at this stage because it enables them to judge what their approach should be in presenting the benefits of their solution.

## Content Scope

This section should go into some detail about the volumes of content to be indexed, in terms of both raw storage and also an indication of the number of documents, the rate of addition of new content, and how quickly this new content needs to be able to be searched. Also in this section the main file formats need to be listed, and, of course, any languages that need to be indexed and/or supported with language-specific interfaces. Even if the initial implementation is going to be for text-based content, almost certainly the development of content analytics and search-based applications is going to create new opportunities and requirements in the not too-distant future.

## User Expectations

The word "expectations" is used rather than requirements because this section needs to cover the expectations of all the stakeholders. These expectations cover not just the search experience but the timescale for the implementation and the ability to customize the search application without any further support from the vendor. If these are not set out in the specification, they will come up at the selection meetings and can easily derail the entire process.

## Information Systems Architecture

Hardware requirements can be a significant cost element, especially storage and network bandwidth, so the current information systems architecture does need to be clearly set out. Again, future intentions need to be clearly signaled, especially intentions to move toward cloud-based applications.

## IT Partnerships

Many companies already have long-term contracts with systems integrators and have outsourced development to companies based in India and some other countries. For a vendor, some of these partnerships could be advantageous, and others may present challenges because they may themselves have agreements with other search vendors. Included in these partnerships should be information about enterprise contracts with major suppliers of software and services, notably IBM, Oracle, Microsoft, and HP, all of whom have significant interests in search technology.

## Internal Development and Support Resources

Internal development and support resources are especially important in the case of open source projects. Vendors want to get a sense of who they will be dealing with at the installation and implementation stages, and it is advisable to present the corporate expertise as short profiles of individual members of staff. Almost certainly there will be other enterprise development projects taking place at the same time as the search implementation, so there could be some potential issues about availability over the six- to nine-month period that it might take to fully deploy the search application.

## Security and Identity Management

The corporate approach to identity management and security management, both of access to the search system and also at a document level, should be set out in detail. It is especially important to assess the security implications of mobile enterprise search and of using hosted search applications for extranet and project management applications where users may not be employees of the organization. This is also the section to highlight issues around data privacy.

## Federated Search Requirements

The vision for enterprise search is that it will be able to search across multiple repositories and applications and present a ranked list of relevant information. This is certainly possible, but the costs and other implications are considerable. Although there may be some fine print in the functional requirements, the expectations for federated search should be highlighted in the initial section of the specification. Also in this section should be a list of all current search applications (e.g., SharePoint 2010) and what the current plans are for upgrading these applications.

## People Databases

One of the most valuable benefits of enterprise search is being able to find individuals by name and by experience and expertise. In this section, details of any HR databases should be given, together with the extent to which the company requires the enterprise search application to meet the requirements of national and EU-level data privacy legislation.

## Project Timetable

The stages of the overall project program should be clearly described, including what will be expected at the proof-of-concept stage, as this will be quite labor-intensive for the vendor.

## Functional Specification

The first piece of advice is not to write a detailed functional specification that ends up with perhaps 500 individual functional requirements. This was certainly the case in a specification produced by a large financial institution a few years ago. There are a number of reasons why preparing a very detailed functional specification is not going to improve the chance of finding the best fit. The functionality of even a medium-specification search application is so comprehensive that it will be able to meet, at a surface level, most of the requirements in the list.

Most search vendors do not have a large team of people waiting around to prepare responses to RFPs that drop into their email inboxes. They will look at the time it is going to take them to respond and they may well decide that filling out a response to 200 or more boxes is not a good return on their investment. Either they will not reply or they will just go through the motions and cut and paste content from the last such RFP they received.

Probably at least 80% of all requirements can be met, to some extent, by all search vendors. The important functionalities will then be lost in the noise of the common features. There is no point asking for a list of all the file formats that the vendor is able to support when what is needed is absolute confirmation that the file formats that are important to the company can be handled with certainty.

The time it is going to take the project team to review each proposal is going to be considerable. By the time the last of 10 proposals have been reviewed, the team has almost lost the will to live and is not looking carefully enough at the proposals.

The approach that should be taken is to focus in on areas where there are some differences of approach between vendors (and this includes open source suppliers) and which therefore enable the team to come up with a well-considered short list for more detailed review.

## Connectors and APIs

If the search application needs to be able to index content from all required repositories or provide federated searches across applications, then any concerns about the capability of the vendor to achieve this need to be identified at the outset. As with so many elements of search, it is not just whether the connectors and APIs are available, but the extent to which they have been deployed successfully in other clients.

## Index Freshness

The challenge for all search vendors is to be able to update the index with new content in a time that matches the requirement by the customer to be able to find recently indexed content.

## Filters and Facets

Most vendors now offer filters and facets to help users drill down into a set of results. It is important to check the extent to which these filters and facets can be modified by the search support team without the need to involve the support team from the vendor.

## Taxonomy and Metadata Management

Most companies will have some form of taxonomy, even if it is just a list of controlled terms or a list of approved abbreviations. Integrating these into the search application can be a challenge, and understanding the way in which this can be achieved is important to clarify right from the outset.

## Search and System Logs

If you cannot measure something, then you cannot manage it, and that is certainly the case with search logs and system logs. Many vendors will have some standard search logs, which are a good place to start, but creating just the views needed to manage your implementation could be a step too far and result in the need for the vendor to develop some customized reports.

## Entity Extraction

This topic has been covered in some detail in Chapter 6. Some vendors buy third-party products from companies such as Teragram and Basis Technologies, and others have developed their own entity extraction algorithms. As is so often the case with enterprise search, it is not what is supplied with the initial install that matters, so much as the ease with which changes can be made to the rules and algorithms on the basis of the experience gained following the initial implementation.

## Questions for the Vendors

So far we have covered the baseline information that vendors will value and some of the functional requirements that need to be clarified. In addition, there are some questions that need to be asked of vendors, the replies to which can be very valuable input into the evaluation process.

## Risk Assessment

Any enterprise search implementation involves risk. The organization itself has never implemented search on this scale before and so has little idea of the specific risks there may be with the implementation of the search application from a specific vendor. However, the vendor will have carried out a substantial number of implementations and should have a good idea of the risks and issues involved. It can be very illuminating to ask what the vendor sees as the main risks to a successful implementation based on the specification provided, and how the vendoe will work with the project team to ameliorate these risks. This can identify a potential lack of knowledge by the vendor in some aspect of the operations of the organization and also assumptions made by the organization about the time and resources the project will require.

## Project Schedule

It is not easy for the vendor to provide a definitive answer to the question about the duration of the implementation project, but the vendor should certainly know how long it took in the case of other clients. Ask the vendor to provide a case study of a similar implementation, setting out the timetable from the time of starting the proof of concept, and including information on the resources that both the vendor and the client contributed to the implementation.

## Project Management Methodology

The company may have its own approach to managing projects, often based on PRINCE2. A clear statement of the methodology used by the vendor will enable potential project management and communications issues to be identified at the earliest possible opportunity. Of particular importance is how red flag issues will be identified and dealt with. It is reasonable to ask the vendor to include some typical project management forms and procedures in the response to the proposal. Search implementation is quite complicated, as you will see from the next chapter, so the project management approach is a critical success factor.

## Upgrade Release Schedule

It is useful to gain some understanding of the application development roadmap of the search vendor. The search business is highly competitive, and there is an understandable reluctance to go too public with a product roadmap. Nevertheless, there should be some element of comfort in seeing what the product roadmap might be and so be able to assess the impact the future development possibilities.

## Supporting a Global Implementation

The challenge with enterprise search is providing installation and implementation across multiple countries. The level of representation of search vendors globally is very variable. US vendors may have a representative in Europe but their task is presales and some customer support. The technical teams are back at headquarters and that could be many time zones away. Global support issues need to be identified at an early opportunity. There is no point in acquiring a complex search application if the technical support cannot be effectively managed by the vendor, or the vendor help desk is based in the United States and there is only a small time slot open for EU-based search operations.

## User Groups

Perhaps surprisingly, there are not currently many vendor user groups. One of the understandable reasons for this is that the install base in any single country is too small to support a national user group. With the technology now available, there should at least be a regular virtual user group. The point here is to be certain about the quality of the dialogue between the vendor and the customer. You want to know about upgrades and bug fixes as soon as possible and also to feel that you have an influence on the way in which the search application is developed.

## Key Employee Strategy

Even in quite large search vendors there are some employees engaged on either development or on installation that have accumulated a significant amount of expertise. It is important to ask whether the search vendor has a key employee strategy in place, so that if one of these employees leaves the company, the vendor is not left exposed in either development or in implementation expertise. Of the two skill sets, implementation expertise is more important because it could directly affect the schedule and the quality of the implementation.

## License and Support Costs

In the initial proposal, it is unlikely that there will be more than an indication of the full cost of the implementation. Fixed-cost contracts are very rare, as there are so many unknowns from the viewpoint of the vendor at the stage of preparing the proposal. The only way to gain at least some sense of the final cost is to set out some scenarios for expansion routes and get at least some estimates from the vendors, but these will all be couched in very vague terms. It is not that they are being difficult, just cautious about committing themselves to a set of unknowns.

## Reference Sites

This is always a sensitive subject. The number of variables is such that it's impossible to compare the implementation and search satisfaction among different customers, as each will have their own unique set of circumstances. What is worth exploring is the way in which the vendor and the reference customer worked together. Were there good channels of communication or were promises made and not kept?

## Training

Some training can be delivered on the vendor premises on a test server, but there will be a need for hands-on training during the installation and implementation stages. There should be a clear statement of how this is going to be carried out, and the prior experience that is expected of the staff being trained. Training employees to be trainers themselves seems to be a smart idea, but it is very important to make sure that the staff have the skills, time, and incentives to be trainers of others.

# Building the Vendor Short List

As the list of vendors in Appendix E of this book illustrates, there are around 60 search software vendors, but in reality the list of potential suppliers is a lot smaller. This is because many of the vendors are not in a position to support a multicountry implementation, or even a large-scale implementation in a single country that is many miles and time zones away. Many of these companies will have partnerships with sales and possibly implementation companies in other countries, but often only limited business is conducted through these channels.

At the other end of the scale, IBM, Oracle, HP, and SAP have global sales, implementation, and support networks, so in theory they should be in an ideal position to provide a multinational offering. The reality can be different, as expertise in the search applications may be biased to major regional markets, in particular the United States. If your company already has enterprise agreements with these companies, then it would be foolish not to consider what they have to offer.

The search industry is also covered by a number of other consulting companies, notably Gartner Group and International Data Corporation. Many consulting companies try to summarize the leaders and laggards in a graphical format, but using these as the basis for a short list is not a good idea because search applications are far too complex to reduce to a two-dimensional diagram.

Another useful approach is to post a request for advice to the LinkedIn Enterprise Search Professionals Group. There are currently over 7,000 members. Many of the replies may come via the private response route!

It has to be said that the search vendor industry is exceptionally good at hiding behind technology jargon. Here are a few examples:

- "The framework and the award-winning technologies provide the ability to transparently identify and tag content with semantic metadata and then classify it to organizational taxonomies aligned to business goals. The use of compound term processing, still unique in the industry, enables organizations to deploy intelligent metadata-enabled solutions that are being used to improve a multitude of enterprise as well as business process challenges."
- "Search is the ultimate killer app"
- "Build high value applications that provide immediate friction-less access to all information sources"
- "Friction-less access"

Many companies score the proposals they receive to help develop a short list to move onto the next stage. In the case of enterprise search, the complexity of the functionality means that the differences between vendors in terms of delivering functionality are going to be small. You may end up with scores of 235, 246, 297, and 299. Dropping off the two low scores does not make sense, as each member of the team will be scoring the proposals with little previous experience of enterprise search.

This is why a multistage approach is the best option, with an initial request for information to come up with a list of perhaps six vendors to whom the request for proposal is then sent out. The aim should be to end up with no more than three vendors for the proof-of-concept stage.

## Using a Consultant

There are consultants who maintain a strictly vendor-neutral approach to vendor selection projects, offering services from the user requirements work right up to supporting the project team in advising on the selection of a search application. These consultants also provide ongoing support post-implementation but very rarely get involved in the implementation work.

## Using an Implementation Partner

As described in Chapter 7, many of the larger systems integration companies offer search implementation services. In addition, there is an increasing number of specialized search implementation companies, some of whom will provide development services for open source search applications. Most of these companies will be in a position to do everything from the user research through to implementation. It may seem a very convenient way to manage a search project, but there is a substantial risk of ending up with an installed application and no knowledge of how it works and how it should be managed. Using an implementation partner should indeed be a partnership, and organizations should be well aware of the benefits and risks of any IT implementation partnership.

If using a partner is attractive in terms of speed of implementation and overcoming a lack of internal expertise, you will first need to determine how important speed of implementation really is and whether the lack of internal expertise will have a negative impact on the long-term management of the search application. It can be instructive to review other implementation partnerships that the organization has established and take the lessons learned into selecting and working with a search implementation partner. If there is no prior experience, then the project risk increases substantially.

The big decision is whether the search application is chosen first and then an implementation partner is appointed, or the partner is asked to advise on the selection of the search application. Most integration companies work with a small number of search applications with which they have partnership contracts and good access to expertise within the search vendor. This is especially the case with the larger general systems integration companies where search implementation has not been a major business for them in the past. One result of this is that they may not have much experience with the current version of the search software.

Many organizations have an incumbent systems integration partnership, perhaps as a result of outsourcing some elements of IT service provision. Care needs to be taken that the partner concerned has an appropriate level of knowledge of search technology and implementation. In addition, if a second partner is appointed just to support the search implementation, there is then a triangle of relationships between the organization, the incumbent systems integrator, and the search integrator that could be a challenging test of the skills of the project manager. This situation is especially likely to arise when the search vendor does not provide implementation services but works through a local partner.

There is no “best solution” to partner selection. The decision will have both benefits and risks that are dependent on the organizational context, and these need to be worked through in detail before any decision is made.

## Open Source Software Procurement

So far this chapter has been concerned with commercial software applications. Chapter 8 looks at open source search procurement in some detail, but to ensure that this chapter provides an overall assessment of selection issues, some of the specific elements of open source search selection are set out in this section.

The relationship with a commercial search vendor is very much about buying a software product with some consultancy services provided to assist with customization and support. The chances that there will ever be a meeting with the team that has written the application are very small indeed. With open source software, the business model is all about buying consulting services, and it is very likely that you will be meeting developers who will go back to the office after the meeting and start writing code. It is possible to carry out the entire development operation in house, especially if the organization has a strong Java development team, but the missing element will almost certainly be enough understanding about how search works to build an application that meets not only current requirements but also future requirements. In-house development can also be hindered by the IT department leaving out the user requirements and statement of requirements steps, and just asking the development team to get on with it. Which they will until a “more important” project comes along and the priorities of the development team are changed overnight.

Finding potential developers is not difficult. There is a list on the Apache Software Foundation site of people who have contributed code to Lucene and Solr, but many of these may work for large IT companies and will not be available for commercial development work. Using Google and Bing will also result in a list of potential developers, but there are probably fewer developers around than might be expected given the high visibility of Lucene, Solr, and Elasticsearch.

Before beginning to approach potential developers, it is important that all relevant managers have signed off on the use of open source search software. Probably the only other example of open source software in the organization is going to be a CMS, and these are far less complex than open source search. With a few exceptions, open source search developers either work for small companies or as members of a virtual team. The statutory accounts of these companies may cause some problems for procurement departments more accustomed to working with large multinational IT companies.

Another aspect of working with small companies and virtual teams is that they may not be able to immediately start work on a project. Indeed, if they can, it is worth finding out why, as competent open source search developers are in short supply.

There is no point in sending off a highly detailed statement of requirements at this early stage. The engagement must be about both sides building a confidence in each other, and the road to defining the requirements is much more of a collaborative process than might be the case with a commercial vendor. The initial discussions should focus on understanding who the members of the development team would be and what their role would be in the project. The development team will certainly be focusing on what the milestones would be for the project, as it will certainly not be a turnkey development approach with the team going off for a few months and returning with the finished software. The milestones are needed to keep the project on track and also to define payment points. Some developers may work on a fixed-fee basis for a small project, but for anything approaching an enterprise development, the contract will be on a time and expenses basis. At the beginning of the engagement, it might be quite difficult for the development team to give more than a broad estimate of the total cost of development.

The development team will largely work off-site and will need good access to the content that needs to be indexed. This can be a procedural challenge for an organization worried about the leakage of confidential data. It must be recognized that any transgression on the part of the development team would be immensely painful to the company, to them personally, and indeed to the open source community. The fewer nondisclosure agreements and complex firewall protocols, the better. Open source development works best when all concerned see it as a win-win partnership. This win-win extends to the developers being able to share innovative code with the community.

As well as small, independent development teams, there are many larger companies, notably LucidWorks, that also offer open source software development services. The business model may be different, but the fundamental elements of a shared commitment to development success remain.

# The Best of Both Worlds?

An organization may feel that it wants to hedge its bets and go out to tender to both commercial and open source solutions. The problem with this approach is that at the evaluation stage, the choice will be between apples and oranges. Using a productized open source solution will make the process a little easier, but at all costs resist the temptation to choose between a commercial solution and a bespoke open source development.

## Proof of Concept

This is sometimes referred to as a “bake-off” and is a very important part of the overall selection process. The objective is to give the potential vendors the opportunity to demonstrate how well their technology works on real corporate information repositories and applications. Preparing for proof-of-concept tests is quite time consuming, and at the stage of writing the proposal, the objectives of the tests need to be set out very clearly. This is as much about gaining a consensus within the organization about what constitutes a successful proof-of-concept as it is about being fair to all the bidders about what might be expected of them.

Two test collections should be developed. One of these should be a collection of perhaps 5,000 documents against which a number of representative use cases can be run. This collection will enable some key performance parameters to be verified, such as speed of indexing, speed of updating the index, server performance, and the default search user interface.

The second should be a collection of documents in every file format that has been identified in the content audit, the objective being to evaluate the performance of the document filters in indexing and in presenting documents in these formats.

A bigger challenge is to set up a federated search proof of concept, because this will require the vendor to have the appropriate connectors available. It is probably not worth the effort, and a more pragmatic approach would be to assess federated search capabilities at some reference sites.

Carrying out proof-of-concept tests is probably the most challenging element of the selection process. A balance needs to be set between a reasonable level of investment on the part of the vendor on the tests and what the reasonable expectations of the company are for the outcomes of the tests. Typically, a proof of concept may take a week to set up, test, run, and evaluate. The conditions for all the vendors need to be the same, and the project team from the company needs to be consistent. Servers need to be provisioned and a set of ACLs developed to assess security handling. The vendor team needs somewhere to work that can be kept secure; a desk in an open-desk area is not suitable. The tests may require the participation of IT staff and users in other countries, and their availability has to be factored in to the schedule. This is why in Table 13-1, the duration of the proof-of-concept tests is shown as two months.

## Contract Negotiation

It is not uncommon for the contract negotiations to take some time to conclude. The full cost of the project will not become apparent until the contract documents arrive. The vendor will have learned a lot during the proof-of-concept tests and will have factored in the impact of discoveries made during the process, especially about the skills of the team that will be responsible for supporting the implementation. It is advisable not to focus solely on the cost of the initial installation and implementation. Enterprise search applications are scalable, but the costs of scalability can be quite substantial—for example, in terms of the costs of developing connectors for specific applications and repositories.

A key factor in the cost structure will be the extent to which the vendor regards the prospective customer as a reference site. If the customer is the first in the sector to invest in a search application, then the potential business that could accrue from being able to make a lot of publicity from the win is worth quite a reasonable discount.

An element of the license cost that often only becomes visible at the contract stage is the number of servers that are required to establish test, development, and production environments, and to be able to scale as more content and applications are indexed. Another factor that is often overlooked is the costs involved through an acquisition or through a divestment. Although it is not possible to foretell the future, if the organization has growth through acquisition or divestment, some scenarios should be discussed at the contract stage that take examples of both into account. This is not just the case of arriving at a "cost per user" number but about understanding changes in support contracts. Search vendors will always be interested in increasing income, but far less so if the acquisition or divestment means a reduction in support income.

## Summary

Although organizations will usually have previous experience of selecting and purchasing enterprise-level solutions, little of this experience will prepare them for an enterprise search project. The project could easily extend over a period of more than a year from the time of the initial decision to upgrade search capabilities to the day of launch. Bringing in specialist expertise from consultants and from systems integrators will reduce many of the risks, but probably not the overall timetable. It is very important to specify what the search application needs to achieve in terms of business impact and not to provide vendors with a long list of features derived from a cut and paste of product documents downloaded from vendor websites. Undertaking a proof of concept is essential. Working out the implications of the lessons learned from the proof of concept and the complexities of negotiating a contract can take a substantial amount of time and effort.

## Further Reading

Johanna Rothman, *Manage It!: Your Guide to Modern, Pragmatic Project Management* (Frisco, TX: The Pragmatic Bookshelf, 2007).

CHAPTER 14

# Installation and Implementation

The way in which the installation and implementation of the search application is conducted will be very specific to a particular company. The milestones for an open source project will also be somewhat different to those for a commercial application. In this chapter, a distinction is made between installation and implementation. Installation covers the provisioning and testing of servers and networks, loading all the modules of the search application, checking that user authentication is being managed correctly, and undertaking user acceptance tests (UAT) that confirm that the base performance criteria are being achieved on a test collection.

Implementation is the process of extending the application to work on live servers and content, and moving the acceptance testing to the search support team and a small group of testers. Overall this could take at least a month to achieve and may require additional time for more complex federated search implementations.

The basic principles of search implementation are the same for a commercial or an open source search application, though the approach to project management may be a little different. The term *vendor* in this chapter applies to any external supplier of search software and development services. Some specific issues around the development of the open source search application itself are covered in Chapter 8.

## Project Management

Installing and implementing an enterprise search application is a complex project with perhaps 40 or 50 individual work packages. There will be little or no previous experience of installing enterprise search, so this project calls for the best project manager the organization employs or hires. The availability of this project manager will decide when the project can begin, and, of course, the project will need to begin perhaps one or two months before the software arrives as the vendor begins the task of fully understanding the content, information architecture, and security management environment.

Search implementation projects are characterized by a fairly large number of short-duration work packages that have a lot of interdependencies. Every dependency is a risk, and excellent risk management is essential if the project is going to deliver to schedule and objectives. There are many different approaches to project management, and a popular methodology is Prince2. However, organizations fail to recognize that a project can be fully compliant with the Prince2 methodology and still fail to deliver the expected outcomes. It is not just about writing a very detailed Project Initiation Document (PID), but having the experience to include all the relevant details. Otherwise, comparison with the PID will not flag up issues that need urgent attention.

Vendors will have their own approach to project management, and at the earliest opportunity there should be a discussion about how the vendor and customer project plans are going to be integrated into a common plan. A critical element of this plan should be a risk register. Risks are conventionally categorized as green (no action needed), orange (a potential risk is emerging but at present it is fully under control), and red (urgent action needs to be taken to manage the risk). There should be a clear escalation procedure for red flag situations, including who is responsible by both the customer and the vendor for devising and implementing a solution.

## Customer Responsibilities

At the commencement of the project, the vendor will identify information and support that they require from the customer. This could include a working area with secure storage, remote access to the customer network and the servers that will be used for the application, and content for indexing. The list will be quite long, and a failure on the part of the customer to deliver to agreed schedules could have some significant knock-on impacts on the project because of the dependencies between the work packages. Many vendors and system integrators are small companies and are working on multiple projects, so a delay of a week in providing a test sample of content could extend the project by several weeks.

## Implementation Schedule

The big question is whether to go for a hard launch (switch off the old and switch on the new) or a soft launch in which both applications are available. The factor that shapes the rollout plan is how much training and support resource is available. If there is a significant upgrade in capability compared to the current application, then users will need support to get the best of the initial version. The initial group of users will also be a test group, so it is not just a question of supporting their use of the application but of collecting feedback on how to improve the search experience.

The launch plan also needs to take into account the business cycle. Expecting users to spend time learning the new search application just as the annual business planning cycle starts is probably not a good idea. There could be other upgrades and system launches taking place which should also be taken into consideration. In public companies, the communications team will have more important priorities to deal with than the launch of a new search application.

Even in quite large vendors, the team responsible for installing the application is quite small, as specialist skills are needed to cope with the intricacies of the enterprise architecture of the customer. The availability of this team has to be matched to a period of stability with both the content repositories and applications that are going to be indexed by the search application. What might seem to be quite an innocuous upgrade could have a major impact on connector performance.

The overall schedule is then fixed by two dates.

Before the project can start, the following tasks must be fulfilled:

- Project manager and search support team in place
- All necessary hardware and access permissions have been established
- Work packages have been agreed on and respective roles of vendor and customer have been established
- A full content scope and audit has been completed
- Vendor and customer are confident that security management issues have been addressed
- A communications plan is in place for all stakeholders

A launch date can only be set under the following conditions:

- Enough content and functionality is available for users to have a good initial search experience
- The outcomes of the initial usability tests have been assessed and changes made as appropriate to the search application
- There are enough resources to support the launch, including training, help desk, and analysis of the initial set of search logs
- No other business applications are being launched
- The business cycle is at a point at which users will have either the incentive or time to make use of the search application (i.e., you should avoid launching during a business planning phase)

- There is still enough time to sort out any major problems before a full release to all employees is undertaken

The most challenging search implementation projects are ones that require content migration to be managed at the same time, perhaps from legacy Notes databases or from Microsoft SharePoint 2007 or SharePoint 2010.

## Minimum Viable Search

One of the elements of the lean startup methodology is the minimum viable product. This is defined by Eric Reis (one of the leaders in lean startup approaches) as a version of a new product that allows a team to collect the maximum amount of validated learning about customers with the least effort. To that definition I would add that the MVP also needs to be at a point where if there is significant difference between user expectation and delivery then making the changes does not significantly change either the budget or the schedule.

Adapting this to a search implementation, a minimum viable search (MVS) could be restricted to:

- Searching only a specific (but representative) repository
- Searching content forward from a specific date
- Only using the lowest level of search security
- Restricting the number of facets and filters

The discussion around the MVS can be a valuable way of focusing in on core user requirements. Out of the discussions might come a roadmap that suggests that the MVS is quite capable of being a near-term solution and not just a project step. The danger is that having implemented the MVS, the decision is taken to wind down the project because the MVS is "good enough."

The chances are that there will then not be enough time to undertake all the work needed to deliver the required search experience, so an iterative approach to project scheduling will be required. The point is that beginning the project without being fully prepared and launching the application without adequate content, functionality, and support is not to be recommended. Nothing travels faster than news that the launch of a new application, especially one that will be used by the majority of employees, has all the hallmarks of a disaster.

## Knowledge Transfer

There is a lot to learn both about the process of search and the technology of the new search application. The worst possible approach is for the vendor team to work in isolation from the search support team. The team needs to have a good understanding not only of what the search application can offer but also of how the search application works. Even more important is a clear understanding of what the search support team can change without needing to pay for additional consulting days from the vendor.

Forget all about search needing to be intuitive. That may be partially true for end users, but certainly not for the IT team working on the installation and implementation, and for the search support team. Search is complicated and there may well be a requirement to provide instruction on how search works before moving on to the specific elements of the search application that is being implemented. It is comparatively rare for organizations to implement a new search application, or even undertake a major upgrade to an existing implementation. As a result, there may be few people in IT who have had any direct involvement either with the current application or with a search implementation in other organizations.

There should be an initial one- or two-day training course for the search support team and the project team on the architecture and functionality of the search application so that issues that arise in the course of the implementation can be put into context. Just providing a pile of documentation is not useful.

In addition to this introduction to the project, a meeting with a reference client whose implementation has been carried out by the same project team is very valuable. This is so important that it should be written into the contract. It is all about ensuring that there is a high level of commitment and competence among the vendor project team members. All projects run into unforeseen problems—the point of interest is to judge how well these problems were addressed. If the vendor is reluctant to come up with a reference client, then a proposal to use the LinkedIn Enterprise Search Professionals site to see if any customers have views on the implementation process they would like to share should result in a change of heart!

The most important knowledge exchange point comes at the time when the project team hands over to the search support team. There ideally should be members in common between the teams, but it may well be that the implementation project team has a strong IT contingent that is unlikely to be required post implementation. It can be useful to agree at what point the transition for responsibility is going to take place. The day that the search application goes live is not the time to transition the responsibility, as it may take perhaps one or two months to eliminate the initial round of bugs, many of which will probably arise from crawl and index management problems, and from capacity management when employees discover just how good the new search application really is.

## The Show Stoppers

Two elements of the implementation project have the potential to be major showstoppers, and if they are not managed well, could jeopardize the success of the project:

- The content to be indexed is not as described in the content audit
- The security model is not as it was described in the initial statement of requirements

The vendor will have carried out some due diligence prior to confirming the contract price, but will almost certainly not have had the time or support from the organization to do a deep dive into either of these two areas. They represent risks of the highest level of impact on the project, and solving problems that were not highlighted in the initial presentations will not only be costly but could mean that the search application cannot deliver to the expectation of users and stakeholders.

The potential problems arising from a failure to have undertaken a full content audit and a full disclosure of security management issues will be especially significant when the plan is to extend the search across multiple applications.

## Get Indexing!

Because of the potential impact of content and security issues, it is essential to start indexing content at the earliest possible opportunity, even if it is with only a base configuration of the software and on open access collections. This approach will have the benefit of showing all the stakeholders that the investment in the search application will pay off.

However, indexing test collections also requires the customer to have some appropriate test queries. Otherwise, all that can be tested is the crawl and indexing performance. At the early stages of implementation, the results may not be encouraging, but without this early benchmarking, it will not be possible to track progress and if required, change elements of the implementation.

The initial implementation of the search application will almost certainly be on development servers. Load and performance testing should be undertaken following the migration to production servers, and this migration is rarely straight-forward.

## User Interface Design

The title is here as a check that due care is taken with the user interface design. The topic of user interface design is covered in Chapter 12.

## Usability and Accessibility Testing

When a project schedule slips, the work packages that almost always get cut are the usability tests on the user interface design. Usability tests are as important as any other element of the search implementation process, and setting high standards for the tests at the implementation stage provides benchmarks for tests undertaken later in the life of the search application when new applications or new facets are added to the search scope. Accessibility is also important to test. This is the extent to which users with a range of visual and physical disabilities are able to use the search application. Often the terms *usability* and *accessibility* are used interchangeably, but they refer to different aspects of the user experience.

It is not enough just to define a period of time when usability tests will be undertaken. The tests may reveal elements of the search implementation that need to be reassessed and possibly changed. Search applications are very susceptible to changes in one element subsequently impacting other elements in unforeseen ways.

## Disaster Recovery Tests

Effective disaster recovery is essential in a search application because there is a significant danger of content being crawled but not indexed. It is important to actually test the disaster recovery procedures under real-life conditions. Search disaster recovery is often put at the bottom of the list of priority applications to restore. However, a search application will still be able to identify information from its index even if the core application is not running.

## Help Desk

The implementation process will touch a lot of other applications and the IT Help Desk team needs to be involved at the earliest possible opportunity. The technical team from the vendor needs to talk in technical jargon not only to the Help Desk team but to other IT specialists. Using the project manager to relay messages is going to confuse rather than communicate. Servers, in particular, tend to have shorthand descriptions linked to a Configuration Management Database.

One of the few pleasures that IT managers have is devising naming conventions that are unambiguous to internal staff but have no meaning at all to external staff to reduce any chance of inadvertent or deliberate hacking. Every server that might somehow be affected by the search implementation needs to be identified. This is especially important when indexes and repositories are maintained on a remote and/or virtual basis. The name that corporate IT uses for a server in India could be very different from its local description. This should not be the case, but all too often it is.

## Metadata Management

This is covered in Chapter 9.

## Communications Plan

From the moment the contract is agreed, a communications plan needs to be implemented, which means that it should have been developed well before the contract is signed. There is a role here for Internal Communications to use every possible communications channel available to spread the news about the objectives and progress of the project. The communications plan needs to include ways in which employees can have concerns answered and be able to make contributions to the progress and outcomes of the project.

Search teams have a tendency to want to stay invisible, usually because they know that they do not have the resources to respond to all the requirements of users. By this stage of the book, I hope that the message that a skilled support team is essential will have been understood and acted upon. A search blog can be a very effective way of communicating the status of a search implementation.

This level of communications activity may well not have been used for the new finance system or the new customer relationship management system, but these applications were only used by relatively few employees. In the case of search, everyone with access to a desktop PC, a smartphone, and/or a tablet will be looking to assess the outcomes of the investment at the earliest opportunity.

## Migration and Search Implementation

The challenges of migrating a website or an intranet to a new CMS, or just a new information architecture, are considerable. There are automated tools that can support the process, but almost inevitably there will need to be a great deal of discussion and work around some specific areas of the site. An emerging requirement is to move file shares and other repositories into a cloud storage and application environment, Google Drive in particular. Any change in intranet structure as the result of a migration will make effective search a vital adoption route until users have found their way around the IA, and possibly new content types and repositories. Although little attention might usually be paid to search in a migration project, as search becomes more business-critical, it is increasingly important to include search-related issues at the outset of planning migration.

There are two content migration options. Moving sections of content to the new site as the content is revalidated certainly spreads the effort out over a period of some months, but the IA and the search application need to be in place from the moment the first content is migrated so that it can be tested in situ. Although a lot of usability testing can be carried out on a sectional basis, there will not be much value in searching the gradual accretion of content until most, if not all, is migrated. However, this approach will provide an opportunity to assess the user interface design and evaluate metadata schemes. There may be particular sections of content where search is especially important (the people directory or sections on corporate policies and guidelines) where an early assessment of recall and precision can be made.

The second option is to work on each section of content "offline," but then move it across in a single process. This will allow additional time for the design and implementation work to be carried out, but any problems with the search functionality may be difficult to pin down. The results for a given query may include some obviously spurious hits, but the reasons for this might take a lot of time and effort. If it is a problem related to inconsistent metadata, there might even be a need to review a substantial section of content. It is also not possible to assess latency and other IT-related performance issues (e.g., security trimming) until the entire site is migrated and perhaps linked via search to other applications.

Moving content into a cloud application also brings up the issue of which search application to use. On-premise search applications may be difficult, if not impossible, to link to a cloud application. In my experience, search requirements are rarely included in the initial specification because the driver is one of reducing cost and "improving collaboration," with little or no thought for how the content is going to be discovered. Sorting out search analytics is also going to be an interesting task!

It all comes down to very careful planning starting before the migration project is scoped out. Waiting to optimize search until all the content has been migrated is potentially risky. The degree of search importance, or perhaps complexity, should be a factor in deciding on the migration schedule and on the training and other support needed during the migration.

The planning needs to extend beyond the notional close of migration because this will probably be several months before the full launch. In the pre-launch period, search will need to be stress tested with a very close monitoring of search analytics, especially if the opportunity has been taken to enhance the scope and quality of the content. Some content may also have been deleted, which may also come as a surprise to some search users, not just because of its disappearance but the resultant changes to the ranking of results.

With enhanced content quality and better metadata, search performance should improve. However, a change in the search application interface or enterprise migration unexpected results appearing on the first few pages of search results could have a serious impact on business performance and undo all the positive expectations of the migration.

In the case of upgrading from SharePoint 2010 to SharePoint 2013, there is much to be said for leading off with the search implementation. This involves establishing a new services farm with just the services needed for search. Content from existing farms is then crawled and connected to the new search service so that users can start to explore and benefit from SharePoint 2013 search as each content farm is migrated.

## Summary

If the installation and initial implementation are not undertaken with a high level of care and resource allocation, the value of the investment in the application will be jeopardized. Implementation planning should be started right at the commencement of the project and not at the point of discussing the contract. The creation of test collections is an important element of benchmarking technical and end-user performance. The implementation project team will gain a very good understanding of the application during the course of their work, and it's very important to ensure that the knowledge gained is transferred to the operational search support team.

## CHAPTER 15

# Search Evaluation

Very rarely is the performance of an enterprise application measured in any detail. Invariably, these applications support a task or workflow and the efficiency is assessed in terms of whether the overall task could be accomplished. These applications are usually built on SQL or a similar database, and methods of optimizing these databases in performance terms are well understood and generally well implemented.

Search is different. Every user at every visit is at a different stage in undertaking a task, and somehow the search application has to deliver what is needed on a highly consistent basis. Recall and precision are both defined in terms of relevance, and yet the "relevance" of a piece of retrieved information is totally dependent on the personal context of the user. Some users might have an inadequate understanding of a concept and so may create an inappropriate query or not be able to judge the relevance of the results presented to them.

Far too many presentations at conferences focus on "improving relevance" but then go no further in considering the difference between recall and precision. Often the emphasis is on precision, seeing the document required on the first page of results, because some other application (typically an intranet) has failed to deliver the information being sought.

Search assessment needs to be undertaken on a carefully considered basis. Ad hoc analysis of search logs will give neither short-term nor long-term benefits. The primary reason for assessing search performance is that search is either a starting point for information discovery or a resource that is used when all other options have failed. In both cases, excellent search performance is essential, and this is why a search support team is so important in ensuring that employees at all levels can find the most relevant and trustworthy information.

One of the problems of managing search is that there is a cycle of circumstances that is difficult to break. There is usually not enough search management resource to allocate enough time to search analytics and yet without the search analytics, it is not possible to make a quantitative business case for additional resources.

Since the beginning of search, the review of search logs to assess performance has often been referred to as search analytics. However, the term *analytics* is increasingly used to refer to the use of analytical applications to identify patterns of information in large text and data sets. When someone remarks that the organization has a great analytics application, you can never be sure if they are referring to search log analytics, text analytics, or data analytics.

## The Five Components of Search Evaluation

There are five components to search evaluation:

- Technical performance
- Query performance
- Usability and accessibility
- Search satisfaction
- Business impact

Each of these needs an appropriate level of assessment experience and resources to build up an overall picture of how well the search application is performing. Even for a small organization, this is an appreciable amount of work, but is essential if users are going to trust the search application to deliver the information they need to make effective decisions.

## Technical Performance

Users expect an acceptable and consistent response time to a query. There are many factors that affect the performance of a search application, which for the purposes of this chapter is defined as providing the user with an acceptable and consistent response to a query. The delay between initiating the query and being presented with the initial page of results is described as the search latency. Latency can be measured on the server side or the client side, where network performance also has to be taken into account. The latter metric, usually referred to as end-to-end latency, is the most appropriate metric to adopt.

Comparatively little research seems to have been published on what might be acceptable and consistent response times for enterprise search applications where one of the main causes of result list delay is the need to check on the user's security permissions. Another factor is the time taken to download a document from a remote server. A consensus view seems to be that a delay of longer than 500 ms starts to be noticeable, and that once the delay starts to get to around one or two seconds, the user starts to be concerned about whether the search application is working correctly.

There are many other aspects of search performance that need to be considered, such as ingestion rates (the rate at which content can be indexed), freshness (the delay between content being posted to a server and the content being indexed and available for search), and throughput (the number of queries that the application can process in a specified period of time). These can all have an impact on search satisfaction, and service levels need to be set and monitored by the search team.

Crawling schedules need special care. The majority of users will be under the impression that content is being crawled and indexed as soon as it is posted to a server. This may be the case if new content is pushed to the application indexer, but the next crawl may not take place until the following day or even week. The result could be that a new content item is visible on an intranet but cannot be found by a search because of the delay in crawling the intranet CMS. Crawling is server-intensive, and crawling on a frequent basis may improve freshness but increase the latency to a point where users start to notice the delay.

## Query Performance

It is important to appreciate that the query terms used in a search do not necessarily indicate what information the user is searching for. The challenge of search is to cope effectively with "exploratory search," where users are not quite sure what they are looking for and just start somewhere, perhaps somewhere they are already familiar with. They may also recall from previous searches that there are useful hits on the home page of even a high-level query term.

In the case of a university, a faculty member may search for *chemistry* because she knows that on the first page of results there will be a list of internal seminars in the Department of Chemistry that are difficult to find by browsing through the site. From her viewpoint, it is a very effective use of search and she is very satisfied with the outcome.

The primary reasons for search performance are that the search application has to try to read the mind of the user when deciding what information to display in response to a vague query, and the search team has to be able to read the minds of multiple users when analyzing search logs.

Search logs are a critically important diagnostic tool. Without adequate log analysis, any search implementation is flying blind, potentially a waste of investment that has been made in the application. However, the capabilities of the analytics applications that are supplied with search applications vary widely in features and ease of use, factors that are often not considered in the evaluation of the application.

The list of possible metrics is quite lengthy:

- Percent of search sessions that result on a click on a search result
- Percent of searches that return zero results
- Percent of searches where users exit from the search results page without clicking on a result
- Percent of sessions that use search
- Average number of searches per session
- Average number of search result pages viewed for a keyword
- Average number of pages viewed after searching
- Average time spent after searching
- Average time spent before searching
- Average time spent on search results pages
- Session duration for all sessions that included searches

Some of these are more relevant to website search than internal search applications, but this is still a very helpful list of metrics. This is only a partial list. The fact that the standard text on search analytics by Louis Rosenfeld, *Search Analytics For Your Site* (Rosenfeld Media), runs to almost 200 pages is a good indication of the range of analytics measures that are available. The choice of which to use on a weekly, monthly, quarterly, or ad hoc basis needs to be considered carefully. The analytics program for a website may well be very different from those for a single internal application. Of course, in the enterprise, there may be multiple search applications, and gaining an overall assessment of search performance is a complex but essential task.

An element of query evaluation that is often overlooked is the value of conducting searches on a regular basis. A member of the search team will look at the process, from query formulation to search result, in different and more analytic ways than most other users. The search queries need to be "real" queries, not just queries dreamt up over a cup of coffee by the project team.

Search queries might include the following:

- A selection of the most highly used search terms
- A selection of search terms with low or zero results returned
- Search terms related to an emerging area of interest for the organization
- Terms that may be cyclic in use but for a short period of time (annual appraisals) may be used quite intensively
- Recent news items

- A query where the documentation may go back a significant period of time
- Corporate policies, to see how many versions are listed!

Arguably, test searches are more important for websites than enterprise applications as a poor search result may result in a loss of a customer engagement opportunity.

## Usability and Accessibility Tests

Usability tests are an important element in developing websites, but there is much less support for usability testing for intranets, search applications, and other enterprise applications. Search usability, in particular, seems often to be totally ignored. One of the excuses that is often given is that because everyone has their own view of what is relevant, there is no point in carrying out usability tests. That misses the point about search being a dialogue and the need to ensure that the search interface supports the dialogue not only up to the point of a set of results being displayed, but then supports the user in reducing the total number of results down to a manageable list to browse through.

The other common excuse is that usability is too time consuming and expensive to carry out. The response to this has to be that the company could be at risk from an employee not being able to find the information that could have made a significant impact on business performance. If the company is willing to take that risk, then certainly there is no point in carrying out usability tests, but that risk could have a very public impact on the reputation of the company.

Because of the complexity of search user interfaces compared to the options offered on most web and intranet pages, usability is not an option but a necessity. The interface situation is going to become even more complex as employees make use of tablet and smartphone devices to access enterprise and web applications.

The major decision to make is whether the usability testing is going to be moderated or carried out remotely. In the case of search usability testing, it is probably better to use moderated tests because of the complexity of the user interface and because there is no "optimal" approach to finding the information required. An important issue with search usability testing is the test subject's evaluation of the list of results. Even if the notional completion of a search task is to find a specific piece of information, perhaps the office address in Vilnius, a user may be annoyed to find that it is on the third page of the search results. This is a good example of the difference between fitness to specification (it can be found with the search application) and fitness to purpose (because the office number is promoted to the top of the results page).

Another challenge with search testing is that it may be necessary to test at various levels of security permissions. Information available to a partner in a law firm is almost certainly not going to be available to anyone, no matter how senior, who is not a partner.

Accessibility also needs to be checked on a regular basis, perhaps by a specialist agency, to make sure that the search interface meets WAI guidelines.

## Search Satisfaction

A challenge with search is that even if the downloads from searches seem to be high, that may not correlate with long-term search satisfaction. It could be that a given document is downloaded many times but that a percentage of users is concerned about the validity of the information or the process of finding it may have been a time-consuming task. At the other end of the spectrum, a failure to find any relevant information on a topic in the past may mean that users are not willing to waste their time again but call or email a colleague for assistance and potentially waste their time.

The first step in achieving and then maintaining high search satisfaction is to provide good feedback channels from users to the search team. A feedback box for search is highly desirable, but for it to be effective, there should be a service-level agreement that indicates that the provider of the feedback will receive at least an initial response within (say) two working days. It might not be possible to fix a problem in two days, but the initial response will at least be a visible recognition that the search team is concerned about the problem that has been identified.

The feedback box ideally needs to capture the search term(s) from the screen code without the users having to enter the information themselves. Ideally, the name of the person to whom the web message is going to be sent should be included. Users dislike feedback forms where there is no channel back to the person who received the message to check that action is being taken or perhaps to report that the problem has been solved.

Consideration should also be given to having either a search blog to report on problems and solutions, or a wiki. Transparency from the search team about both problems and solutions (especially if the solution is not a relatively immediate one to implement) will encourage users to report not only problems but also when the search application has exceeded their expectations.

Another approach is to set up a panel of search users across the organization and poll them on perhaps a quarterly basis about their views on the search applications using a small set of questions that explore in qualitative terms issues around search performance, usability, content quality, search team support, and the contribution that search has made to their own business performance. The scoring could be either on a scale of 1–10 or as a delta change to their views three months previously. This panel could be very useful in monitoring the impact of changes to the user interface or the addition of new features, but in both cases, surveys of this type are no substitute for usability testing.

## Business Impact

A particular challenge with search is that in the short term, a user may be satisfied with the performance of a search application. The latency is low, the search interface is good, and in general, the results identify relevant documents. However, the true value of a document may not be apparent at the time, and indeed in due course it may become apparent that one or more of the relevant documents is not in fact relevant and/or credible. This is often the case when working with teams.

Every organization has team meetings, though increasingly these are virtual team meetings, which require substantially more planning. Teams tend to have regular tasks, such as providing monthly status reports on new projects, revising corporate policies, and tracking the activities of competitors. Sitting in on these meetings can help identify the types of searches that are carried out and what would be the desirable outcomes of the search process. The benefit of teams over focus groups is that team members will feel comfortable with each other and have a collective focus on certain corporate objectives that may well determine career development opportunities or compensation awards.

However, there is no point just turning up at the meeting and asking for input on search requirements in the Any Other Business section of the meeting. The program of attendances at the team meetings needs to be planned in advance. It is also advisable to have the discussion about search fairly high up on the agenda, so that it is positioned as an important topic. Having the discussion on the agenda also (hopefully!) ensures that attendees come prepared.

In the case of project teams, arranging the meeting at the start of the project will give an opportunity to discuss how best to go about searching for information in the course of the project. The project manager could be asked to keep track of the successes and failures of finding information and share these with the search team as well as including them in the project report. Always offer members of the team the option to talk individually about their search experience and requirements. They may not wish to disclose to their colleagues that they are having difficulty with the search applications.

As mentioned before, teams increasingly work and meet on a virtual basis, and this requires more preparation, as the attention span of participants may well be lower when taking part in a meeting which may have been scheduled at a time that is not totally convenient for them. On the positive side, as the attendees will be participating through a networked computer, it may be possible for them to demonstrate some of the aspects of the current search application that they would like to see enhanced.

Conducting user surveys with web-based survey tools has transformed the effort required to carry out large-scale surveys and have the results available in a short period of time.

There are some guidelines that should be taken into account in designing the search survey:

- Start out with no more than 10 questions, which will probably take a user around 10 minutes (or a cup of coffee) to complete. Anything longer will need very careful design.
- The questions should be intuitive, so that respondents gain an immediate understanding of why the question is being asked.
- Ideally provide an indication of how far through the survey a respondent has reached.
- Don't ask questions that rely on feats of memory about what the respondent did over a past period of time. "Do you use search now more than you did a year ago?" has no value at all.
- Don't expect respondents to write essays in a text box. Invite respondents to contact you if they would like to talk through issues in more detail.
- Recognize that it may be better to send out different surveys to specific user groups than try to accommodate the views of the entire workforce with a single set of questions
- If using Likert or Likert-like surveys, do not average out the scores. Use the median.
- Commit to summarizing the outcomes by a given date, and invite respondents to comment on the results.
- Test the survey, and then test it again.

Many organizations carry out an employee engagement survey on an annual basis, and it may be possible to add a question or two to this survey to assess the ability of employees to find the information they need to carry out their responsibilities. Internal communications teams are usually overwhelmed with suggestions from questions from departments, so it would be wise to have a good business case. One factor in favor of a question about information discovery is that it would apply to every employee.

As well as looking for evidence that the search application is working well, these surveys and interviews are invaluable in providing good stories for sharing across the organization or with specific senior managers. "Did you know that Susan Palmer was able to find some test results from that company we bought last year that saved us several weeks of work carrying out the testing ourselves?"

## Search Evaluation Planning

Something has to be done to increase the ease of search from just 15%–20% of users regarding it as easy to undertake a search, and a well-supported program of search assessment is going to be a better investment than the costs and implementation challenges of replacing the current search application and then still finding low levels of search satisfaction.

The starting place should be:

- Develop a set of search user personas and use cases
- Read *Search Analytics for Your Site*
- Ensure that the role and responsibility for search evaluation are written into the job description of the search support team
- Make sure there is a visible feedback box with a statement of when a response will be forthcoming
- Be clear about the balance between the five dimensions of search evaluation
- Decide on a core set of evaluation measures that would have a short impact on improving search

On a monthly basis:

- Ensure that search query latency is acceptable
- Check that the crawl and indexing processes are working correctly
- Create a report from the evaluation measures that focuses on the actions that need to be taken to continue to improve search performance
- Report on the outcomes of these actions

On a quarterly basis:

- Poll the group of users on their views on how search delivers what they are looking for
- Look carefully at search logs, taking into account search personas and use cases so that the context for search queries is understood
- Conduct test searches for the top queries to see how they could be improved
- Dig deep into searches that returned few or no results
- Attend some meetings of employees who are likely to make extensive and/or business-critical use of search

- Provide a concise report for managers on the outcomes of the assessment, the actions that are being taken or need to be taken, and the likely impact if no action is taken
- Publicize the results on the intranet or an internal social media channel and invite feedback and suggestions

There are two important sets of skills that are required to undertake even the most basic of evaluation programs. One set of skills involves taking large sets of data and presenting them in a way that clearly highlights trends. This will involve working with both the log applications from the search applications and probably moving them into an Excel spreadsheet for further analysis.

The second set of skills is more akin to mind reading. The trends have to be assessed and actions agreed based on a very good understanding of the organization. The search team member responsible for highlighting actions that need to be taken based on the statistical and survey evidence has to know the language of the organization and be able to spot search query terms that seem odd and then know who to contact to see if these new or curious terms represent a new area of business where search needs to play an important role.

In addition to the business language of the organization, the range of languages used by employees needs to be supported. Some of these language issues can be quite subtle. For example, Brazilian Portuguese and the language as used in Portugal have different words for the same topic. A bus is *ônibus* in Brazil, but *autocarro* in Europe.

It is difficult to make a general statement of search team size but for an organization with more than 5,000 people, search assessment needs to be a full-time task for one person. If there is a substantial use of search in a second or third language, then each of these languages needs the full-time attention of a search assessment because employees may also be searching in English with poor query construction, and that also needs to be identified and addressed.

Over the next couple of years, the complexity and value of search assessment is going to increase substantially as enterprise mobile search and collaborative search both start to be widely adopted in organizations, and also as search becomes even more important as a means of coping with the very rapid growth in data and information even in smaller organizations.

## Summary

For most enterprise applications, either it works well and supports a specific work process, or it is broken and needs fixing. Search is very different, as it may not be at all obvious what has broken. By the time someone calls the help desk to find out what has gone wrong with search, the organization could have lost out on a major contract opportunity. Time is of the essence in identifying how search can be enhanced on perhaps even a daily basis. Search analytics are important but (as with Big Data) may not provide guidance on a solution. That will come from the qualitative surveys and feedback routes in discussion with the search team and perhaps subject experts in the organization.

## Further Reading

Nielsen Norman Group, "Turn User Goals into Task Scenarios for Usability Testing", January 12, 2014.

Jeff Sauro, "20 Tips for Your Next Moderated Usability Test", April 10, 2012.

In addition to these resources, the entire Rosenfeld Media catalog is concerned with user experience and user requirements research.

## CHAPTER 16

# Website Search

It is not uncommon to find that there are no strategic, operational, or managerial connections between the choice and implementation of website search and the enterprise search team. This is either because the website search function is embedded in the chosen CMS and/or because the department owning the website has its own budget and support team. All too often the search feature of websites looks as though it has been added as an afterthought rather than as an integral part of the design. This is most notable in the poor quality of the presentation of search results, with usually inadequate information to enable a site visitor to select a few results that seem to be most relevant to their requirements.

Website search is likely to be of most value to a site visitor when the content on the site meets one or more of the three characteristics often associated with Big Data:

*Volume*

The amount of content on the site is so large that any information architecture is going to struggle in presenting the content within a reasonable number of clicks (this could well be the case for research sites that are online archives for public sector organizations with content dating back many years into the past)

*Velocity*

The content changes at such a rate (news or financial information) that it is very difficult to refresh the information architecture to take account of new topic areas and a topic that was hot news two days ago has all but vanished

*Variety*

The content of the site is so diverse that the information architecture is going to struggle to maintain a reasonable level of performance

Most of the chapters in this book apply to both enterprise-facing and customer-facing sites, so this chapter just picks up on some issues that are of especial importance to customer-facing sites. The way in which search is embedded in ecommerce sites is a very important topic, but we will not cover it here (for more information, refer to the books listed in "Further Reading" on page 223).

There are a number of obvious differences between website and internal search that have an important bearing on the way in which the search functionality on the site is specified and managed:

- For website search, the user population is infinite, potentially global, and unknown. In an enterprise environment, it is very easy to survey and talk to users, using the range of performance evaluation methods in Chapter 15. This is why developing personas is so important for a website, and arguably there should be an even greater emphasis on targeting specific groups of users.
- Site visitors are probably less likely to have a good command of English than might be the case for users of enterprise search applications.
- The technology options will be different. Rarely will a case be able to be made to invest in a standalone search application, and so the range of functionality and of metrics may be less than ideal.
- There will almost always be another way that frustrated site visitors will be able to find the information. They could use Google or go to another site. You have no way of knowing which.
- Feedback from site visitors will be negligible.
- A high level of accessibility is very important with website search because of the wide range of users.
- For website search, ensuring that search functions well on a mobile device is paramount.
- In general, site visitors will not have enough familiarity with the organization to use filters and facets which presuppose that site visitors are also employees.

All these need to be taken into account in the strategy for website search.

The UK Government Digital Service has done a great deal to improve the quality of government and other public sector websites and has published an otherwise excellent Government Service Digital Manual, but there is no specific reference in the manual to website search.

The US government does have a website dedicated to usability. The guidelines section was relaunched in May 2015, but there are no specific guidelines on search.

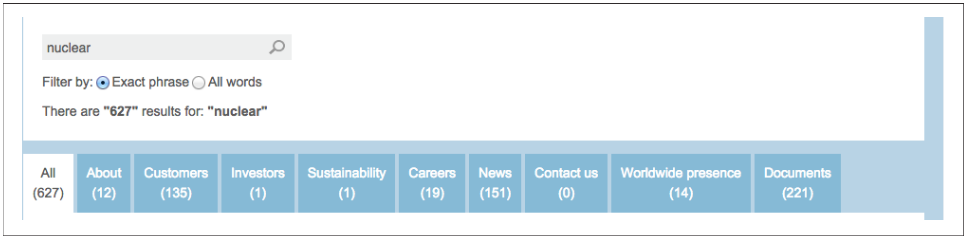
This lack of focus on search is quite symptomatic of website managers failing to understand the importance of website search. Much of the content of this book is broadly relevant to website search. The aim of this chapter is just to highlight some issues that are of particular relevance to website managers. The special case of ecommerce site search is not covered. I have put some good references on this topic in "Further Reading" on page 223.

# First Impressions Count

In the case of a website, the competition to website search comes from Google, Bing, and other web search services. For good reason, site visitors have a strong case to ask why the search feature on a website does not work as well as Google. They might well have found the site via Google and then want to drill down further with the website search without feeling they are moving back through time to the early days of search. The evidence suggests that site visitors make up their minds about the site experience within a minute or so of arriving at either the home page or destination page from a web search. The first experience of the search feature might well be the last. To illustrate good and poor elements in website search design, I have selected three organizations which in the United Kingdom could be regarded as exemplars in their industry sectors.

Rolls-Royce has a global reputation for engineering quality, but judging from the first page of search results, the desire for quality does not extend to its website.

A search for *nuclear*, a specialization of Rolls-Royce, results in 627 results. As Figure 16-1 shows, like many websites, the counts by category are organized in a set of horizontal tabs.



nuclear

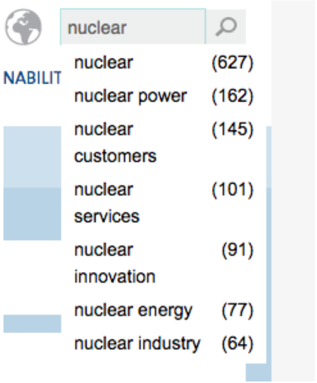
Filter by: ● Exact phrase ○ All words

There are **"627"** results for: **"nuclear"**

*Figure 16-1. Tab of results by category on the Rolls-Royce website*

The sum of the tab counts is 554, not 627, so an initial question might be to ask what happened to the missing 73 results. When organizing results by category, the two obvious options are to list in alphabetical order or by the number of results. In the case of Rolls-Royce, the sequence is random. Moreover, the same tab width is given to Contact Us (0) and Investors (1) as Documents (221). These tabs are, in fact, common to all search results pages.

The issue here is that the tabs do not help a site visitor move down into more detail about the nuclear industry. If there are 221 documents, what do the other 406 results represent? To be fair to Rolls-Royce, there is a typeahead feature on the search query box (Figure 16-2).



nuclear

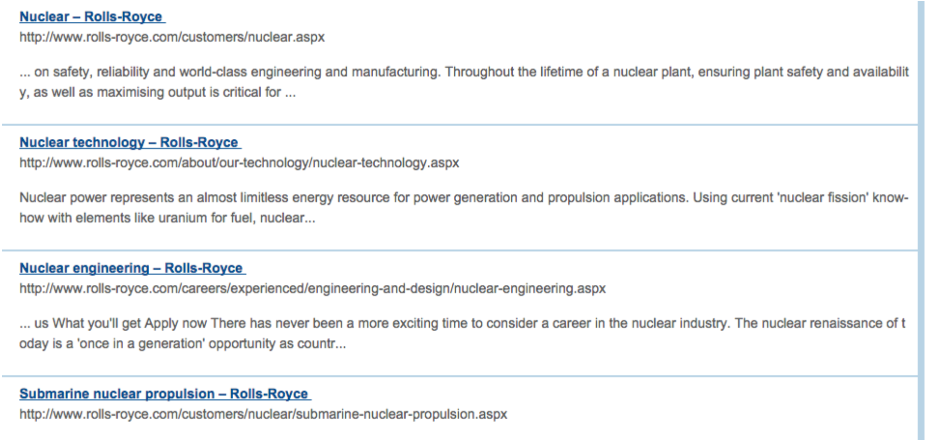
NABILIT

| | |
|---|---|
| nuclear | (627) |
| nuclear power | (162) |
| nuclear customers | (145) |
| nuclear services | (101) |
| nuclear innovation | (91) |
| nuclear energy | (77) |
| nuclear industry | (64) |

*Figure 16-2. Typeahead options box on the Rolls-Royce website*

But why aren't these options available once a top-level list of results has been displayed?

The results page provides no useful information about the content that has been declared as relevant. Figure 16-3 shows the lower section of the first page of results. The most obvious omission is any indication of the date of the content. Moving on down to the bottom of the first page of 10 results, two identical results are presented, differing only by the addition of *.aspx* on the URL.



Nuclear – Rolls-Royce
http://www.rolls-royce.com/customers/nuclear.aspx

... on safety, reliability and world-class engineering and manufacturing. Throughout the lifetime of a nuclear plant, ensuring plant safety and availabilit y, as well as maximising output is critical for ...

Nuclear technology – Rolls-Royce
http://www.rolls-royce.com/about/our-technology/nuclear-technology.aspx

Nuclear power represents an almost limitless energy resource for power generation and propulsion applications. Using current 'nuclear fission' know-how with elements like uranium for fuel, nuclear...

Nuclear engineering – Rolls-Royce
http://www.rolls-royce.com/careers/experienced/engineering-and-design/nuclear-engineering.aspx

... us What you'll get Apply now There has never been a more exciting time to consider a career in the nuclear industry. The nuclear renaissance of t oday is a 'once in a generation' opportunity as countr...

Submarine nuclear propulsion – Rolls-Royce
http://www.rolls-royce.com/customers/nuclear/submarine-nuclear-propulsion.aspx

*Figure 16-3. Section of Rolls-Royce website search results page for "nuclear"*

The next result in Figure 16-3 is, perhaps, the most surprising of all. Remember that the query was *nuclear*, indicating a very exploratory search. This result is for a brochure on a very sophisticated valve control mechanism for nuclear power stations, none of which are being built in the United Kingdom at present. For some reason, the relevance ranking has positioned this brochure at level 8 in 627 results. It is probably a reasonable assumption that anyone who knows what this valve does already has the technical brochure.

The important issue here is that once a user begins to question the logic behind the relevance ranking, they will be disinclined to continue looking at further pages of search results. Rolls-Royce is rightly regarded as one of the world's leaders in advanced engineering and production quality of the very highest level, but the search functionality of the website could certainly be better.

Moving on to the public sector, the Office of Communications is a UK non-governmental organization that regulates the TV and radio sectors, fixed-line telecoms, mobiles, postal services, plus the networks over which wireless devices operate. Figure 16-4 shows a list of search options on its website.

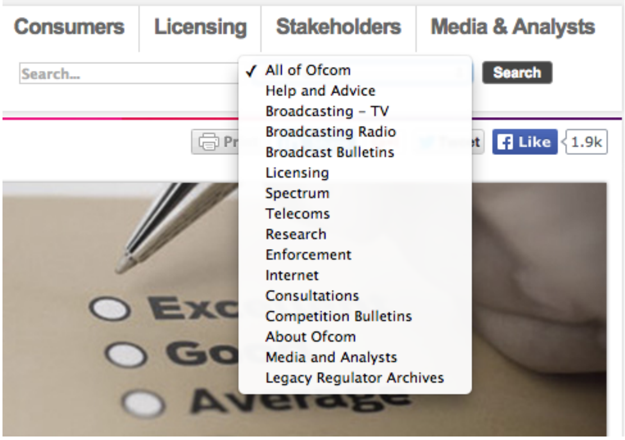
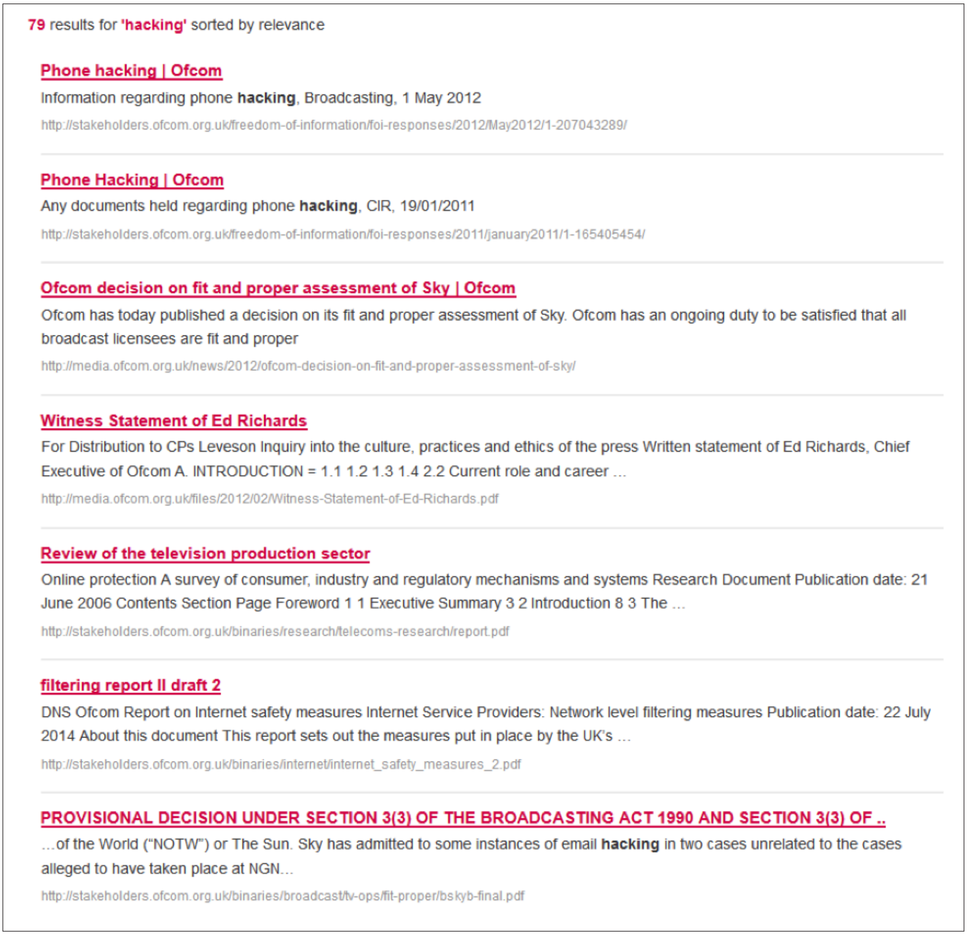


*Figure 16-4. List of search options for the Ofcom website*

Over the last few years, Ofcom (as it is known) has been heavily involved in the fallout from the discovery that some journalists were hacking into the mobile phone message services of celebrities, politicians, and the Royal Family. Beginning a search for *hacking* generates a list of possible places that a visitor might wish to search through.

One fundamental problem with any list of this type is that the average visitor has no idea of where the information they are looking for might be found. Indeed, that is the reason why they are using the search function. The other problem is that there is no capability to search across more than one category of documents. Users are faced with the requirement to conduct several separate searches and then somehow de-duplicate and integrate the results. This is an unreasonable requirement.

Ignoring this option brings up the first page of search results shown in Figure 16-5.



79 results for 'hacking' sorted by relevance

Phone hacking | Ofcom

Information regarding phone **hacking**, Broadcasting, 1 May 2012

http://stakeholders.ofcom.org.uk/freedom-of-information/foi-responses/2012/May2012/1-207043289/

Phone Hacking | Ofcom

Any documents held regarding phone **hacking**, CIR, 19/01/2011

http://stakeholders.ofcom.org.uk/freedom-of-information/foi-responses/2011/january2011/1-165405454/

Ofcom decision on fit and proper assessment of Sky | Ofcom

Ofcom has today published a decision on its fit and proper assessment of Sky. Ofcom has an ongoing duty to be satisfied that all broadcast licensees are fit and proper

http://media.ofcom.org.uk/news/2012/ofcom-decision-on-fit-and-proper-assessment-of-sky/

Witness Statement of Ed Richards

For Distribution to CPs Leveson Inquiry into the culture, practices and ethics of the press Written statement of Ed Richards, Chief Executive of Ofcom A. INTRODUCTION = 1.1 1.2 1.3 1.4 2.2 Current role and career ...

http://media.ofcom.org.uk/files/2012/02/Witness-Statement-of-Ed-Richards.pdf

Review of the television production sector

Online protection A survey of consumer, industry and regulatory mechanisms and systems Research Document Publication date: 21 June 2006 Contents Section Page Foreword 1 1 Executive Summary 3 2 Introduction 8 3 The ...

http://stakeholders.ofcom.org.uk/binaries/research/telecoms-research/report.pdf

filtering report II draft 2

DNS Ofcom Report on Internet safety measures Internet Service Providers: Network level filtering measures Publication date: 22 July 2014 About this document This report sets out the measures put in place by the UK's ...

http://stakeholders.ofcom.org.uk/binaries/internet/internet_safety_measures_2.pdf

PROVISIONAL DECISION UNDER SECTION 3(3) OF THE BROADCASTING ACT 1990 AND SECTION 3(3) OF ..

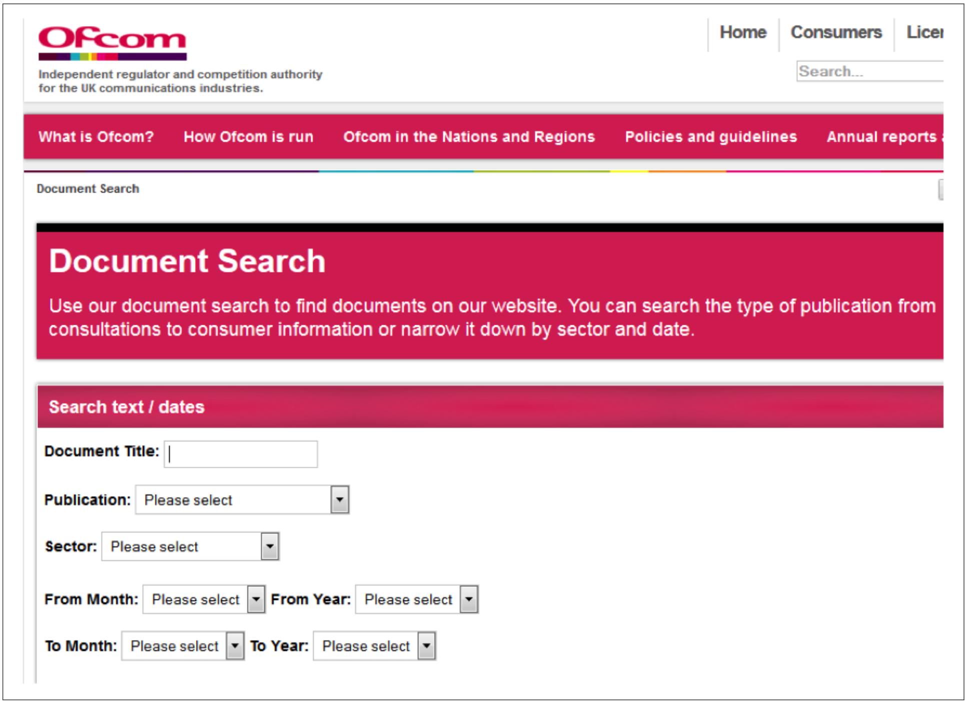
...of the World ("NOTW") or The Sun. Sky has admitted to some instances of email **hacking** in two cases unrelated to the cases alleged to have taken place at NGN...

http://stakeholders.ofcom.org.uk/binaries/broadcast/tv-ops/fit-proper/bskyb-final.pdf

*Figure 16-5. Search results for "hacking" from the Ofcom website*

The first two look promising, but in fact are PDFs of responses to Freedom of Information requests. In both cases, they are copies of letters written to people who had filed a Freedom of Information request telling them that Ofcom did not hold the information they were asking for. Yet these are deemed to be the two most relevant results. The fifth most relevant result refers to a document on the television production sector dating back to 2006! My favorite result (not displayed here) has the title of 97.

I have been using this site as an example of poor search practice throughout 2014 and 2015. However, recently, I noted that a document search was now available, as shown in Figure 16-6.



Independent regulator and competition authority for the UK communications industries.

Document Search

Use our document search to find documents on our website. You can search the type of publication from consultations to consumer information or narrow it down by sector and date.

Search text / dates

Document Title:

Publication: Please select

Sector: Please select

From Month: Please select From Year: Please select

To Month: Please select To Year: Please select

*Figure 16-6. Document search query page from the Ofcom website*

However, this search requires users to know the document title, publication, sector for the document, and when it was published. The Publication option lists document types in random order, including Direction, License Process, and N/A. N/A?

The website of the scientific journal publisher, Nature Publishing Group, is perhaps the best example of a search application with a deep understanding of site visitor requirements and high-quality delivery. You probably need to be a scientist to fully appreciate the site, but as a trial, search for *nanotechnology* and consider both the way a range of resources is presented and the scope of the advanced search feature.

## Personas and Use Cases

Many websites are built around the concept of personas. The definitive book on the development of website personas, *The User Is Always Right*, authored by Steve Mulder and Ziv Yaar (New Riders), defines them as follows:

> A realistic character sketch representing one segment of a website's target audience. Each persona is an archetype serving as a surrogate for an entire group of people. Personas summarise user research findings and bring that research to life in such a way that a company can make decisions based on these personas, not based on their own perceptions of what users require.

Typically, four or five personas are developed for a website and can be immensely valuable in helping to shape the overall look and feel of the website. However, the extent to which personas are able to be used to define search requirements may often be limited.

The following are some excerpts from personas taken from what in other respects were a well-developed set of personas created by a design agency for a UK public sector organization.

- "Search facility that recognises the terms he is using"
- "Sophisticated and reliable search offering a variety of data cuts and option to drill deep"
- "Sophisticated search facility allowing variety of information cuts"
- "Facility to search by topic with search results indicating the file format"

Words such as "sophisticated search" may well have been used in the research process, but just what the user means by "sophisticated search" is rarely explored due to a lack of time and understanding of search. Design agencies may not be fully aware of the opportunities and complexities of search, and so may fail to probe deeply enough in the research process or focus just on user stories rather than on more rigorous persona stories.

## Integrating Search into the Design Process

To pick up on this design issue, it is important that the process of developing a new website or enhancing a current site is not totally focused around optimizing the navigation. The design process will begin with using the personas, and associated use cases, to develop an initial information architecture. This will then be tested iteratively with sample groups of users as wireframes and then through tree testing. At no point will there be any consideration of how search and navigation will work together, and, of course, wireframes cannot replicate the outcome of a search. In my experience, when user testing on the site is carried out, the design team gets very upset if a tester decides to use the search feature and immediately rushes off to redesign the architecture so that it never happens again.

Search may well be used to find a section of the website that can then be explored in more detail through browsing. There are no easy ways to do this, but a good place to start is to ensure that either the personas include a specific reference to search, or that a separate set of search personas are created, in both cases using one of the many possible information discovery modes that are outlined in Chapter 3.

It is also advisable to test out the search application as early as possible in the redesign or upgrade process, not just to assess the relevance of the results achieved but to optimize the search page itself and the level of detail that will be included in each search result.

## Technology Options

A significant difference between websites and enterprise applications is the widespread use of hosted search services, either from Google or from the increasingly wide range of smaller hosted search vendors, many of whom are using open source search applications. Another option is to use the search application that is embedded in the website CMS, such as EPiServer. The website CMS search application is inevitably optimized to search content contained within the CMS, so finding information contained in other integrated repositories (perhaps a staff directory or a list of publications from a SharePoint server) may be difficult, if not impossible. Sometimes the integration is possible at a technical level to produce a single index, but the ranking of the results could well be unhelpful to a site visitor.

Search latency is also something that needs to be assessed if the search is outsourced to a third-party service. This need not add substantially to the search result display time, but is still worth checking at various times in the working day, especially if the service is located in a different time zone and therefore a different range of access peaks.

It is not usual to find comparatively little experience of evaluating search technology options for a website, especially in smaller organizations. Aside from recommending the use of a consultant, it is essential to look at some reference sites and talk to the web team about their experiences. The discussions should extend to the way in which customized search logs can be generated, the ease with which search logs and website click logs can be integrated, and the level of professional services support from the vendor. This is just as important to do with Google as with any other vendor. Google technology is certainly impressive for websites, but does it deliver what you have decided would be the optimal search experience for your site visitors? The Google Enterprise Search Appliance is also an option for a website. It does bring search in house, but care needs to be exercised over the total operating costs of the appliance.

Another option would be to have the search function built around an open source search application. This would be a very standard type of project for any open source search development team, and would be the ideal approach to adopt if there were any plans to search content that is not held in the CMS.

## Advanced Search

The Nature Publishing Group example discussed earlier illustrates the fact that in certain circumstances some form of parametric search (drop-down options) may be helpful. The example from Ofcom shows that unless it is very carefully designed, it will be of no assistance at all. The default should always be to satisfy the most of the search requirements of site visitors with the primary search box. If there is a specific persona that would benefit from using some form of parametric search, a business case should be made for doing so. In a public sector website, it is quite possible that lobbyists and policy makers may wish to take the time and effort to find all the relevant documents.

## Mobile Search

There can surely be no website manager that regards the effective use of the corporate website on a mobile device as irrelevant. The options for displaying the website are either to create a mobile version (the approach adopted by the BBC) or to use responsive page design coding to progressively adapt to mobile devices. Responsive page design will work well for standard web pages, but providing even a basic search capability is another matter. The fundamental problem is that the search box is on the far righthand side of the screen and yet usually the most important information for site guidance is on the lefthand side, given the Western requirement to read from left to right. It is interesting to note that *The Guardian* website (Figure 16-7) has the search icon on the lefthand side of the screen so that it is more visible and usable on a mobile device.

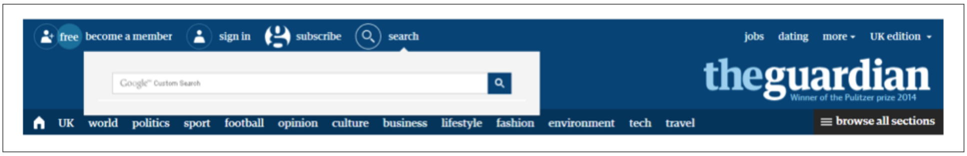


*Figure 16-7. Search box on The Guardian website*

However, even this is only a partial success, because the autosuggest feature provides quite long query suggestions, and selecting these with a finger on a mobile device while standing in a London bus (as an example) is a challenge in dexterity.

Only fairly recently has research started to be undertaken on the relationship between searching mobile and desktop devices for a common query. A team from Microsoft has looked into the transition from desktop to tablet and then phone, and a team from Yahoo! has investigated the relationship between the use of mobile search and the use of apps. There will undoubtedly be more in the future, but at present, there is a lot of anecdotal advice available, some of it well grounded in large-scale implementations, but as yet there is no A/B testing research.

## Search Your Own Site

The most valuable piece of advice I can offer to website managers is to carry out searches on the website on a regular basis. There is a tendency to focus on adding in new content and revising existing content, which is, of course, very important. There is little point in doing this work if it is difficult for the content to be discovered, especially through the search box. At a regular time each week, it can be revealing to carry out searches on the following types of queries:

- Queries on the most frequently used terms, as these may indicate that important information is otherwise difficult to find through the site architecture
- Queries on current hot topics, such as hospital waiting times in the United Kingdom
- Queries that are likely to result in a high number of hits and so push filters and facets to their limit
- Queries that are listed as having zero or very few hits, to make sure there are justifiable reasons for the low level of user satisfaction that is likely to be a consequence

Ideally, the searches should be congruent with the personas used on the website, which is the principal reason for ensuring that the personas have enough search-related content to assess the outcomes.

However, it is not just about the quality of the search results, but about presenting a coherent web branding for the business that reflects an appropriate level of professionalism and/or a concern about ensuring that site visitors can find the information they need by either browsing the site or by searching the site. The UK website for fashion retailer Burberry is just one example of a site that does not carefully consider customer expectations. If you search for *shoes* on the site, you get 370 images of shoes with a mouseover giving the price but no way of categorizing them by evening wear, leisure, or sportswear, for example. If you walked into a Burberry shop and asked for the shoe department, the first question you would be asked (hopefully!) is what type of shoe you were looking to buy.

## Search Support

For smaller websites, it would be difficult to make a case for a full-time search manager, but at a minimum, there should be a section in the responsibilities of the web team for monitoring search performance, making recommendations for enhancing the search application, and managing the actions needed to deliver the recommendations. Web teams, like any enterprise team, will be short of resources, but the business case is simple. If there is a case to be made for offering website search, then with this case comes the acceptance that managing the search experience is a responsibility for the web team, and not the IT team. If that resource cannot be supported, then the search box should be removed so that there is no expectation from site visitors. Putting this as the option when the business case is being made will usually focus minds on the value of a search box.

The requirement for support by a team with specialist skills is why bringing together these skills in a Search Center of Excellence is a smart approach. Any enterprise-level search application, such as the corporate intranet, will without doubt require at least one dedicated search specialist, for whom looking after the website search would be interesting without the need for a substantial amount of time and effort.

## Fit for Purpose

In general, the quality of website search is poor, even from a company like Rolls-Royce which takes considerable care over corporate communications that reflect corporate values of quality and innovation. It is sadly much easier to find examples of poor search performance than good search performance.

The following are key questions that need to be addressed by all organizations:

- Who are the most important groups of visitors to our website and why?
- What are the typical journeys through the site for these visitors, and therefore the balance between the use of the site architecture and the search functionality?
- Do the website statistics for both clicks and searches confirm that we have achieved the optimal balance between architecture and search?
- How do we compare with our competitors or related sites?
- To what extent are site visitors using Google or Bing to search the site, and why?
- In the case of search, are we providing a good search experience, especially with regard to the information provided about each search result?
- If the experience is not at an acceptable level, is this a result of poor content quality, poor technology, or inadequate staffing levels?
- Do we have the skills and experience to maintain site search at an appropriate level of performance and user satisfaction?

- Do we need a search strategy that also takes into account internal enterprise search applications to ensure that the skills and experience in search can be used across all our search applications?

## Summary

Website search is often not included in an enterprise search strategy. It is important that this is not the case, as there is always a good business case to be made for sharing skills and experience between the people with responsibility for the search applications. Often search is given a low priority on a website, invariably a result of designers focusing on browsing (which can be replicated in card sorting and wireframes) and not on search. There is no better way of assessing web search performance than conducting searches on some of the most popular search terms and considering whether the search experience creates a positive view of the website and the organization.

## Further Reading

Shahriyar Amini, Vidya Setlur, Zhengxin Xi, Eiji Hayashi, and Jason I. Hong, "Investigating Collaborative Mobile Search Behaviors," *MobileHCI '13 Proceedings of the 15th International Conference on Human–Computer Interaction with Mobile Devices and Services*, Munich, Germany, August 27–30, 2013.

Jamie Appleseed and Christian Holst, "E-Commerce Search Usability", Baymard Institute, 2014.

Michael Bendersky, W. Bruce Croft, and Yanlei Diao, "Quality-Biased Ranking of Web Documents," *WSDM '11 Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Hong Kong, China, February 9-12, 2011.

George D. Montañez, Ryen W. White, and Xiao Huang, "Cross-Device Search," *CIKM '14 Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, China, November 3-7, 2014.

Peter Morville, Louis Rosenfeld, and Jorge Arango, *Information Architecture: For the Web and Beyond, Fourth Edition* (Sebastopol, CA: O'Reilly, 2015). (Chapter 9 is on website search.)

Theresa Neil, *Mobile Design Pattern Gallery, Second Edition* (Sebastopol, CA: O'Reilly, 2014).

Greg Nudelman, "Design Patterns for Mobile Faceted Search: Part II", May 3, 2010.

Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang, “Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance,” *WWW ’13 Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, May 13–17, 2013.

# CHAPTER 17
# eDiscovery

Even in organizations that pride themselves on operational excellence, things can go wrong. In 2013, there was a public outcry in the United Kingdom when it was discovered that meat products that were ostensibly beef had been adulterated with horsemeat. All the major supermarket chains found themselves under immense public and governmental pressure to explain how this had happened and to remove the adulterated products from sale. Tracking back through the electronic document trails was an immense task and yet essential to being able to reassure customers that this was a one-off occurrence and that it could not happen again. Motor manufacturers have faced similar problems with tracking down the source of parts that do not meet design standards. The scale of these document trails is immense. In November 2014, a court case in London heard that Rebekah Brooks, the former editor of the *News of the World*, had ordered the deletion of over three million emails sent during the period of her editorship.

When things do go wrong, a court case may be the outcome. At this point, lawyers for both sides of the case will want to find out exactly who knew about the issue, when they knew about it, and what decisions were made. However, the requirement to maintain a complete inventory of documents relevant to a particular business issue might not necessarily stem solely from legal concerns. Internal compliance audits may also require the identification of documents related to regulations on insider trading, money laundering, and export restrictions on certain (usually defense) technologies. eDiscovery has a much higher visibility in the United States, as a result of pretrial discovery under the Federal Rules of Civil Procedure. There are broadly similar requirements in for courts in England and Wales.

There are two reasons for including this short chapter on eDiscovery in a book on enterprise search. First, from a corporate search strategy perspective, there should be visibility on the investment being made for eDiscovery and whether there is an opportunity for cross-exploitation of the skills and expertise of the eDiscovery and enterprise search teams.

The second is that the risks in a court case of failing to find all relevant documents could be damaging in terms of both litigation costs and reputation. As a result, a considerable amount of research into optimizing the eDiscovery process is being undertaken. Although this is of particular interest to law firms and large organizations that may be subject to e-disclosure requirements, the techniques delivered could have wider implications. An example might be in patent searching where there is an equally important requirement to identify all relevant patents in order to confirm novelty or to challenge a patent.

There are many aspects and definitions around the use of the term *eDiscovery*. It is used both for the process of locating documents and managing them through the litigation process, and also for the technology being used to support the process. At a base level, the technology is the same as that used in other enterprise search applications, but vendors of eDiscovery technology provide highly specialized applications for legal use, and very few vendors also provide general-purpose enterprise search applications. These eDiscovery applications are usually the responsibility of a corporate legal department or a law firm undertaking the process of discovery on behalf of a client.

There is a broadly accepted model of the eDiscovery process (Figure 17-1) that has been developed by EDRM, a US trade association. EDRM members are mainly service and software providers and major law firms involved with eDiscovery and information governance.

Ensuring that processes are in place to mitigate risk and expense should eDiscovery become an issue, from initial creation of electronically stored information (ESI) through its final disposition.

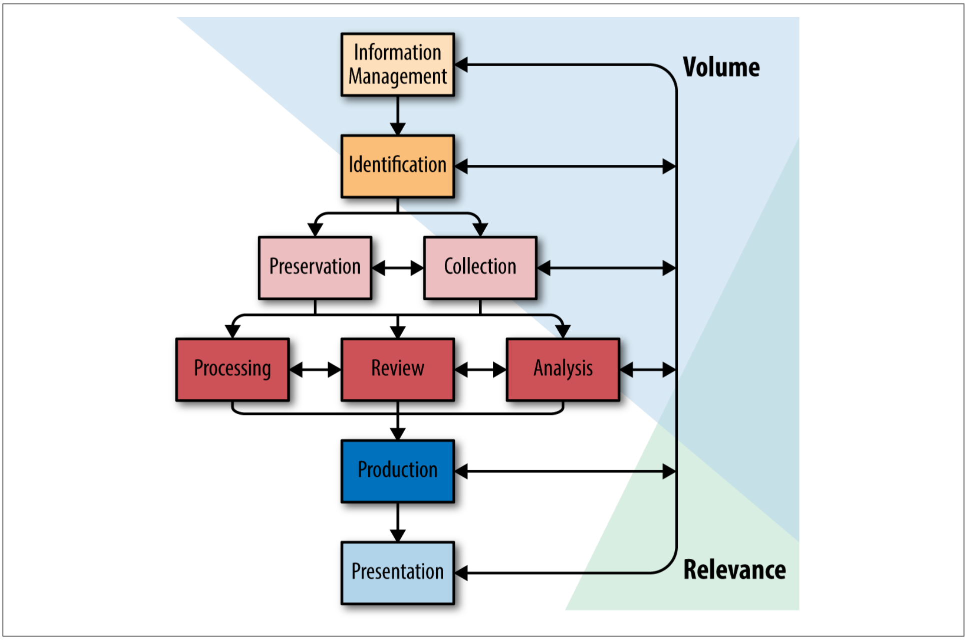


*Figure 17-1. Schematic of the Electronic Discovery Reference Model*

The following notes give a brief scope on each element of the model:

*Identification*
: Locating potential sources of ESI and determining its scope, breadth, and depth.

*Preservation*
: Ensuring that ESI is protected against inappropriate alteration or destruction.

*Collection*
: Gathering ESI for further use in the eDiscovery process (processing, review, etc.).

*Processing*
: Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review and analysis.

*Review*
: Evaluating ESI for relevance and privilege.

*Analysis*
: Evaluating ESI for content and context, including key patterns, topics, people, and discussion.

*Production*
Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms.

*Presentation*
Displaying ESI before audiences (at depositions, hearings, trials, etc.).

## Email Discovery

One of the major differences between most enterprise search applications and eDiscovery applications is that most organizations do not provide users with the ability to search emails despite the wealth of information and knowledge that they contain. There is a tendency for employees not to realize that email is a corporate application and that they are using it for the purposes of the business of the organization. However, emails often contain comments on people or the company that the sender of the email hopes to keep confidential with the person they have sent it to. In the process of writing this chapter in late 2014, the news was full of the hacking of the computer systems and email servers of Sony, which revealed some less than polite comments about a film star.

A survey from Mimecast in 2012 indicated that:

- Only 25% of emails are considered essential for work purposes, with an additional 14% of critical importance
- 13% of work email is personal, not related to work at all
- 40% of work email is either functional or of low-level importance
- On average, 63% of email is internal, employee-to-employee communication

In late 2014, IBM announced IBM Verse as an application to integrate the many ways employees connect each day—via email, meetings, calendars, file sharing, instant messaging, social updates, video chats, and more—through a single collaboration environment. IBM Verse also offers faceted search, based on the Apache Solr platform. In the announcement, IBM reported that industry analysts estimate that 108 billion work emails are sent daily, requiring employees to check their inboxes an average of 36 times an hour.

Searching emails is not as simple as it might seem. Most emails contain attachments, and these need to be indexed and associated with a specific email. The titles of emails are invariably a poor indication of the subject, or often subjects, within an email, and an email discussion might well be extended over several weeks and a range of addressees. These are just a few of the reasons why organizations tend not to try to index and search emails, but developments such as IBM Verse may start to change this situation.

However, there are software applications that can identify groups of messages involving similar conversations or discussions, listed in chronological order and with duplicate content removed. The groupings can be based on metadata (the sender, recipients, date) or on the text in the body of the email.

In an eDiscovery context, it is not only important to be able to identify emails which need to be examined further in the context of an investigation, but to be able to place a hold on the email so that it cannot be deleted. This places a focus on the email archiving policies of the organization. In a survey conducted by AIIM in 2014, the outcomes of the questions about email hold was that 35% of organizations in the survey admitted that their email retention policies and practice are insufficient to ensure reliable discovery and hold.

## Predictive Coding

In principle, a lawyer needs to review each document found by an eDiscovery application to determine whether it is relevant to the case. It is not uncommon for the number of documents to be well in excess of a million items. This is time consuming and (given the business of law) expensive. Predictive coding involves automatically assigning a rating (or proximity score) to each document based on concepts and terms found in documents that have already been deemed by the legal team to be relevant to the matter. There are a number of different approaches to assigning the term codes:

- Lawyers select documents from an initial set of highly relevant documents for the eDiscovery application to analyze. Sometimes a set of irrelevant documents are used as a control set.
- Lawyers use a set of randomly selected documents from the review set.
- An initial keyword search is carried out to provide a list of relevant documents.

The predictive coding software then algorithmically creates a filter that can then be used to screen other documents, allocate a score that reflects the probability that they are relevant to the matter, and present the documents found to the lawyer for an initial review. Based on the review, the filter may be modified and either used on the initial set or on a new set of documents.

Other important elements of eDiscovery technology include the requirement to identify duplicates or near duplicates—for example PDF conversions of Word documents. This can reduce the number of documents that need to be reviewed. The use of semantic text analytics technologies also enables clusters of related documents to be identified.

The use of these technologies is still relatively novel, but as the scale of organizational repositories grows (and potentially extends into Big Data collections), there will be an increasing need to reduce as far as is possible the requirement for the individual review of documents. However, there are some unresolved issues in the US courts about the level of reliance that can be placed on these technologies. Predictive coding works well with large document sets that are largely text based, but is far less reliable for documents (e.g., an Excel spreadsheet) where the amount of text is small. Data quality is also an issue, as many of the documents may be scanned images.

In solving these and other related problems, the requirements for effective eDiscovery could be of benefit to other search applications. It is of note that in early 2015, Microsoft acquired Equivio, arguably one of the leaders in eDiscovery applications.

## The Business of eDiscovery

Of the 22 eDiscovery vendors listed in the 2014 Gartner Magic Quadrant for this market sector, only HP Autonomy and Recommind also offer enterprise search applications. All the other vendors specialize in eDiscovery applications. The main market for these vendors is the United States, but the market in other countries is now starting to develop. For good reason, the department responsible for specifying eDiscovery applications will be a corporate legal department, which in the past may have had little exposure to enterprise search applications.

## Summary

eDiscovery makes use of the fundamental technologies of information retrieval optimized to meet a primary requirement for very high levels of recall. eDiscovery applications are mainly used by large law firms or by the legal departments of large organizations working in areas where there is a requirement to comply with requests for information disclosure.

## Further Reading

Emily Shaw, “Out, Damned [Metadata]!” *Cornell Law School Graduate Student Papers*, Paper 31, 2014.

*The Proceedings of the Annual Conference of the International Association for Artificial Intelligence and Law* include many papers on the use of search and eDiscovery applications in law, patents, and trademarks.

Another important forum for discussions about the technology and implementation of eDiscovery applications is the Sedona Conference, which publishes a wide range of working papers and conferences, including the following:

- *The Sedona Conference Glossary: E-Discovery and Digital Information Management, Fourth Edition* (2014).

- *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* (2013).

## CHAPTER 18

# Text and Content Analytics

The reason for including this chapter on text analytics is because there is a high degree of commonality in the underlying technologies of search and text analytics, and as a result, any enterprise search strategy should take into account the current and potential adoption by the organization of text analytics solutions.

Over the last few years, there has been a substantial amount of hype around Big Data, with most of the major players in the IT industry promoting the view that a significant investment in Big Data is all that an organization needs to make in order to achieve its business objectives. There is no doubt that managing Big Data can make a substantial contribution, but gradually it seems that a balance is emerging around the extent to which Big Data applications need to be complemented by search and text analytics technologies.

The core reason for this is that organizations have a mix of both structured and unstructured information. A survey undertaken by Unisphere Research in 2014 with sponsorship from IBM shows that 25% of respondents indicated that structured data represented no more than 50% of the data under management. Even where the majority of the data under management is structured data, as may be the case in financial services, retail services, and telecommunication services, the unstructured data may provide very important information that enables the data to be placed in a business context. As a result, the value might be substantially greater to the organization than might appear to be the case in terms of volume of data.

## Text Analytics Applications

Text analytics (often referred to as content analytics) examines relationships between topics in a document, often presenting these relationships in a visual display that can be managed interactively. Text analytics dates back to the late 1990s when it was more commonly referred to as text mining. The use of the terms *text analytics* and *content analytics* can be confusing when *search analytics* refers to the use of search logs to track search use and performance. As with search, the technology has quite a history, but has only come into prominence as data and text volumes have increased so rapidly and computing power enables real-time analysis and visualization.

Text analytics uses most of the core elements of a search application, such as information retrieval, computational linguistics, natural language processing, and entity extraction. The technique extracts concepts and patterns from unstructured data to identify relationships and trends that would not be apparent from reading through all the sources being mined, even if one person had the time to do so. In so doing, text mining, in effect, is transforming unstructured data to structured data.

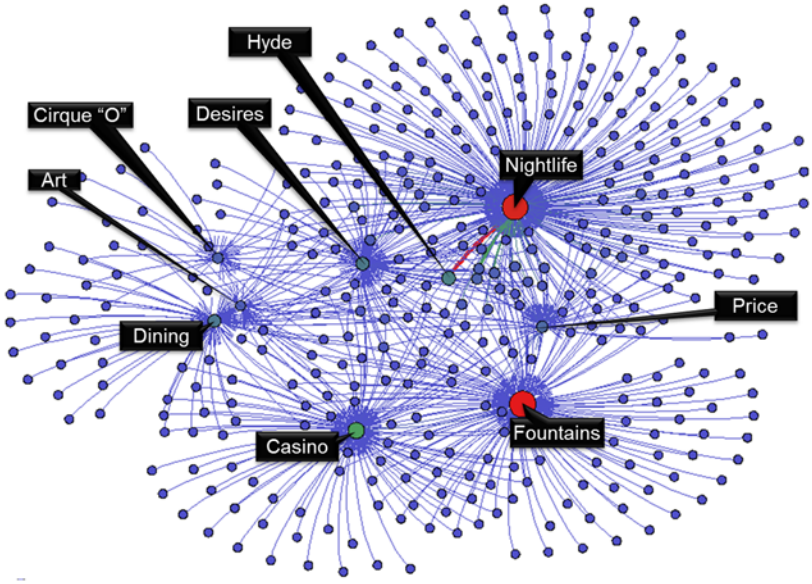


*Figure 18-1. An illustration of the use of visualization in text analytics*

The range of sources that can be mined by text analytics applications is very wide, as illustrated by the chart shown in Figure 18-2 from "Text Analytics 2014: User Perspectives on Solutions and Providers" by Seth Grimes of Alta Plana.

The 216 respondents chose a total of 962 textual information sources, with an average of 5.6 sources per respondent, up from 4.5 in 2011. The big news is not news at all: social sources are by far the most popular, and six of the top eight categories—of the categories chosen by at least 30% of respondents—are social/online (as opposed to in-enterprise) sources.

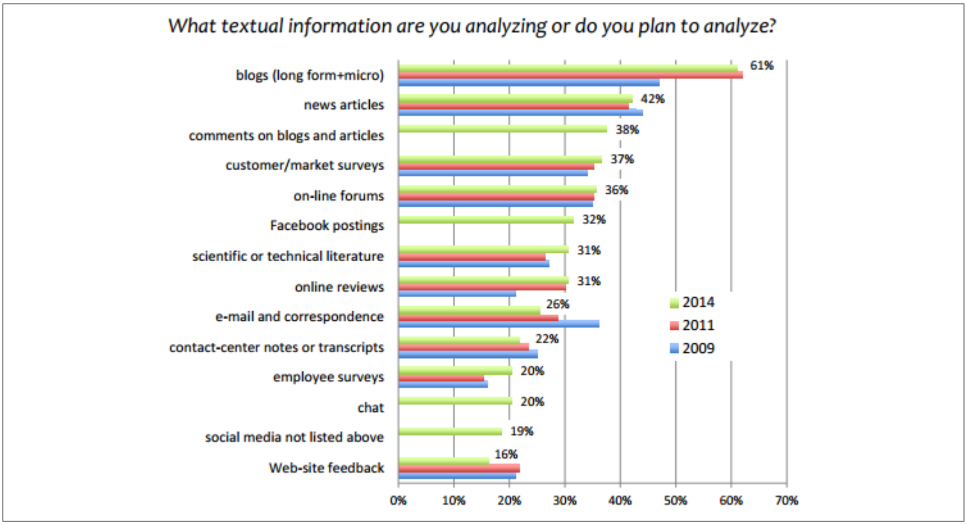


*Figure 18-2. Pattern of change in use of text analytics 2009–2014*

Search works well when the user is able to frame a query based on prior experience and then use that experience to assess the relevance of the information being presented to her in the form of document titles and summaries. With text analytics, there needs to be a starting point for the journey, but then the journey is directed by relationships between topics. The two approaches are complementary.

## Integrating Big Data, Search, and Content Analytics

In 2014, AIIM published the results of a survey on a range of search and discovery topics. One of the questions in the survey asked respondents to comment on the relative importance of enterprise search projects and Big Data/content analytics/visualization projects. The author of the report comments that many aspects of enterprise search have an overlap with content analytics or Big Data. Certainly connectivity to multiple repositories is important, along with context sensitivity within document content. Presentation of the results will be quite different, and when it comes to priorities, there is a philosophical view in that search is of benefit to the everyday jobs of most users, whereas content analytics and Big Data is likely to be a corporate initiative to extract very specific information.

As the survey results in Figure 18-3 demonstrate, only 11% of respondents were prioritizing analytics, and 23% are giving analytics and search equal priority. Half of the respondents feel that search projects should take priority over Big Data projects. Only 5% already have both capabilities.

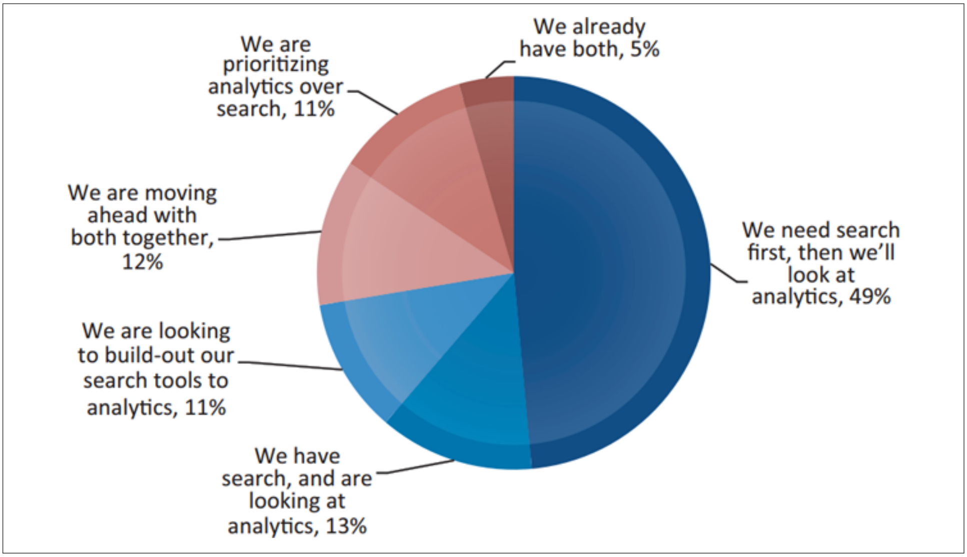


*Figure 18-3. In your organization, how are you prioritizing enterprise search projects and Big Data/content analytics/visualization projects? (N=332)*

In an additional question, 19% said they are moving to a unified Big Data and search strategy, but only 2% said that they are already there. 21% have separate strategies, and 59% have no Big Data strategy at all.

The importance of integrating Big Data with the ability to search unstructured data is emphasized in Figure 18-4 from the report, *Unstructured Data and the New Frontier of Fact-Based Insight*, by the Aberdeen Group in late 2014.

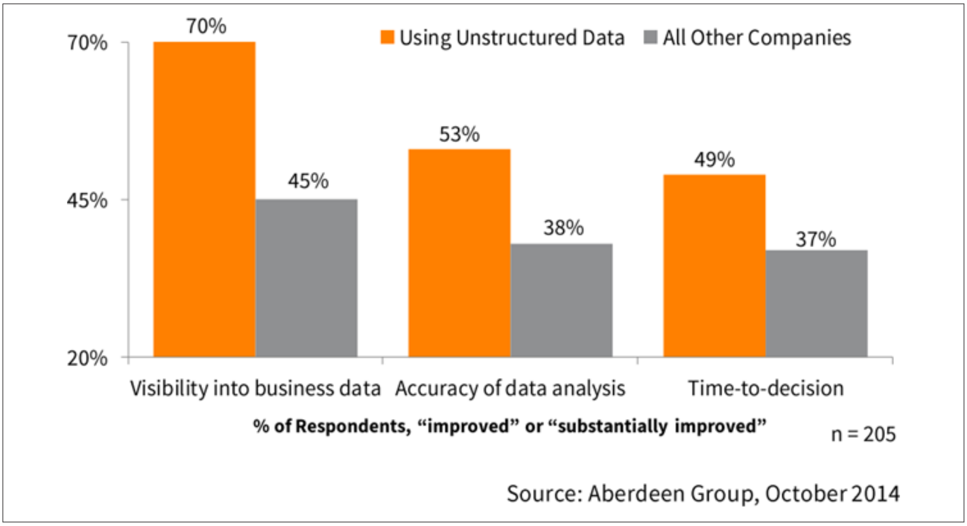


*Figure 18-4. Key aspects of the decision process*

The chart shows that there was a significant improvement in the visibility into business data, the accuracy of data analysis, and a reduction in time to decision.

In 2015, AIIM published another survey on the use of content analytics, and it is interesting to note that in Figure 18-5, using analytics for text searching is the second most popular use.

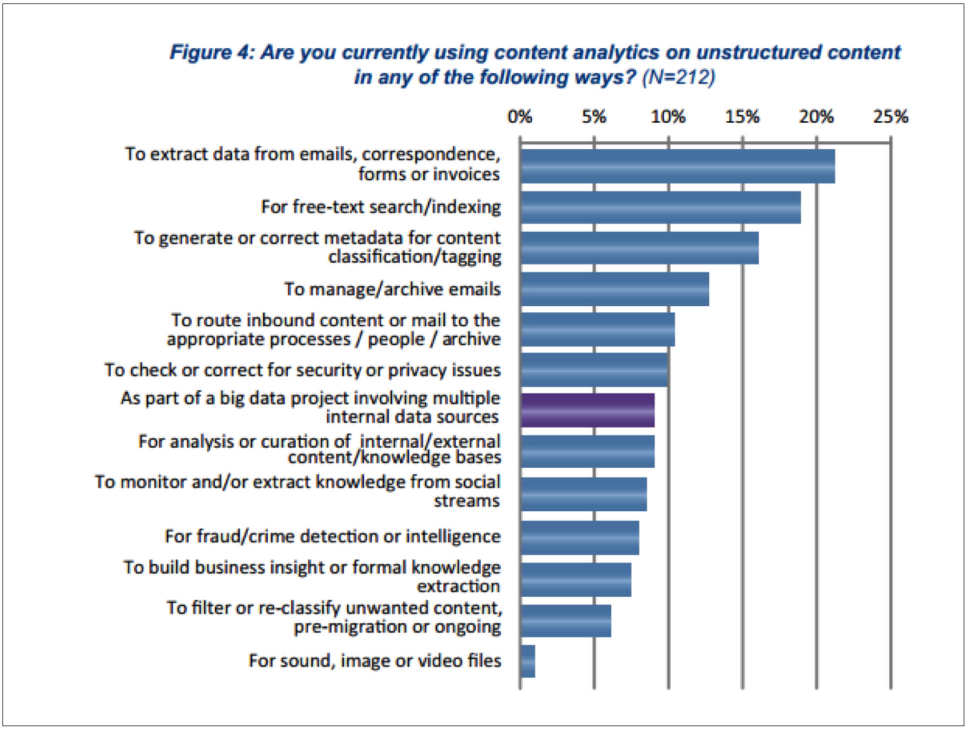


*Figure 18-5. Use of content analytics and unstructured content*

## Sentiment Analysis

There is a lot of attention being paid, at present, to sentiment analysis. An auto manufacturer may well wish to track comments in periodicals, blogs, and other social media about customer reactions to a newly-released model. Ideally, they would be looking for adjectives, such as "excellent," "great drive," and "superb." However, one of the strengths of language, especially English, is the use of sarcasm and irony. "That is the best they can do," is not the same as, "If that is the best they can do then…" They are the same words at the core of the statement, but a very different perspective. "If" is a stop word in most search applications and the addition of these two letters makes all the difference to the statement. However, the efforts to meet the need for dependable sentiment analysis will be of benefit to all types of search applications.

Sentiment analysis is a special case of text analytics, as there is a substantially lower requirement to identify and present relationships between technical terms but a much greater requirement to understand the semantic "sense" of sentence or phrase. The requirement is made even more challenging by the use of irony and sarcasm, especially in English.

It can be helpful to see three broad categories of sentiment analysis:

- Document level, where the objective is to determine the overall sentiment of a document. A company might wish to provide case studies in its marketing literature of successful projects, but what is the definition of "successful"? If it is just profit margin, then the finance department should be able to provide a list. However, a project could have had a very poor financial outcome but the lessons learned could have ensured a new stream of work. Comments about the outcomes might only be in the project summary.
- Sentence level, where a sentence could have a positive, negative, or a neutral (or no) opinion. This might often be the case in a review of a hotel or restaurant, or reviews of new products on industry websites. In the course of a short review, the comments on the food could have been very complementary, but there was a lot of disappointment about the levels of service.
- Entity and aspect level, best illustrated by "This car is not the most comfortable but is perfect for carrying our four children to school." It is difficult enough for a human to work out what the dominant sentiment is, and that would depend on knowing if the car was used on long journeys. In that case the lack of comfort could be a significant negative.

There is a very significant amount of research being conducted into improving the performance of sentiment analysis applications because of its very direct commercial impact. A company could take a product off the shelves in a few days as the result of negative sentiment identified from social media.

# Taking a Strategic Perspective

As with eDiscovery, most of the technology used in text analytics applications is very similar to that used in enterprise search applications. The vendors of text mining solutions do not compete in the enterprise search market. They include Attensity, Clarabridge, IBM Cognos, Lexalytics, Linguamatics, SAP, SAS Teragram, and TEMIX. As with enterprise search vendors, each offers a slightly different pack of modules and algorithms and may specialize in one or more vertical markets.

One important similarity is that none of the technologies works well without a commensurate investment in the staff responsible for managing the applications. There also needs to be ownership of the applications. In the case of eDiscovery, it may well be a general counsel, well grounded in the business requirements of eDiscovery. Customer service and marketing departments would be typical large-scale users of text analytics. Search applications tend to end up with IT, usually by default.

There is undoubtedly going to be a significant increase in the use of text mining as the volumes and potential value of unstructured information escalate over the next few years. In view of the technology overlap with the core elements of enterprise search, there would be a very good case to bring together the support teams for enterprise search and text mining into a common Search Center of Excellence, as set out in Chapter 8.

## Summary

Big Data and contents analytics applications are capable of providing a substantial amount of business intelligence. More benefits are likely to be gained from an integrated approach in which the benefits of enterprise search are recognized and the applications are managed under a common strategic plan. As with eDiscovery, the skills to take advantage of all these technologies have a high degree of commonality.

## Further Reading

Ronen Feldman, “Techniques and Applications for Sentiment Analysis, Communications of the ACM,” April 2013, 56(4), 82-89.
Seth Grimes has published a list of nine examples of how semantic search technologies can improve search quality.

Seth Grimes, “Text Mining and Visualization”, June 13, 2010.

KDnuggets, “Text Analysis, Text Mining, and Information Retrieval Software” (a list of text mining software applications).

Bing Liu, *Sentiment Analysis and Opinion Mining* (San Rafael, CA: Morgan and Claypool, 2012).

# CHAPTER 19
# The Next Five Years

The primary reason for writing a second edition of this book is that so much has changed since I started to write the first edition in 2011. This is especially true of the search business, with many companies being acquired, others disappearing, and open source search being accepted as an enterprise solution. In 2011, we knew very little about how enterprise search was being viewed and used by organizations, but with the help of Findwise, and more recently AIIM, we now have a fairly good indication of the state of the enterprise search environment.

It is still in the early stages of development despite the fact that the core technologies date back to the 1960s. However, as a result of the global financial crisis in 2007 and 2008, many organizations embarked on major redundancy programs in an effort to reduce costs. The place they started was at corporate headquarters with staff in administrative functions without appreciating the role these employees played in being the corporate memory of the organization. They knew where information on past and current projects was stored, and they had excellent networks to tap into the knowledge of the organization. At the same time, the amount of information being generated and collected by these organizations was increasing rapidly. Storage costs were low, so no one worried about the vast collections of multiple versions of documents that were being established. Just press the Save button and let someone else have the problem of finding the information.

The "solution" to business performance problems emerged in 2010 and 2011. It was referred to as Big Data, and it was going to transform the performance of every organization. There are excellent case studies of where this has been the case, but equally there are many organizations that are still struggling to get a return on the investment. The main reasons for this is that structured data only represents part of the data and information stored by the organization, and that the respective volumes of structured and unstructured data bear no relationship to their respective values. The optimal strategy is to invest in proportion to the value, but this is very difficult to determine in any mathematical way. There are as yet no accepted (by accounting authorities) ways to calculate the financial value of data and information, and so it is also not possible to calculate a mathematical return on the investment.

The focus is now shifting toward looking at information in risk management terms, and recognizing the importance of information governance. In the words of the Information Governance Initiative, information governance is the activities and technologies that organizations employ to maximize the value of their information while minimizing associated risks and costs. It is important to note that the word "activity" comes before the word "technology." The value of knowledge management has been appreciated for over two decades and data management moved center stage with Big Data. Now information management is arguably where the action is, and it is in supporting effective information management that search has a very important role to play. In my view, we are entering a golden age for search, and this chapter outlines some of the directions that search needs to take over the years leading to 2020 if the emerging requirements of digital workplaces are going to be met. Of course, the danger with committing this forecast to a book is that readers in 2020 may have cause to doubt my sanity, but hopefully you will find something in this chapter to excite you about the future possibilities of search. The topics are not in any specific order of priority or imminent availability.

## Collaborative Information Seeking

Nowadays, everyone works in many different teams, often taking a different role in each team. All our IT systems are designed to support team working, to the extent that typically the investment in an enterprise-wide SharePoint implementation is justified on the basis of helping people to work together more effectively. These IT systems increasingly support the simultaneous use of desktop PCs, laptops, tablets, and smartphones by members of what are often virtual teams.

However, search remains a solitary exercise: one person sitting at a desktop PC and conducting a search and then emailing the URLs or the documents retrieved to other members of the team. In the 1970s and 1980s, there was a very rapid growth in the use of online search services to databases of scientific, engineering, legal, and business information from companies such as Lockheed Dialog and LexisNexis. Using these systems was not easy, and it was very common for searches to be carried out as a partnership between an information professional and search requester around the same terminal. This approach was overtaken when the World Wide Web arrived and everyone became an expert searcher almost overnight.

From a research perspective, work on information retrieval and work on collaboration and human–computer interaction (HCI) has traditionally been undertaken in different departments. Now there is a recognition of the need for joint research projects in an area that is referred to either as collaborative information retrieval (CIR) or collaborative information seeking (CIS). This does not make it easy to track down research papers! There are benefits in using the term CIS as this moves the focus away from just considering collaborative search applications, instead putting search as just one way of seeking information within the context of an overall information management strategy.

It is important to note that collaborative information seeking is not just about search but about how information is discovered and shared between teams, especially when they are separated by time, distance, and fluency in a common language. A good place to start in understanding the potential benefits and challenges of CIS is Coagmento. Set up by Professor Chirag Shah, Rutgers University, the Coagmento project is in the forefront of research in this area.

One of the major challenges in providing collaborative information seeking is that in a team not all the members may have the same access privileges. That may not affect the way in which the team works together, but could cause some problems when one member of the team seems to have access to documents that are not available to another member of the team, and this is immediately obvious from the search results list that is on the tablet that is being passed around. Members of the team, and indeed the team leader, may not be aware of these security problems.

Nevertheless, given the high level of teamworking in organizations, especially virtual teamworking, the opportunity for vendors to support these teams with collaborative search applications is a very significant one.

## Further Reading for "Collaborative Information Seeking"

Robert Capra, Javier Velasco-Martin, and Beth Sams, "Collaborative Information Seeking by the Numbers," *CIR '11 Proceedings of the 3rd International Workshop on Collaborative Information Retrieval*, Glasgow, UK, October 24–28, 2011.

Jonathan Foster (editor), "Collaborative Information Behavior: User Engagement and Communication Sharing," IGI Global, Hershey, PA, USA, 2011.

Meredith Ringel Morris, "Collaborative Search Revisited," *CSCW '13 Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, San Antonio, TX, February 23-27, 2013.

Chirag Shah, *Collaborative Information Seeking* (New York: Springer, 2012).

Chirag Shah, "Evaluating Collaborative Information Seeking: Synthesis, Suggestions, and Structure," *Journal of Information Science*, 40:4 (2014): 460–475.

## Mobile Search

As I started to write this chapter in 2015, Apple announced the largest recorded quarterly profits in the history of commerce, a result of shipping millions of iPhone 6 smartphones. No one would disagree that mobile is now the dominant way of accessing the World Wide Web, but the situation inside an organization is somewhat different. Much is made of the benefits of responsive design in ensuring that the investment in a website can be recovered from making it universally useful to PCs, tablets, smartphones, and, no doubt, in due course to watches as well. Where responsive design is less effective is in managing search pages on websites because of the widespread use of facets and filters. In addition, the primary method of entering a query is to type it in from a virtual keyboard which obscures a significant area of the screen.

In a consumer search situation, the mobile phone can make some assumptions about the search scope based on immediately previous searches, location, and possibly information from a calendar application. The environmental information is very important, and distinguishes mobile search from desktop search in the same way that web search (based around page links and views) differs from enterprise search.

Delivering enterprise mobile search involves:

- Ensuring that it is device independent when employees are using their own devices
- Making the search box/query interface easy to use in mobile circumstances
- Providing an adequate level of security management
- Displaying results in a way that they can be opened, assessed, and closed down
- Presenting ways of reducing large result sets through the mobile equivalent of filters and facets
- Displaying nonstandard (i.e., non-Microsoft Office) documents
- Enabling users to store result sets and documents so that they can be printed out later on

Over the last few years, a number of search vendors have promised enterprise mobile search applications, but currently only Coveo is making a significant push in this direction. However, at the time of writing, only Apple and Android devices are supported.

Another important issue is managing the potential transition between a user undertaking a search on a PC in her office, wanting to show some of the results to colleagues using a tablet, and then emailing documents discovered through the search to other colleagues. The transition from PC to tablet to mobile has not been studied in an enterprise context, but in a consumer context, research from Microsoft suggests that there could be differences between a journey that starts with a PC and a journey that starts with a search on a smartphone.

One technology that will play an increasingly important role in mobile search is voice recognition technology. The advances in this technology have been dramatic over the last decade, and issues about strong dialects and complex query terms are well on their way to being solved, especially as 4G networks provide the bandwidth that unleashes the computer power needed to transform voice to text and text to voice. These technologies are patented and supported by major IT companies, notably Microsoft, Google, and Apple. Android is only nominally open source—it is as open source as Google wishes it to be. The point is that these technologies may not be widely available to the open source community to build voice input query systems.

Much of the development work on mobile search has been for public rather than enterprise applications. Nevertheless, the work of Greg Nudelman on design patterns for mobile search is a very good starting point even though the examples are mainly from ecommerce applications.

Expect there to be a substantial increase in mobile search solutions over the next couple of years. The user demand will eventually stimulate search vendors to offer solutions that will make use of the voice applications and swipe interface potential of mobile devices.

## Further Reading for "Mobile Search"

Shahriyar Amini, Vidya Setlur, Zhengxin Xi, Eiji Hayashi, and Jason I. Hong, "Investigating Collaborative Mobile Search Behaviors," *MobileHCI '13 Proceedings of the 15th International Conference on Human–Computer Interaction with Mobile Devices and Services*, Munich, Germany, August 27–30, 2013.

George D. Montañez, Ryen W. White, and Xiao Huang, "Cross-Device Search," *CIKM '14 Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, China, November 3-7, 2014.

Theresa Neil, *Mobile Design Pattern Gallery, Second Edition* (Sebastopol, CA: O'Reilly, 2014).

Greg Nudelman, "Design Patterns for Mobile Faceted Search: Part II" May 3, 2010.

Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang, "Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance," *WWW '13 Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, May 13–17, 2013.

## Search + eDiscovery + Text Analytics

The underlying technologies of search, eDiscovery, and text analytics are fundamentally the same, certainly to an 80% level. The devil is in the detail of the remaining 20%, and so far, the companies providing solutions in each of these areas have not ventured into the other two. Recommind and HP Autonomy are the only exceptions. In all three sectors, the vendors are relatively small businesses and do not have the marketing, sales, and support investment to move into another sector. However, customers will increasingly look for a total search platform, and it would seem likely that there will be consolidation in the market. Early in 2015, Microsoft acquired Equivio, and that could signal the start of the consolidation process.

This consolidation does not mean that keyword search is dead. Electric cars offer very considerable benefits over conventional cars, but at present the charging infrastructure is not there to support the widespread adoption of electric cars, the technology is certainly going to change, servicing requires considerable expertise and different diagnostic techniques, and it is not clear what the total cost of ownership will be over an extended period. This is the situation in search. The majority of organizations are still learning how to get the best of what some might regard as low-technology solutions. Moreover, if these organizations read this book, they will realize that the immediate solution is not technology, but investing in the people who know how best to take advantage of the technology. It's more about engine tuning than installing an electric motor.

## The Cost Model for Open Source Search

The core justifications by the open source community to adopt open source search solutions were cost savings and choice. Certainly the license cost savings can be considerable, but good developers, either internal or external, will be a significant investment and the search support team requirements are the same. As yet, there are no good Total Cost of Ownership models available.

A bigger problem is the level of venture capital investment in LucidWorks and Elasticsearch. In the case of the latter, the disappearance of Sense, a zero-cost Chrome plug-in for Elasticsearch that supported auto-completion, code highlighting, and formatting, and its incorporation into the subscription-based license for Marvel is an indication of the future. Venture capital funds need to obtain a return on their investment. They know that not every investment will work out, but that does not mean that they will allow companies they have invested in to work on free-to-use products. It is always interesting to take note of the change of management when an investment fund comes on board.

Certainly there is a vast choice of developers who can use Apache Lucene and Solr to create very effective search applications, but in effect the choice of core code is either

Lucene/Solr or Lucene/Elasticsearch. There are other software libraries but they have a very limited or specialized use.

These are observations, and do not imply that I see no future in open source search. Nevertheless, the dominant vendors in enterprise search are Microsoft and Google, with perhaps IBM as an option for companies with an enterprise commitment to the company. Persuading enterprise IT managers to switch out their SharePoint 2013/2016/cloud applications, or a Google appliance, for an open source search option is going to take an enormous amount of marketing and pre-sales investment from vendors who are still marketing to the developer community. In my view, this is not going to happen, even though there will be substantial adoption of open source search in every other sector of the search market. I think that it is more likely that open source search will be used for specific applications where there is a need to develop in stages as requirements and solutions emerge.

## Advances in Information Retrieval

The basis of all information retrieval software is a combination of computational linguistics (to develop a semantic model of a piece of content) and the mathematics of probability (to determine the extent to which the content satisfies the query). Research into new approaches to information retrieval is growing in scope, as the scale of the annual SIGIR (the Special Interest Group for Information Retrieval of the Association for Computing Machinery) conference is a testimony to. IR research is also featured in many other conferences around the world. What is interesting about IR is that in general it is not technology-specific, and involves mainly brain power. So the costs of IR research are not as high as, for example, genomic research.

All the major IT companies are currently involved in various aspects of IR research, and much can be gleaned from the published papers about some the directions that these companies are taking. They can see the potential benefits of solving the problems faced by their customers with novel approaches, mainly at an algorithmic level, to improving search performance. It is also entirely possible that over the next few years new companies will be set up to take commercial advantage of this research, possibly with the view of selling out to a major IT company in due course.

The challenge here is finding a mechanism to bridge between the research interests of academic groups and creating business ventures with a realistic business plan. Although these are quite common in many areas of research (chemistry and biochemistry in particular), there is little experience of doing so in the information retrieval sector. If these transfers can be well supported, then the benefits to enterprise search could be substantial.

A good place to start to get a sense of the opportunities and challenges is to read the report of a conference that took place in Australia in 2012. Entitled “Frontiers, Challenges, and Opportunities for Information Retrieval," it sets out the views of most of the leading research teams in information retrieval.

To keep aware of developments in information retrieval and search, the best approach is to gain access to the Digital Library of the Association for Computing Machinery, the US professional organization for computer science. Members of the British Computer Society have access rights and that may be the case for other national computer science professional societies. This Library can be searched online and lists not only ACM publications and conference proceedings but also those from a wide range of commercial and learned society publishers. The ACM also sponsors the Annual Conference of the Special Interest Group for Information Retrieval.

## Further Reading for "Advances in Information Retrieval"

James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson (Editors), "Frontiers, Challenges, and Opportunities for Information Retrieval", Report from SWIRL 2012, The Second Strategic Workshop on Information Retrieval in Lorne, February 2012.

Microsoft Research.

# Cross-Lingual Search

Despite the extensive use of English in organizations, even multinational organizations, there are a great many other languages that are used in an individual country or even in many countries (e.g., Spanish). The requirement to search across multiple languages remains an important one, and again there is a significant amount of research taking place.

There are three approaches:

- The query is translated into the language of the document
- The document is translated into the language of the query
- Both the query and the document are translated into a third, common, language

This area of work has been transformed by the speed of development of machine translation applications, notably from Google and Microsoft. Microsoft Skype already offers real-time translation between English and Spanish, and no doubt other languages will be offered in due course.

At present, the focus is on approaches that translate queries into the same language as the target document collection. Comparatively speaking, there has been little interest in working to translate the document collection into the same language as the query. It may seem counterintuitive, but it is easier to translate a full document than a query. This is because documents contain contextual information that is not present in queries. In addition, it is very important that each term in a very short query is correctly translated. An error in the translation of even a number of individual words in a much longer section of text will have less impact on the quality of retrieval. Until recently, these advantages have been offset by the costs (in machine processing time) of translating the entire collection of documents. We may now be close to the point where these costs are such that the improvement in retrieval quality can be offset. In the initial stages, this may only be in a few languages, the most important of which (and the most difficult) would be Chinese.

## Further Reading for "Cross-Lingual Search"

Carol Peters, Martin Braschler, and Paul Clough, *Multilingual Information Retrieval: From Research to Practice* (New York: Springer, 2012).

# Multistage User Interfaces

It is generally accepted that search is a multistage task, as has been discussed already in Chapter 10. One of the early models for search tasks was proposed by Carole Kuhlthal and is set out as follows:

*Phase 1: Initiation*
: Becoming aware of a lack of knowledge or understanding, often causing uncertainty

*Phase 2: Selection*
: Identifying and selecting a general area, topic, or problem; sense of optimism replaces uncertainty

*Phase 3: Exploration*
: Exploring and seeking information on the general topic; inconsistent info can cause uncertainty

*Phase 4: Formulation*
: Focused perspective is formed; uncertainty is reducing, while confidence increases

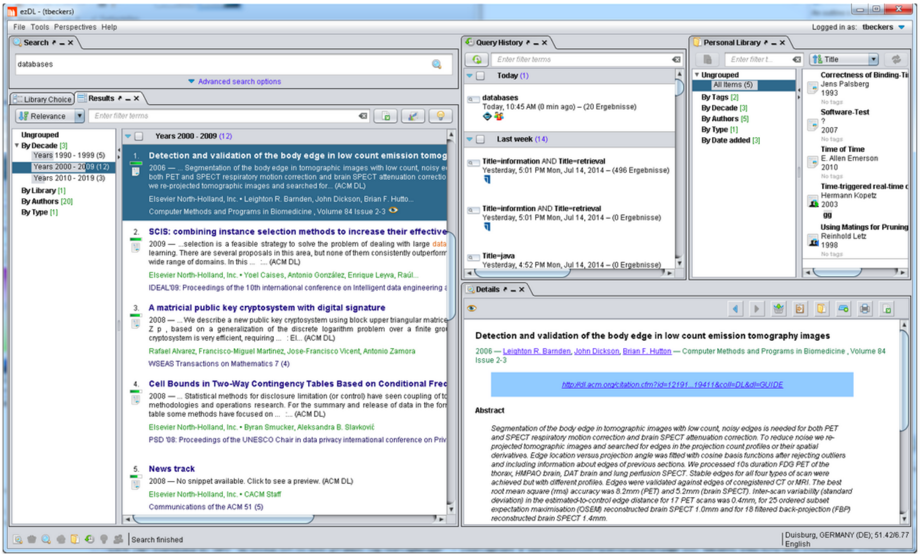
*Phase 5: Collection*
: Gathering pertinent information to focused topic; less uncertainty, more interest/involvement

*Phase 6: Presentation*
: Reporting and using results of the completed search

At present, all these stages are managed by a single interface, but this is almost certainly a compromise resulting from a lack of research into the benefits of different user interfaces and a concern about increasing the complexity of the search application without any immediate benefit. As already discussed, users are now accustomed to different screen layouts on desktop, tablet, and smartphone devices and may be less averse to coping with different layouts on a desktop. Max Wilson has identified four elements of a search user interface (Max Wilson, *Search User Interface Design* (San Rafael, CA: Morgan and Claypool, 2012)):

- Input features which help the user to express what they are looking for
- Control features which help users to modify or restrict their input
- Informational features which provide results or information about the results
- Personalizable features which relate specifically to searchers and their previous interactions

The ezDL (easy access for digital libraries) interface has already been developed for use in digital library search management, and the website for ezDL illustrates a number of different screens that can be used in the course of complex searches. Figure 19-1 shows the main screen for ezDL, while Figure 19-2 shows one of the other screens.



databases

Ungrouped
By Decade [3]
Years 1990 - 1999 (5)
Years 2000 - 2009 (12)
Years 2010 - 2019 (3)
By Library [1]
By Authors [20]
By Type [1]

Years 2000 - 2009 (12)

1. **Detection and validation of the body edge in low count emission tomog**
2006 — ... Segmentation of the body edge in tomographic images with low count, noisy e
both PET and SPECT respiratory motion correction and brain SPECT attenuation correctio
we re-projected tomographic images and searched for... (ACM DL)
Elsevier North-Holland, Inc. • Leighton R. Barnden, John Dickson, Brian F. Hutto...
Computer Methods and Programs in Biomedicine , Volume 84 Issue 2-3

2. **SCIS: combining instance selection methods to increase their effective**
2009 — ...selection is a feasible strategy to solve the problem of dealing with large data
learning. There are several proposals in this area, but none of them consistently outperform
wide range of domains. In this ... :... (ACM DL)
Elsevier North-Holland, Inc. • Yoel Caises, Antonio González, Enrique Leyva, Raúl...
IDEAL'09: Proceedings of the 10th international conference on Intelligent data engineering a

3. **A matricial public key cryptosystem with digital signature**
2008 — ... We describe a new public key cryptosystem using block upper triangular matrice
Z p , based on a generalization of the discrete logarithm problem over a finite gro
cryptosystem is very efficient, requiring ... : El... (ACM DL)
Rafael Alvarez, Francisco-Miguel Martinez, Jose-Francisco Vicent, Antonio Zamora
WSEAS Transactions on Mathematics 7 (4)

4. **Cell Bounds in Two-Way Contingency Tables Based on Conditional Frec**
2008 — ... Statistical methods for disclosure limitation (or control) have seen coupling of to
methodologies and operations research. For the summary and release of data in the for
table some methods have focused on ... :... (ACM DL)
Elsevier North-Holland, Inc. • Byran Smucker, Aleksandra B. Slavković
PSD '08: Proceedings of the UNESCO Chair in data privacy international conference on Priv

5. **News track**
2008 — No snippet available. Click to see a preview. (ACM DL)
Elsevier North-Holland, Inc. • CACM Staff
Communications of the ACM 51 (5)

Today (1)

**databases**
Today, 10:45 AM (0 min ago) – (20 Ergebnisse)

Last week (14)

**Title=information** AND **Title=retrieval**
Yesterday, 5:01 PM Mon, Jul 14, 2014 – (496 Ergebnisse)

**Title=informtion** AND **Title=retrieval**
Yesterday, 5:01 PM Mon, Jul 14, 2014 – (0 Ergebnisse)

**Title=java**
Yesterday, 4:52 PM Mon, Jul 14, 2014 – (0 Ergebnisse)

Ungrouped
All Items (5)
By Tags [2]
By Decade [3]
By Authors [5]
By Type [1]
By Date added [3]

**Correctness of Binding-Ti**
Jens Palsberg
1993
No tags

**Software-Test**
?
2007
No tags

**Time of Time**
E. Allen Emerson
2010
No tags

**Time-triggered real-time c**
Hermann Kopetz
2003
gg

**Using Matings for Pruning**
Reinhold Letz
1998
No tags

**Detection and validation of the body edge in low count emission tomography images**

2006 — Leighton R. Barnden, John Dickson, Brian F. Hutton — Computer Methods and Programs in Biomedicine , Volume 84 Issue 2-3

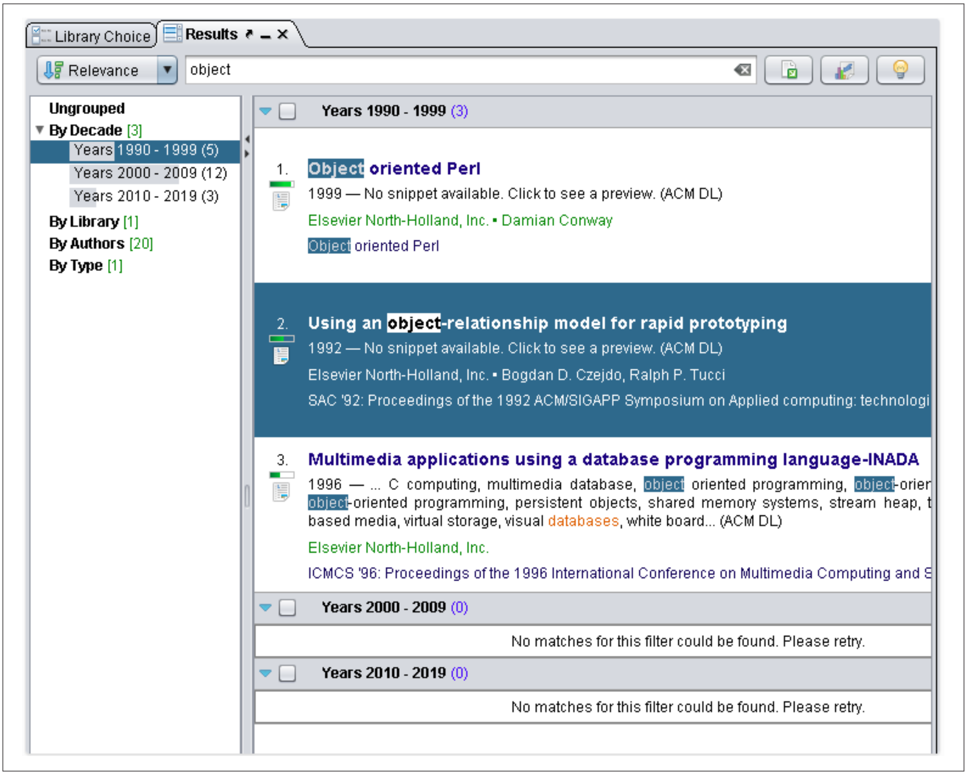
http://dl.acm.org/citation.cfm?id=12191...19411&coll=DL&dl=GUIDE

**Abstract**

*Segmentation of the body edge in tomographic images with low count, noisy edges is needed for both PET and SPECT respiratory motion correction and brain SPECT attenuation correction. To reduce noise we re-projected tomographic images and searched for edges in the projection count profiles or their spatial derivatives. Edge location versus projection angle was fitted with cosine basis functions after rejecting outliers and including information about edges of previous sections. We processed 10s duration FDG PET of the thorax, HMPAO brain, DAT brain and lung perfusion SPECT. Stable edges for all four types of scan were achieved but with different profiles. Edges were validated against edges of coregistered CT or MRI. The best root mean square (rms) accuracy was 8.2mm (PET) and 5.2mm (brain SPECT). Inter-scan variability (standard deviation) in the estimated-to-control edge distance for 17 PET scans was 0.4mm, for 25 ordered subset expectation maximisation (OSEM) reconstructed brain SPECT 1.0mm and for 18 filtered back-projection (FBP) reconstructed brain SPECT 1.4mm.*

Duisburg, GERMANY (DE); 51.42/6.77
English

*Figure 19-1. Main desktop from the ezDL application*



Ungrouped

▼ By Decade [3]

Years 1990 - 1999 (5)

Years 2000 - 2009 (12)

Years 2010 - 2019 (3)

By Library [1]

By Authors [20]

By Type [1]

□ Years 1990 - 1999 (3)

1. Object oriented Perl
1999 — No snippet available. Click to see a preview. (ACM DL)
Elsevier North-Holland, Inc. • Damian Conway
Object oriented Perl

2. Using an object-relationship model for rapid prototyping
1992 — No snippet available. Click to see a preview. (ACM DL)
Elsevier North-Holland, Inc. • Bogdan D. Czejdo, Ralph P. Tucci
SAC '92: Proceedings of the 1992 ACM/SIGAPP Symposium on Applied computing: technologi

3. Multimedia applications using a database programming language-INADA
1996 — ... C computing, multimedia database, object oriented programming, object-orier object-oriented programming, persistent objects, shared memory systems, stream heap, t based media, virtual storage, visual databases, white board... (ACM DL)
Elsevier North-Holland, Inc.
ICMCS '96: Proceedings of the 1996 International Conference on Multimedia Computing and S

□ Years 2000 - 2009 (0)

No matches for this filter could be found. Please retry.

□ Years 2010 - 2019 (0)

No matches for this filter could be found. Please retry.

*Figure 19-2. The result list filtering screen from ezDL*

This application is available for both desktop and Android smartphone versions.

## Further Reading for "Multistage User Interfaces"

*http://www.slideshare.net/TimelessFuture/from-multistage-information-seeking-models-to-multistage-search-systems-web*

*http://humanities.uva.nl/~kamps/publications/2014/huur:from14.pdf*

# Federated Search

In principle, federated search should be the Holy Grail of search, representing the ability to use one search query to search a very wide range of search applications, both internal and external. The major problem is one of providing the user with a ranked list of results that can be relied upon to be representative of all the applications which have been searched. Research is being carried out into how best to fuse the individual results sets together in a federated search which parses a single query across multiple applications, rather than creating a single index.

This is a good place to highlight the TExt Retrieval Conference (TREC), which is cosponsored by the National Institute of Standards and Technology (NIST) and the US Department of Defense. The main purpose of TREC, which began in 1992, is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

The TREC workshop series has the following goals:

- To encourage research in information retrieval based on large test collections;
- To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- To speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- To increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

In 2013, TREC initiated a Federated Search track to reflect the growing interest in solving the problems inherent from searching across multiple applications. It is not just a question of resolving the ranking of results, but also in determining which applications should be searched so that the search process is not unduly prolonged by searching through applications which are very unlikely to contain highly relevant information. Although the Federated Search track was continued in 2014, a lack of funding and of interest has led to it being discontinued.

At present, research has focused on federated search across websites and across digital library collections, and there has only been a limited amount of research into enterprise applications. This will change over the next few years and it is very likely that there will be some significant advances in federated search applications in both commercial and open source search applications.

## Search-Based Applications

Search-based applications vie with enterprise portals as a term without any clear meaning. In essence, a search-based application uses search technology but without a specific query box to support access to data and information. It is all done in the background. The query box is replaced by a set of APIs that provide the interface from a user to the application. A good example would be media monitoring, in which a high-volume stream of news stories is being filtered through a set of profiles.

The example I use when talking about search-based applications is the UK real estate site Right Move. Real estate sites are also a good example of parametric search, with an almost endless set of selection criteria in drop-down lists. What is unusual about Right Move is Draw-a-Search. Real estate sites either offer to list sites within a radius of a specific location or make use of postcodes to define an area. Draw-a-Search enables a buyer to draw a polygon around an area, perhaps avoiding a busy road but taking in an area where there is a good school. The process of drawing the polygon defines the area in a sequence of searches in which the position of the arrow on the screen is converted to latitude and longitude and houses are displayed which have a geolocation that is within the polygon. A very simple interface but a very complex search-based technology solution.

## Further Reading for "Search-Based Applications"

Gregory Grefenstette and Laura Wilber, *Search-Based Applications: At the Confluence of Search and Database Technologies* (San Rafael, CA: Morgan and Claypool, 2011).

# It's All in the Mind

Fred Hoyle (1915–2001) was a distinguished British mathematician and astronomer who also wrote several seminal science fiction books, among them *The Black Cloud* in 1957. The black cloud was an intelligent cloud-like body that had arrived in the solar system. The hero of the book, Chris Kingsley, is able to find a way of communicating with the cloud, which agrees to share with Kingsley its knowledge of the universe and how it works. The outcome is that Kingsley suffers a major neurological failure as he tries to replace the knowledge in his brain with the revised version of science that comes from the cloud, ultimately resulting in his death.

So what does this have to do with enterprise search? As an information scientist, I thought I ought to end this book with some reflections on the cognitive aspects of retrieval. A few days before completing the text, I came across a paper by Professor Reijo Savolainen from the School of Information Sciences at the University of Tampere, one of the largest and most active research groups in the world in the area of information retrieval.

The paper, "Cognitive Barriers to Information Seeking: A Conceptual Analysis," is a very important contribution to the understanding of how we undertake the process of information seeking. The author addresses a challenging issue: the characterization of the impact of cognitive barriers on information seeking. This was undertaken by a very thorough review of the literature, and the bibliography extends to over 50 citations.

From the analysis, six barriers are described:

- Unwillingness to see needs as information needs
- Inability to articulate information needs

- Unawareness of relevant information sources
- Low self-efficacy, where the user feels that it will be difficult to obtain the documents
- Poor search skills
- Inability to deal with information overload

From my own experience working as an enterprise search consultant, none of these barriers would be recognized by what is often a technology-led enterprise search team, but as I talk to users and stakeholders, these issues all emerge.

Searching for information is a very challenging process. A piece of technology must do its best to read our minds and determine what we might regard as relevant to our query. As the scale of information repositories grows, it will become increasingly difficult for users to cope with the six barriers to information seeking outlined by Professor Savolainen unless, as an enterprise, the organization recognizes that we are now in the Information Age—information is business critical, and we need to be training every employee not just about how to search but also about how to manage information at a corporate and personal level. The appropriate use of technology, good quality information and metadata, high-quality user interfaces, and a skilled search support team are just a small element of what is a much wider issue, that of managing information as an asset of the business (which is where I started in Chapter 1).

Good luck. It will be worth the effort.

## Further Reading for "It's All in the Mind"

Nigel Ford, *Introduction to Information Behaviour* (Facet Publishing, 2015).

Daniel Levitin, *The Organized Mind* (Penguin Books, 2015).

Reijo Savolainen, "Cognitive Barriers to Information Seeking: A Conceptual Analysis," *Journal of Information Science*, 41(5), pp 613-623 (2015).

# APPENDIX A
# Search Strategy

This Appendix sets out (in alphabetical order) a checklist of the topics that should be considered for inclusion in a search strategy.

***Accessibility***
Sets out the extent to which the search applications meet the Web Accessibility Initiaitives

***Acquisition***
The extent to which the search applications could be extended in an acquisition or merger situation

***Architecture***
Server and network architecture requirements and server availability

***Best bets***
The user requirements for best bets and how they will be reviewed and revised

***Big Data***
Integration of the search and Big Data strategies, especially around common metadata schemas

***Budget***
License costs, vendor maintenance and support, and staff costs

***Business cases***
Summary of the evidence from user requirements research to support and prioritize specific business cases, including the potential business impact

***Cloud search***
The potential benefits and challenges from implementing cloud (or hybrid) search

***Communications***

The communications strategy and forward communications program for stakeholders and users

***Connectors***

Requirements for connectors and associated support from suppliers

***Content analytics***

The extent to which the organization will benefit from implementing content analytics solutions and the relationship of these solutions to search

***Content quality***

Requirements for content quality and content curation to enhance search performance, ideally placed within an information life cycle framework

***Content scope***

A list of the content being crawled and indexed

***Crawl management***

Optimal crawl schedules to balance user requirements with any architecture/ performance constraints

***Dependencies***

Business or technical dependencies that could impact search performance and search satisfaction

***Development plan***

The opportunities for enhancing the search environment over the following two years, based on user requirements research and business objectives matched against resources

***Disaster recovery***

Disaster recovery plans with Recovery Time Objective (RTO) and Recovery Point Objective (RPO) requirements

***eDiscovery***

If appropriate, the touch points between the eDiscovery strategy and the search strategy, especially regarding the sharing of skills

***Expertise search***

Linking the requirements for expertise search from a knowledge management strategy with the search strategy

***External search***

The requirements for search access to external information resources on e.g. research, competitors, and market opportunities

***Federated search***

Current and potential opportunities and challenges for implementing federated search

***Feedback***

How users will be able to feedback comments and suggestions to the search team

***Governance***

The ownership of the search budget and search strategy, together with roles, responsibilities, and lines of reporting for members of the search team

***Help desk management***

The relationship between the IT Help Desk team and ticket system and the search help desk

***Information management***

A summary of the organization's information management strategy with particular reference to the requirements and objectives for the search strategy

***IT liaison***

Service-level agreements with IT departments for support and development, including the requirement for staff with specific skills to be available

***Key performance indicators***

Definition of a set of periodic key performance indicators that relate to the business cases and business impact requirement

***Language***

Setting out any requirements for indexing and searching in languages other than the nominal corporate language.

***Legal conformance***

Requirements to conform to data privacy, Freedom of Information, and export license controls

***Licenses***

List of licenses by vendor and license renewal date so that the implications of a merger or acquisition of the vendor can be quickly assessed

***Metadata***

A summary of metadata schema, controlled term lists, thesauri, and relevant master data schema

***Metrics***

A summary of the suite of performance, discovery, satisfaction, and impact metrics, together with required benchmark levels

***Migration***

The implications for search as an element in a content migration strategy

***Mobile***

How the search applications will be implemented on mobile devices, together with an assessment of the need for cross-device support

***Open Source***

Sets out the organization's approach to using open source applications

***People search***

The requirements for people search

***Performance***

The technical (network/server) performance benchmarks for crawl, index, query, and result display

***Risk register***

A risk analysis relating both to operational and strategic risks for the search application, and the consequential risks to the organization

***Roadmap***

Release dates for upgrades to search applications, the basis on which they would be implemented, and development roadmaps for other enterprise applications

***Scope***

Confirmation of the repositories to be crawled and indexed in order to meet user and business requirements, the search applications to be included in the strategy, and the search applications that are being excluded

***Search support team***

Operational responsibilities and reporting lines for the search support team, including requirements for training

***Security***

Summary of security requirements covering confidentiality, integrity, and availability in line with ISO 27001

***SharePoint strategy***

As appropriate, sets out the scope of Microsoft SharePoint adoption and development

***Stakeholders***

Confirmation of the stakeholders and other members of the search community, using the RACI model

***User training***

Provision of training courses for search users, especially new joiners and staff in search-intensive roles

***Usability tests***

The scope and schedule for on-site and remote usability testing

***User requirements***

Summarizes the core user requirements

***User interface***

Sets out any proposed changes to the user interface to meet user requirements, including the development, testing, and implementation schedules

***Website search***

Sets out the management and operational links between internal and external search

# APPENDIX B
# Critical Success Factors

After more than a dozen chapters and over 90,000 words, I thought you might find it useful to have a list of critical success factors:

*#1: Content quality is essential for quality search*

Good search technology will quickly reveal poor content. There should be guidelines for content and metadata quality. It is of little benefit to the organization if a search lists 20 algorithmically relevant documents with a content quality that renders them unfit to be trusted. Moreover, with poor quality content, the relevancy may be irrelevant to the user.

*#2: Invest in a search support team*

Without a search support team with the appropriate skills, enthusiasm, and organizational knowledge, any investment in search technology or the creation of content will be wasted. If content can't be found, it can't be used or shared. A good search team will make a very significant difference to search satisfaction and business impact.

*#3: Get the best out of the current investment in search*

There is usually much that can be done to improve the current search applications once the search team and users work together to consider options and define priorities. Only when it is clear that the current technology is not able to support business requirements should a new search application be considered.

*#4: Recognize that enterprise search is an approach and not a technology*

Enterprise search is about creating a managed search environment that ensures employees find the information they need to achieve organizational and/or personal objectives. Be aware that the promise of federated search across multiple applications and repositories has yet to be fulfilled. A strategy for enterprise search is essential to be able to define priorities, allocate resources, and track progress toward objectives.

*#5: Understand user requirements and monitor user satisfaction*

Search logs will indicate the queries that have been used but not the information that was being sought. It is important to understand the business and information context of users, and to monitor user satisfaction with search on a continuous basis using a wide range of methods. Try to obtain evidence of the positive impact of search on business performance.

*#6: Recognize that information discovery involves searching, browsing, and monitoring*

Users need to be able to search when needed, browse when needed, and monitor as needed. These three processes need to be linked together to provide an effective information discovery environment. Any work on creating navigational hierarchies for intranets and other repositories needs to take into account the value of search as well as browse.

*#7: Train and support your users*

Search is not intuitive. It is far more than entering words into the search query box. Make sure that there is a range of online and face-to-face advice available. The process of training will highlight areas for improvement for other users.

*#8: Remember that search is a dialogue*

In an enterprise environment, users will have complex and often ill-defined queries that require them to be able to refine their query and re-evaluate the results with the minimum of effort.

*#9: Pay attention to people search*

It is just as important to understand how well people search is working as the core application, and yet few organizations track people search failure despite highlighting people search in a second search box.

*#10: Regard achieving search excellence as a journey, and not a project*

The process of ensuring that search is meeting user requirements never comes to an end. Every day, there are new employees, new business challenges, new business opportunities, and new developments in search technology. Search should never be a "project"—it should be a way of working.

# APPENDIX C
# Search Blogs

This is a list of blogs that track developments in enterprise search technology and implementation. Many of these blogs are published by search vendors. Others, such as the Real Story Group and Xerox PARC, cover a range of topics, including enterprise search and content analytics. Most of the search vendors have a blog which highlights new product developments and often provides a perspective on industry developments. This list was reviewed in July 2015.

- **Aghy Blog** (Agnes Molnar)
- **All About Search** (Ronald Baan)
- **Attensity**
- **Basis Technology**
- **Beyond Search** (Stephen Arnold)
- **Bing**
- **Breakthrough Analysis** (Seth Grimes)
- **Concept Searching**
- **Coveo Insights**
- **Daniel Tunkelang**
- **Data Dexterity** (Attivio)
- **Do More With Search** (BA Insight)
- **Elastic**
- **Enterprise Search** (Miles Kehoe)
- **Exalead**
- **Flax**
- **Findability** (Findwise)
- **Funnelback**
- **Google Enterprise**
- **IBM Research**
- **Information Interaction** (Tony Russell-Rose)

- **Intranet Focus** (Martin White)
- **LucidWorks**
- **Microsoft Research**
- **Mindbreeze**
- **Mind Over Matters** (Recommind)
- **Opensource Connections**
- **Ravn** (Ravn Systems)
- **Real Story Group Blog**
- **Searchblox**
- **Search Chronicles** (Paul Nelson, Search Technologies)
- **Search Nuggets** (Comperio)
- **Sematext Blog**
- **Sinequa**
- **State of Enterprise Search** (Edwin Stauthamer)
- **Systems Thinking** (Paul Cleverly)
- **Xerox Parc**
- **Yahoo!**

## APPENDIX D

# A Core Library for Enterprise Search

This is a personal selection of books on information retrieval and search that represent a core library for a search support team. The list excludes books on specific search applications so that the list is vendor neutral.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search, Second Edition* (Boston: Addison-Wesley, 2011).

Marcia J. Bates (editor), *Understanding Information Retrieval Systems: Management, Types, and Standards* (Boca Raton, FL: Auerbach Publications, 2011).

Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines* (Cambridge, MA: MIT Press, 2010).

G. G. Chowdhury, *Introduction to Modern Information Retrieval* (London: Facet Publishing, 2010).

W. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice* (Boston: Addison-Wesley, 2010).

Susan Feldman, *The Answer Machine* (San Rafael, CA: Morgan and Claypool, 2012).

Nigel Ford, *Introduction to Information Behaviour* (London, UK: Facet Publishing, 2015).

Gregory Grefenstette and Laura Wilber, *Search-Based Applications: At the Confluence of Search and Database Technologies* (San Rafael, CA: Morgan and Claypool, 2010).

Marti A. Hearst, *Search User Interfaces* (Cambridge, UK: Cambridge University Press, 2009).

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, *Introduction to Information Retrieval* (Cambridge, UK: Cambridge University Press, 2006).

Peter Morville, *Ambient Findability* (Sebastopol, CA: O'Reilly, 2005).

Peter Morville and Jeffery Callender, *Search Patterns* (Sebastopol, CA: O'Reilly, 2010).

Louis Rosenfeld, *Search Analytics for Your Site* (Brooklyn, NY: Rosenfeld Media, 2011).

Tony Russell-Rose and Tyler Tate, *Designing the Search Experience* (Burlington, MA: Morgan Kaufmann, 2012).

Martin White, *Enterprise Search* (Sebastopol, CA: O'Reilly, 2012).

Ryan White, *Interactions with Search Sytems* (Cambridge, UK: Cambridge University Press, 2016).

Max Wilson, *Search User Interface Design* (San Rafael, CA: Morgan and Claypool, 2012).

There are also three important series of books on information retrieval and enterprise search:

Gary Marcionini (Editor), Synthesis Lectures on Information Concepts, Retrieval, and Services (San Rafael, CA: Morgan and Claypool).

Graeme Hirst (Editor), Synthesis Lectures on Human Language Technologies (San Rafael, CA: Morgan and Claypool).

Douglas Oard and Mark Sanderson (Editors), Foundations and Trends in Information Retrieval (Delft, The Netherlands: Now Publishing).

Springer Publishing is another major publisher of books on database management and information retrieval.

## APPENDIX E
# Vendor List

This table lists 60 vendors. In addition to search software vendors, this list includes some vendors that provide a range of ancillary services and some that offer data analysis and business intelligence software. In the case of Google, IBM, Microsoft, Oracle, OpenText, and SAP, no URL is listed, as the pages for search products tend to be far from permanent. This list was created in July 2015.

No warranty, explicit or implied, is given by Intranet Focus Ltd as to the quality of the products listed.

| Company | Country | URL |
|---|---|---|
| Active Navigation | United Kingdom | *http://www.activenav.com* |
| Alcove9 | United States | *http://www.alcove9.com* |
| Ankiro | Denmark | *http://www.ankiro.com* |
| Apache (Lucene/Solr) | Community | *http://lucene.apache.org* |
| Attensity | United States | *http://www.attensity.com* |
| Attivio | United States | *http://www.attivio.com* |
| Autonomy | United States | *http://www.autonomy.com* |
| BA Insight | United States | *http://www.bainsight.com* |
| Basis Technology | United States | *http://www.basistech.com* |
| Cogito | Italy | *http://www.expertsystem.net* |

| Company | Country | URL |
|---|---|---|
| Concept Searching | United Kingdom | *http://www.conceptsearching.com* |
| Constellio | Canada | *http://www.constellio.com* |
| Coveo | Canada | *http://www.coveo.com* |
| Dieselpoint | United States | *http://dieselpoint.com* |
| dtSearch | United States | *http://www.dtsearch.com* |
| ElasticSearch | Community | *http://www.elasticsearch.org* |
| ElasticSearch | Netherlands | *http://www.elasticsearch.com* |
| Exalead | France | *http://www.exalead.com* |
| Exorbyte | United States | *http://www.exorbyte.com* |
| Funnelback | Australia | *http://www.funnelback.com* |
| Google | United States | *http://www.google.com/services* |
| Inbenta | Spain | *http://www.inbenta.com* |
| Intelligenx | United States | *http://www.intelligenx.com* |
| IntelliSearch | Norway | *http://www.intellisearch.com* |
| IntraFind | Germany | *http://www.intrafind.de* |
| Lexalytics | United States | *http://www.lexalytics.com* |
| LTU | France | *http://www.ltutech.com* |
| LucidWorks | United States | *http://www.lucidworks.com* |
| MarkLogic | United States | *http://www.marklogic.com* |
| MaxxCAT | United States | *http://www.maxxcat.com* |
| Mindbreeze | Austria | *http://www.mindbreeze.com/en* |
| Ontolica | Denmark | *http://www.surfray.com* |
| OpenSearchServer | France | *http://www.open-search-server.com* |

| Company | Country | URL |
|---|---|---|
| Perfect Search | United States | *http://www.perfectsearchcorp.com* |
| Perceptive Software | United States | *http://www.perceptivesoftware.com* |
| Q-Sensei | United States | *http://www.qsensei.com* |
| Ravn | United Kingdom | *http://www.ravn.co.uk* |
| Recommind | United States | *http://www.recommind.com* |
| SchemaLogic | United States | *http://www.schemalogic.com* |
| SearchBlox | United States | *http://www.searchblox.com* |
| Searchdaimon | Sweden | *http://www.searchdaimon.com* |
| Simplexo | United Kingdom | *http://www.simplexo.com* |
| Sinequa | France | *http://www.sinequa.com* |
| SLI Systems | New Zealand | *http://www.sli-systems.com* |
| Smartlogic | United Kingdom | *http://www.smartlogic.com* |
| Sphinx | Community | *http://sphinxsearch.com* |
| Synaptica | United States | *http://www.synaptica.com* |
| Temis | France | *http://temis.com* |
| Teragram | United States | *http://www.teragram.com/oem* |
| TeraText | United States | *http://www.teratext.com* |
| Terrier | United Kingdom | *http://terrier.org* |
| Thetus | United States | *http://thetus.com* |
| Thunderstone | United States | *http://www.thunderstone.com* |
| Vivisimo | United States | *http://www.vivisimo.com* |
| Wand | United States | *http://wandinc.com* |
| Xapian | Community | *http://xapian.org* |

| Company | Country | URL |
|---|---|---|
| X1 Technologies | United States | *http://www.x1.com* |
| ZyLab | United States | *http://zylab.com* |

# Glossary

**Adjacent result**
A result that is comparable or analogous to a searched term; often produced by "Find more like this" searches

**Absolute boosting**
Ensuring that a specified document always appears at the same point in a results set, or always appears on the first page of results

**Access control list (ACL)**
Defines permissions to access a specific repository, a set of documents, or a section of a document

**Ambiguity**
A search involving one word with many different meanings, or in a search for an object that can be described many different ways

**Appliance**
A search application pre-installed on a server ready for insertion into a standard server rack

**Approximate pattern matching**
A process in which an algorithm determines the similarity between items—for example, in spellchecking

**Automatic Indexing**
An entirely automated process of converting information into an index

**Auto-categorization**
An automated process for creating a classification system (or taxonomy) from a collection of nominally related documents

**Auto-classification**
An automated process for assigning metadata or index values to documents, usually in conjunction with an existing taxonomy

**Average response time**
An average of the time taken for the search engine to respond to a query, or the average end-to-end time of a query

**Bayesian Inference or Bayesian Statistics**
A probability technique based on the work of Thomas Bayes (1702–1761) and used to determine the relevancy of a given document against a particular query

**BigTable**
A highly scalable database technology that is proprietary to Google

**Boolean Operators**
A widely used approach to create search queries; examples include And, OR, and NOT—for example, *information* AND *management*

**Boolean search**
A search query using Boolean operators

**Boosting**
Changing search ranking parameters to ensure that certain documents or categories of documents appear in the results

**Case-based reasoning**
A technology that allows a system to "learn" by gathering past instances into a "case base" that it can use to solve future problems

**Categorization**
The placing of boundaries around objects that share similarities (e.g., taxonomy)

**Clustering**
A process employed to generate groupings of related words by identifying patterns in a document index

**Collection**
A group of objects methodically sorted and placed into a category

**Computational linguistics**
The use of computer-based statistical analysis of language to determine patterns and rules that aid semantic understanding

**Concept extraction**
The process of determining concepts from text using linguistic analysis

**Connector**
A software application that enables a search application to index content in another application

**Controlled vocabulary**
An organized list of words, phrases, or some other set employed to identify and retrieve documents

**Corpus**
A collection of objects with a defined scope (e.g., all annual reports)

**COTS**
Commercial off-the-shelf software

**Crawler**
A program used to index documents

*See also* Spider

**Description**
A brief statement in a document that effectively summarizes the meaning of a document, often employed to annotate search results

*See also* Key sentence

**Document**
A structured sequence of text information, but often used as a generic description of any content item in a search application

**Document processing**
The deconstruction of a document into a form that can be tokenised and indexed

**Document repository**
A site where source documents or other content objects are stored, generally a folder or folders

*See also* Information source

**Early binding**
A search conducted only across documents that a user has permission to access

*See also* Late binding

**Entity extraction**
The automatic detection of defined items in a document, such as dates, times, locations, names, and acronyms

**Exact match**
Two or more words considered mutually inclusive in a search, often by enclosing them in quotation marks—for example, "United Nations"

**Extract-transform-load (ELT)**
The process of migrating content between databases when undertaken by a single specialized software application

**Facet**
Presentation of topic categories on the search user interface to support the refinement of a search query

**Fallout**

A quantity representing the percentage of irrelevant hits retrieved in a search

**Federated search**

A search carried out across multiple repositories and/or applications

**Field query**

A search that is limited to a specific field in a document (e.g., a title or date)

**Filter**

A function that sets specific criteria for search results

**Free text query**

A search enabling a user to input words in any form, without following any query language criteria

**Freshness**

The time period between a document being crawled and the index being updated so that a user will be able to find the document

**Fuzzy search**

A search allowing a degree of flexibility for generating hits (i.e., matches that are phonetically or typographically similar)

**Golden set**

A set of documents  used to benchmark search performance that is representative of content that will be searched on a regular basis

**Guided search**

A search in which the system prompts the user for information that will refine the search results

**Hit**

A search result matching given criteria; sometimes used to denote the number of occurrences of a search term in a document

**Index**

List containing data and/or metadata indicating the identity and location of a given file or document

**Index file**

A file that stores data in a format capable of retrieval by a search engine

**Indexer (automatic)**

A program that collects data on a given set of files or documents and provides results for a user search

**Indexer (human)**

A person who assigns metadata to a given set of files or documents and makes results available for a user search

**Information source**

The location of indexed documents

*See also* Document repository

**Ingestion rate**

The rate at which documents can be indexed, usually specified in Gb/sec

**Inverse document frequency (IDF)**

A measure of the rarity of a given term in a file or document collection

**Inverted file**

A list of the words contained within a set of documents, and which document each word is present in

**Inverted index**

An index whose entries identify a given word and the documents in which it appears

**Iterative calculation**

A calculation utilizing a recursive and self-referential algorithm

**Key sentence**

A brief statement that effectively summarizes a document, often employed to annotate search results

**Keyword**

A word used in a query to search for documents

**Keyword search**

A search that compares an input word against an index and returns matching results

**Keyword targeting**
A process that helps to ensure the inclusion of given websites in a search for a specific object

**Knowledge extraction**
The extraction of metadata from a given set of objects

**Late binding**
Access permission checking carried out immediately before the presentation of the document to the user

*See also* Early binding

**Lemmatization**
A process that identifies the root form of words contained within a given document based on grammatical analysis (e.g., run from running)

*See also* Stemming

**Lexical analysis**
An analysis that reduces text to a set of discrete words, sentences, and paragraphs

**Linguistics**
The study of the structure, use, and development of language

**Linguistic indexing**
The classification of a set of words into grammatical classes, such as nouns or verbs

**Meta search engine**
A class of search engine that generally retrieves information to user queries by utilizing other search engines

**Meta tag**
An HTML command located within the header of a website that displays additional or referential data not present on the page itself

**Metadata**
Data that provides information about other data (i.e., is data about data)

**Morphologic analysis**
The analysis of the structure of language

**Natural language processing**
A process that identifies content by attempting to adhere to the rules of a given language

**Natural language query**
A search input entered using conventional language (e.g., a sentence)

**Parametric search**
A search that adheres to predefined attributes present within a given data source

**Parsing**
The process of analyzing text to determine its semantic structure

**Pattern matching**
A type of matching that recognizes naturally occurring patterns (word usage, frequency of use, etc.) within a document

**Phrase extraction**
The procurement of linguistic concepts, generally phrases, from a given document

**Precision**
The quantification of the number of correct documents returned in a given search

**Proximity searching**
A search whose results are returned based on the proximity of given words (e.g., *pressure* within four words of *testing*)

**Query by example**
A search in which a previously returned result is used to obtain similar results

**Query performance**
A measure of performance based on the speed a system can receive a query and return results

**Query transformation**
The process of analyzing the semantic structure of a query prior to processing in order to improve search performance

**Ranking**
A value assigned to a specific result returned for a query—the first item listed

has a ranking of 1, the second has a ranking of 2, and so on

**Recall**

A percentage representing the relationship between correct results generated by a query and the total number of correct results within an index

**Relevance**

The value that a user places on a specific document or item of information

**Search results**

The documents or data that are returned from a search

**Search terms**

The terms used within a search field

**Semantic analysis**

An analysis based upon grammatical or syntactical constraints that attempts to decipher information contained in a document

**Sentiment analysis**

The use of of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in documents

**Soundex search**

A search in which users receive results that are phonetically similar to their query

**Spider**

An automated process that provides documents to a data extraction or parsing engine

*See also* Crawler

**Statistical Indexing**

Probabilistic methods relying on mathematics, not "linguistics"

*See also* Bayesian

**Stemming**

A process based on a set of heuristic rules that identifies the root form of words contained within a given document (e.g., run from running)

*See also* Lemmatization

**Stop words**

Words that are deemed to have no value in an index

*See also* Word exclusion

**Structured data**

Data that can be represented according to specific descriptive parameters—for example, rows and columns in a relational database, or hierarchical nodes in an XML document or fragment

**Summarization**

An automated process for producing a short summary of a document and presenting it in the list of results

**Synonym expansion**

Automatically expanding a search by adding synonymns of the query terms derived from a thesaurus

**Syntactic analysis**

An analysis capable of associating a word with its respective part of speech by determining its context in a given statement

**Taxonomy**

In respect to search, the broad categorization of objects (typically a tree structure of classifications for a given set of objects) in order to make them easier to retrieve and possibly sort

**Term frequency**

A quantity representing how often a term appears in a document

**Text Retrieval Conference (TREC)**

A conference held by the National Institute of Standards and Technology in which participants search a collection of documents and present results on various search applications

**Thesaurus**

A collection of words in a cross-reference system that refers to multiple taxonomies and provides a kind of meta-classification, thereby facilitating document retrieval

**Tokenizing**

The process of identifying the elements of a sentence, such as phrases, words, abbreviations, and symbols, prior to the creation of an index

**Truncation**

Removal of a prefix of suffix

**Unstructured information**

Information that is without document or data structure (i.e., cannot be effectively decomposed into constituent elements or chunks for atomic storage and management)

**Vector space**

A model that enables documents to be ranked for relevance against a query by comparing an algebraic expression of a set of documents with that of the query

**Weight**

A value applied to a given area of a search system (e.g., term weighting, which represents its importance with respect to other factors)

**Wildcard**

A notation, generally an asterisk or question mark, that when used in a query, represents all possible characters (e.g., a search for *boo** would return *book*, *boom*, *boot*, etc.)

**Word exclusion**

A list containing words that will not be indexed—this usually is comprised of words that are excessively common (e.g., *a*, *an*, *the*, etc.)

**Word proximity analysis**

An analysis that measures the distance between searched words in a document

# Index

## About the Author

**Martin White** is an intranet and information management strategy consultant, running Intranet Focus Ltd since 1999. He has been a Visiting Professor at the Information School, University of Sheffield since 2002, and was elected a Fellow of the Royal Society of Chemistry in 2006 for his work in information management for the pharmaceutical industry. He is also the author of *The Intranet Management Handbook* and *The Content Management Handbook* (Facet Publishing).

## Colophon

The animal on the cover of *Enterprise Search* is a purple martin (*Progne subis*). It is part of the swallow family and can be found throughout North America. They nest in cavities in temperate climates.

The purple martin has a distinct look amongst other swallows. Males are black in color, but have a noticeable dark blue sheen in the light. Females are of similar coloring, but have lighter undersides on their bodies, mostly consisting of gray shading. Though fairly small in size compared to other birds, the purple martin is the largest of the North American swallow family, averaging almost eight inches from bill to tail.

Like most swallows, the purple martin's diet relies heavily on insects caught while in flight. They have been known to ascend to great heights in order to catch meals, but will also hunt right above water surfaces to satisfy their insectivorous appetite. There has been some evidence that they will forage on the ground for meals, but this is most likely due to inclement weather or other unusual circumstances.

Breeding for the purple martin takes place during the spring months. Males will find nesting territory first (usually man-made bird houses). Females lay three to six eggs per mating season, which are mostly incubated by her. Males may have more than one mating partner at a time. Chicks typically leave the nest around one month after hatching. There has been a decline in the purple martin population due to limited nesting sites, which have been taken over by encroaching starlings and house sparrows.

Many of the animals on O'Reilly covers are endangered; all of them are important to the world. To learn more about how you can help, go to *animals.oreilly.com*.

The cover image is from *Lydekker's Royal Natural History, Vol. 3*. The cover fonts are URW Typewriter and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.