



The Skeptical Searcher

or

How to Give Up The Full Review and Discover What Matters

Daticon EED

Discovery is overrun by the sheer volume of email and electronic files. The numbers of emails and files potentially that contain responsive information make full review impractical. These numbers continue to grow at a rate that will not abate any time soon.

One study estimates that the volume of new information grows at about 30% per year. Ninety two percent (92%) of all new information is generated by computer and stored primarily on hard disk drives. Only one tenth of one percent (0.1%) of all new information is stored on paper.¹ Our own experience across hundreds of litigation support discovery projects is that the volume of electronic data now exceeds the volume of paper by at least five to one. Only three years ago, only one in five projects even had an electronic discovery component.

Recent surveys illustrate an electronic discovery market at one half to one billion dollars per year or more, with growth estimated as fast as 60% per year for several years. Driving these rates is a growing awareness that computer files, particularly emails, are a rich source of evidence. But because of the growing size of that source of evidence, all industry participants need is looking for cost containment, more capacity, greater knowledge, and better standards in managing electronic discovery.²

Litigators and investigators need precise and efficient tools to find and review critical evidence in corporate email, file systems, and backup tapes. Reviewing it all is no longer an option.

Why Not Review It All?

First, a full review for most projects would likely include a large volume of irrelevant and unnecessary emails. Emails are stored in most corporate systems by individual account, not by subject. Most of us with email accounts are either poor organizers of the important, notorious packrats of the unimportant, or at best, not consistent among us in what we keep or how we keep it. Most email software and

systems make it difficult to comprehensively search, review and tag emails. As a result, for most discovery projects, it is necessary to preserve and collect whole accounts lest important relevant information be left behind, but then cut out unnecessary review. The case studies below demonstrate some outcomes.

Second, courts and regulators are encouraging parties to embrace the technology, for economy to clients and for speed and efficiency in the decision making process.³

Third, and most compelling, is that it serves our clients well to handle discovery efficiently, from preservation, interviews, scope negotiations and disclosures right through review, evaluation and production of evidence. In all of these areas, the courts, the regulators, and the legal marketplace penalize the technologically slow-to-react.

Real Savings Using Search Technology

Two projects using search methodologies illustrate the potential savings in time and cost.

Case Example 1 below illustrates a mergers & acquisitions second request in which more than half a terabyte of data was collected on behalf of more than 300 employees. The data was loaded and indexed over a ten-day period. The legal teams submitted lists of search terms to identify potentially relevant as well as potentially privileged documents. In addition, search criteria were applied to custodian identity, document types, and date ranges, as negotiation on those criteria were continuous.

Nearly 120 attorneys took part in the review. Less than three months elapsed from the beginning of data loading and indexing until the applicants were in substantial compliance with the second request. The production was made available to various regulators by an online review application (ASP service) without paper, TIFFs, CDs, or subset databases.

The size of the Case Example 1 data collection, the number of custodians, and the need for speed drove the application of technology. The legal teams had the ability to search and to test search terms, the ability to identify potentially responsive and privileged documents, the ability to shuffle particular custodians into and out of the review set, and the ability to make separate date cutoffs for various custodians. With these capabilities, the legal team could tailor and negotiate the scope of review and production even while reviewing and producing important materials on a rolling basis.

Compared to Case Example 2 below and many other projects, the average document size in Case Example 1 is rather larger than usual. Typically, we expect across emails and electronic files an average document size of about five pages. Typically, across many collections, we estimate page equivalence at between 50,000 to 75,000 pages per gigabyte.

Case Example 2, below, illustrates how potentially 1.2 million pages or 480 boxes of emails and files, if all printed, can reduce to 136,800 pages or the equivalent of about 55 boxes to review. Probably because it involves an

older data collection with older versions of email (circa late nineties), it does not have the huge volumes of email per person that we see more typically today. Still, it contains a typical mix of emails, laptop and desktop files, and network shared files. It also included selective restoration and inclusion of folders from network servers, rather than restoration and inclusion of the contents of the entire server.

In Case Example 2, legacy data was pulled from a set of backup tapes for the relevant time period. Email accounts for forty individuals were extracted from legacy email system post offices for the relevant custodians. In addition, files were pulled from individuals' network shares and from a key departmental shared drives. The collection consisted primarily of small emails, short spreadsheets and other office documents. The email accounts averaged about 1,800 messages and 600 attachments, or the equivalent of about 15,000 pages. The file systems on computers and network accounts for each of the forty averaged another 1,000 computer files, or about another 5,000 pages. Shared network servers department server or two with another 80,000 stored files, or 400,000 pages if printed.

Case Example 1

320 business clients

Average 1.6 gB email each

Average 140 mB files each

Average 11 pages per document in all

Estimated Volumes:

	Gigabytes	Documents	Page Equivalent
Volume submitted to Search Capability	550	2,700,000	33,000,000
Submitted for review after application of searches	160	875,000	9,600,000
Produced after review	135	750,000	8,250,000

Estimated Volume Saved in Review 1,825,000 23,400,000

Case Example 2

40 business clients

Average 200 mB email each

Average 100 mB files each

Estimated 8gB shared file data

Average 2.8 pages per document in all

Estimated Volumes:

	Gigabytes	Documents	Page Equivalent
Volume submitted to Search Capability	20.00	420,000	1,200,000
Remaining after removal of non-documents	19.00	399,000	1,140,000
Remaining after identification of duplicates	15.20	319,000	912,000
Submitted for review after application of searches	2.28	47,880	136,800
Produced after review	1.71	36,000	102,600

Estimated Volume Saved in Review 271,120 775,200

The automated steps were as follows: First, the data was subject to a first-pass automated filtering. The filter detected

and set apart from further processing whatever files that may be software or system files not containing user-created data.

Second, search technology was used to identify and quantify duplicates. De-duplication was performed by mathematical algorithm called MD5 hash, in which a calculation is used to assign a practically perfectly unique value to the content of a file. In some collections, the percentage of duplicates is as high as 55%. In others, it is as low as 8 to 10%. Duplicates are identified and tracked so that multiple copies of identical files can receive the same treatment during review, but need not actually be reviewed.

Third, the collection was searched with a refined list of search terms. For large collections in which clients work on many different projects, it is typical to see only 5-15% of emails and files relevant to a particular subject or project. In the illustration above, by the use of search terms, the collection is reduced by 85% of the remaining total.

If printed, the collection would have yielded 1.2 million pages for review. Instead, the legal team reviewed about 11-12% of that number. By any industry or professional standards, the costs for review (let alone printing!) 1.2 million pages greatly exceeds the litigation support costs of searching and culling 20 gigabytes and putting the equivalent of 136,800 pages into a litigation support software for review. This scenario is typical of the application of search terms to email and office files.

Courts Support the Pre-Review Search

The productivity and cost savings leveraged by the technology are attractive to the savvy litigator. But many attorneys still cling to the notion that all of a client's potentially relevant material must be reviewed. The notion is not as much skepticism of the technology as it is a matter of discovery compliance, lest some single piece of responsive information be overlooked.

Opposing counsel, courts, and regulators are willing to accept or even require the use of searches to identify potentially relevant materials and to limit the scope of review. There is substantial legal authority for the use of search technology to find potentially relevant material and to narrow the volume of review. Consider at least these authorities, and forge ahead:

The clearest and most recent judicial pronouncement comes in Zubulake V, though specifically on the issue of preservation. "To the extent that it may not be feasible for counsel to speak with every key player, given the size of a company or the scope of the lawsuit, counsel must be more creative. It may be possible to run a system-wide keyword search; counsel could then preserve a copy of each "hit." Although this sounds burdensome, it need not be. Counsel does not have to review these documents, only see that they are retained. For example, counsel could create a broad list of search terms, run a search for a limited time frame, and then segregate responsive documents. When the opposing party propounds its document requests, the parties could negotiate a list of search terms to be used in identifying responsive documents, and counsel would only be obliged to review documents that came up as "hits" on

the second, more restrictive search. The initial broad cut merely guarantees that relevant documents are not lost."⁴

The Sedona Principles directly address use of search terms as a means to limit review. Principle 11 states that "[a] responding party may satisfy its good faith obligation to preserve and produce potentially responsive electronic data and documents by using electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data most likely to contain responsive information." Comments to Principle 11 elaborate with illustrations methodology for sampling, collection, and search "The scope of terms employed must be reasonably calculated to return relevant data."⁵

Zakre v. Norddeutsche Landesbank Girozentrale cites The Sedona Principles toward a ruling that a respondent's form of electronic data production was sufficient. "Nord/LB has conducted a review of the emails for privileged documents, but has not conducted a review for responsiveness to Zakre's specific document requests. *The Sedona Principles: Best Practices, Recommendations & Principles for Addressing Document Discovery* (Sedona Conference Working Group Series 2004). Principle 11 states that: A responding party may satisfy...."⁶

Courts have directly or implicitly been involved in directing parties in the use of search terms. In Rowe v. William Morris et al., after describing the basis for cost shifting, the Court laid out the protocol for using search terms while protecting potentially privileged materials. "Plaintiffs' counsel shall formulate a search procedure for identifying responsive e-mails and shall notify each defendant's counsel of the procedure chosen, including any specific word searches. Defendants' counsel may object to any search proposed by the plaintiffs.... Once an appropriate search method has been established, it shall be implemented by the plaintiffs' expert. Plaintiffs' counsel may then review the documents elicited by the search on an attorneys'-eyes-only basis. The plaintiffs may choose the format for this review; they may, for example, view the documents on a computer screen or print out hard copy."⁷

In re Ford Motor Company, Ford appealed an order allowing opposing parties to search Ford's databases directly, without restriction or condition. On appeal, the Eleventh Circuit found that the requesting parties had found no discovery abuses, that Ford had already produced relevant non-privileged materials from the database, and that the discovery rules did not allow the requesting party unconditional access. It went further and as a practical matter, indicated that "[t]he district court established no protocols for the search. The court did not even designate search terms to restrict the search."⁸

Regulators Are Prepared To Negotiate

The Federal Trade Commission's pronouncement⁹ on the use of search terms has effectively set the standard for dealing

with the FTC, the Department of Justice, and the Securities and Exchange Commission in many kinds of investigations. As a practical matter, it sets the ground rules for negotiation among parties in civil litigation. The letter states, in part,

“...b. Use of Term Searches. Staff frequently is asked to agree that a production resulting from a term search will be sufficient, regardless of the number of documents the search produces. Staff also has been asked to edit and provide input on the acceptability of terms to be searched. While a prohibition against term searches for parties' increasingly voluminous electronic document databases seems unreasonable, so does a request to agree, in advance, that a specific term search is all a party need do, regardless of the search's efficacy. We view term searches as a mechanism which, if used properly, may enable parties to respond adequately, up to and including substantial compliance. A thorough and well-executed term search of electronic files may be an efficient way to respond to a second request, just as a thorough and well-executed physical review may be.”

To make search term lists more complete, the Staff advises parties to consult organizational charts and industry and company glossaries in devising search terms. It suggests consulting on sample searches and results, and providing rolling productions, so that any deficiencies in the methodology can be recognized early.

Recognizing both the importance of email in its investigations but also the huge and duplicative volumes of email to be found on backup tapes, the Staff indicates that it may forego requiring backup tapes unless it appears that important files are inexplicably missing from the “live” email collection. Still, it may negotiate “significant limitations to the portion of archives and backups that needs to be searched and produced (either by person, dates, or terms/specifications)”

The Staff specifically recognized (a) use of search terms, (b) limiting the number of people to be searched, and (c) reducing the time period of the search of emails for different levels of employees within an organization all as an effective means of limiting the volume to be produced. Accordingly, the organization's structure and the employees' roles and responsibilities are critical to understanding which people are considered for search.

The staff recommends in negotiating names, dates and search terms a “stipulation to produce any responsive documents not identified in the term search, but identified by other means, before certifying substantial compliance.” This condition serves notice on parties that compliance, not merely execution of negotiated search methodology, still governs the proceedings.

What Do I Need For Successful Searching?

There are three basic requirements to achieving the optimal speed and economy of pre-review search: (1) a targeted

collection strategy, (2) a defensible search strategy, and (3) a flexible search capability.

Targeted Collection Strategy: The basis for everything searched and reviewed, and in fact, the basis by which to discover what matters in any particular case is limited by the potential evidence that is collected. There are preservation concerns and strategies to address, but they are discussed in more detail in other papers. Here, it is important to note the following: Under-collection may miss important evidence. Over-collection (by far more common in electronic discovery) burdens the client with excessive costs and the service providers and legal team with unnecessary processing or review. While search strategies can make short work of large volumes of irrelevant information, there is still a cost impact to submitting those volumes to the process.

How Do I Build A Defensible Search Strategy?

A search strategy consists of the methods by which the legal team will search the data, evaluate the results, feed information back into the search mechanism, record progress and results, and be prepared to explain to a regulator, an opposing counsel, or a court how the search was designed to find potentially relevant information, and how it did, in fact, accomplish its design.

Clearly, performing a search prior to review represents a big change from the review of everything. In the paper world where information is evaluated first onsite during collection, then reviewed page by page, the legal team could assure itself of a high degree of discovery compliance so long as it took reasonable care and well-worn steps during the sweep and the review. The search prior to review very likely will miss documents that can be relevant. The courts recognize that there must be a tradeoff to getting through the mountain of data, and that search prior to review represents a reasonable effort. But what steps must be taken to execute a reasonably compliant search?

In any search of structured or unstructured information, there are documents that are “hit” that are relevant, documents that are hit that are not relevant, and documents that are relevant that are not hit. These concepts are important in determining the effectiveness of searching unstructured information.

Relevance is typically considered to be a subjective quality. An attorney knows it when s/he sees it, so long as s/he knows enough about the discovery request and the context of the case. A whole document is relevant to an information need if and only if it contains at least one sentence relevant to that need.

The *precision* rate is the proportion of retrieved documents that are relevant. In our common Internet-search experience, precision represents the documents you seek when you search that you actually receive as search results. Precision is expressed as a percentage: what percentage of the documents in my search hit set are relevant? *Fallout* is the proportion of

retrieved documents that are not relevant. *Recall* is the proportion of relevant documents retrieved. The recall rate is the proportion of relevant documents retrieved, compared to the total number of relevant documents that exist.¹⁰

Attorneys using search terms as part of a litigation support solution want their recall rates to be high—to capture in the search and review set as many of the relevant documents as possible that are “out there”.

Effort toward a high recall rate is a key element of defensibility. It indicates that a serious effort was made to get a high proportion of relevant documents included in the search results and into the review. Presumably, if these documents are not privileged, then they are produced.

The typical means to achieve a high recall rate in complex collections is to add more keywords. But the effect of more keywords is to contribute to fallout and to sacrifice precision. The net effect is that as more relevant documents are added to the search hit set by use of more keywords, considerably more irrelevant documents are added as well. This contribution to fallout works against the savings in time and money that the search strategy is intended to achieve.

So how do you use search terms defensibly but guard against reviewing more than necessary? Here is a basic strategy:

- ✓ Use search terms derived from client interviews, lists of names and email addresses, from paper productions, from company organizational charts and glossaries, and from key documents. Be specific about file types, dates, sources, key names, & key words, with their roots, stems and variations. Purely brainstormed lists of search term lists can be overly broad or narrow.
- ✓ Test-search, review the statistics by term or custodian, and determine whether the results are what you expect, given what you know about the case, the custodians and data sources.
- ✓ Sample-review portions of the collection where no hits are found, among the data of custodians who otherwise have responsive data. Looking outside the expected data helps to validate your lists of search terms.
- ✓ Revise and re-test the search terms and document the testing and results, particularly if you are negotiating the search with opposing counsel or regulators. Add Boolean logic or other restrictions to manage search terms with multiple meanings in context.
- ✓ Employ separate sets of search terms to help you find both responsive and privileged data. An experienced analyst can help facilitate your review by segregating for review the responsive, non-responsive, and potentially privileged materials.

This strategy employing *feedback*¹¹ effectively sets the standard for keyword or Boolean searching before review. Having an experienced analyst within the legal team helps to guide the process and to document the strategy. This is a strategy by which many attorneys have become comfortable with a pre-review search.

How do you get the feedback you need? First, you can search interactively, look at the results, and build your search list interactively. The procedure is reasonably exact, but tedious and time consuming.

More useful to the electronic discovery setting is a set of statistical reports that indicate where and how many search hits resulted: by custodian, by type of document, and whether among emails, stand-alone files, databases, or other sources of information. An attorney who has had a hand in client interviews and development of a list of key custodians has a good idea of who should have potentially relevant documents and on what subjects. A well-crafted set of counts and tallies provides the attorney with the feedback required to make intelligent adjustments to the search.

Defensibility and Newer Technologies

Used properly, keyword searching with Boolean logic has proven savings and defensible results. New tools advance efforts toward automated feedback and visual display of documents related by concept. These will continue to play an increasingly important role in pre-review search strategy so long as we (1) “look under the hood” to see how they work and (2) use that information to assess our rates of recall and precision.

Clustering is a means by which documents that are related by some classification are visually grouped together. The groupings may be done by user input or user feedback, by word or term weighting, or by any of a number of techniques or algorithms for automated classification or “concept” assignment. Clustering is in and of itself a form of search, with graphical depiction of search results. To understand the utility of clustering for purposes of defensibility in pre-review searching, it is important to understand the means by which clustering is accomplished. Here are a few:

Some search engines use a technique called parsing and natural language analysis. They apply rules of grammar and lexicons are applied to try to explicitly understand textual information. These have a high degree of success in automated classification by weighting of the frequency of words in context, but face some difficulty when presented with words with multiple meanings. The validity of weightings suffers when significant (relevant) unique words show up among more heavily used insignificant words. Natural language analysis and parsing is effective in highly controlled vocabularies (like case law or even deposition transcripts) but less effective with highly structured data.

Collaborative filtering is a form of automated feedback with user input. It attempts to allow computers to make personal recommendations to users based on their similarity to other users. The basic principle is quite simple: by getting a large number of users to give information about their preferences the system endeavors to make recommendations. Collaborative filtering requires intensive participation from its usership. It may not be particularly scalable over either a large unstructured collection or a large number of users.

Shannon's principles¹² assume that the less frequently an expression occurs, the more meaning it has. The principle mathematically takes into account the uniqueness of words and phrases in assessing relevance or importance among more commonly used words and phrases. Shannon's theory is the cornerstone of communications engineering and has proved itself many times over. Bayesian Inference is a mathematical structure behind a form of automated feedback. It is used to assess how well the concepts in a document match the query an agent is executing. This also provides a mechanism for the technology to "learn", making subsequent queries even more accurate. Engines using Shannon's principles and Bayesian inference can be used to produce meaningful results with any type of content and in any language, without metadata or tagging.¹³

How Do I Build A Search Capability?

Flexible Search Capability: The search capability consists of the search software, the algorithms and analysts used to devise the search and generate the searches, the means by which searches and results are fed back to the legal team, and the infrastructure required potentially to manage large volumes of information over many aspects of a single case or over many cases.

- ✓ **Search Software:** Will the search capability accomplish what you need to find the relevant documents? Can you perform iterative searches? Can you search separately for privileged documents? Can you intersect search sets to look for specialized information? Can you disjoin search sets, so that if new issues arise or new cases arise using the same collection, you can search anew without including previously searched and reviewed materials?
- ✓ **Search Infrastructure:** What will the hardware and database management system (DBMS) bear? Can it support the full text of 100,000, a million, or ten million documents or more? Can it support the database, plus indexes for a reasonably large group of simultaneous users?
- ✓ **De-duplication strategy:** What methods are there for identifying duplicates across the collection? What methods are there for identifying duplicates within individual custodians' collections? Can you or are you looking for "near-duplicates" or is their uniqueness important?

- ✓ **Metadata:** Does the search capability extract metadata into searchable categories you may need for a litigation support application? Does it allow you to independently search particular metadata fields. In the context of native file preservation and review, does it sufficiently store metadata in a way that will be useful for parsing by, e.g., names, dates, locations, custodians?
- ✓ **Chain of Custody:** In the context of native file production, or production by means not yet foreseeable, do you have the ability to trace the review and production set, first, back to the native file, and second, back to the original source of that native file?
- ✓ **Rolling Productions:** For the larger projects, can you subset your review to accomplish rolling productions?
- ✓ **Decision and Case Tracking:** What mechanisms do you have to track your search and hit history, review and edit history, and production and exhibit history?
- ✓ **Feedback Mechanism**

What About Computer Forensics?

The same principles apply to computer forensics, though in a microcosm. Cases in which there are contests over the control and content of hard drives lead to negotiations over search protocol and limited review for privileged and confidential information.

First, no one would collect or agree to collect *all* of the hard drives forensically unless there was an overriding case reason for it, such as allegations of widespread fraud, theft of trade secrets, or some evidence of spoliation. Here, as in the megacases, there are tactical decisions about how and how much data to collect.

Second, it is not feasible to review *all* of the recoverable files and fragments even from a single drive; a single modern hard disk drive can yield tens or even hundreds of thousands of pages. So the legal team carefully directs the forensic examiner in what to search and how to report.

Discover What Matters

Corporations involved in litigation and regulatory investigations and the corporate law departments who manage outside counsel are looking for legal and tactical skill as well as efficiency in managing discovery: crisp interviews that cover both paper and electronic data; solid communications with IT staff about email servers and backup tapes; a preemptive and tactical approach to scope, preservation and disclosure; targeted strategies for collection, search and review.

Law firms whose litigation departments handle big cases or smaller firms with specialty discovery practices compete successfully for clientele when they demonstrate technical savvy combined with a plan: *a collection strategy, a search capability, a reasonable and defensible search strategy, and a strategy for review*, rather than say, “We’ll decide when we get in there and see the paper and the data. We’ll let you know what we come up with”.

Insurance companies footing the bill understand that parties willing and able to pursue and to negotiate these strategies can

Endnotes

¹ In 2002, the volume of *newly created data alone* exceeded 5 exabytes, or 10^{18} bytes, or 5 million terabytes. “How Much Information? 2003”, University of California Berkeley School of Information Management and Systems. Regents of the University of California, October 23, 2003. www.sims.berkeley.edu/research/projects/how-much-info-2003/

² Surveys by Socha-Gelbmann and EDDix, Inc. illustrate a huge impact on the discovery process: Demand for electronic discovery services may be between \$400 million and \$1.3 billion in the past year, and may continue to grow at a rate of 60% per year over the next several years. Socha & Gelbmann, “EDD Showcase: It’s an Explosion”. Law & Technology News, August 2004. “The 2004 Socha-Gelbmann Electronic Discovery Survey”, www.sochaconsulting.com. “Electronic Discovery In Litigation – EDD Supplier Landscape”, www.eddixllc.com, September 30, 2004.

³ See e.g., the discussions on disclosures, searchability and forms of production in proposed changes to F.R.C.P Rules 16, 26 and 34 and Committee Notes, “Proposed Rules Amendments Published for Comment “ August 2004. www.uscourts.gov/rules/newrules1.html; www.uscourts.gov/rules/comment2005/CVAug04.pdf

⁴ Zubulake v. UBS Warburg LLC et al. (“Zubulake V”) 02 Civ. 1243 (SAS) SDNY (July 20, 2004).

⁵ “The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production.” www.thesedonaconference.org, January 2004

⁶ Zakre v. Norddeutsche Landesbank Girozentrale, No. 03 Civ. 0257(RWS). S.D.N.Y. April 9, 2004. 2004 WL 764895 See also In re Lorazepam and Clorazepate Antitrust Litig., 300 F.Supp.2d 43, 46 (D.D.C.2004).

⁷ Rowe v. William Morris et al. 51 Fed. R. Serv. 3d (West) 1106; *aff’d*, 53 Fed. R. Serv. 3d (West) 296 (S.D.N.Y. 2002).

⁸ In re Ford Motor Company 2003 U.S. App. LEXIS 19531; 345 F.3d 1315 (11th Cir. September 22, 2003) No. 03-10440, D.C. Docket No. 01-01759 CV-S.

⁹ Statement of the Federal Trade Commission’s Bureau of Competition On Guidelines for Merger Investigations (December 2002, as searched on January 12, 2004 <http://www.ftc.gov/os/2002/12/bcguidelines021211.htm>)

save a lot of time and money by not unnecessarily collecting, processing or reviewing, and by looking for opportunities to share production costs. They are looking for outside counsel and service providers who *get it*.

To deal with the volumes of information currently facing us in electronic discovery (not just in the “mega-cases”) we do, in fact, have powerful tools and defensible procedures to make shorter work of it.

¹⁰ VanRijsbergen, C. J. *Information Retrieval*, 2nd © Van Rijsbergen, C. J., 1999. Information Retrieval Group, Department of Computing Science, University of Glasgow. <http://www.dcs.gla.ac.uk/~iain/keith/>

¹¹ Ibid., VanRijsbergen, C. J, at 105. “A user confronted with an automatic retrieval system is unlikely to be able to express his information need in one go. He is more likely to want to indulge in a trial-and-error process in which he formulates his query in the light of what the system can tell him about his query. The kind of information that he is likely to want to use for the reformulation of his query is: (1) the frequency of occurrence in the data base of his search terms; (2) the number of documents likely to be retrieved by his query; (3) alternative and related terms to be the ones used in his search; (4) a small sample of the citations likely to be retrieved; and (5) the terms used to index the citations in (4). All this can be conveniently provided to a user during his search session by an interactive retrieval system. If he discovers that one of his search terms occurs very frequently he may wish to make it more specific by consulting a hierarchic dictionary which will tell him what his options are. Similarly, if his query is likely to retrieve too many documents he can make it more specific.”

¹² Shannon, Claude “A Mathematical Theory of Communication” (1948), outlines what we now know as Information Theory, described the measurement of information by binary digits representing yes-no alternatives - the fundamental basis of today’s telecommunications.

¹³ Butler, Martin “Automated Management of Unstructured Content” Butler Direct Limited, February 2003 <http://www.autonomy.com/content/News/Releases/2003/0225.en.html>