THE REGNET PROJECT

Manual and Technology-Based Approaches to Using Classification for the Facilitation of Access to Unstructured Text

By Charles H. Heenan

Engineering Informatics Group
Department of Civil and Environmental Engineering
Stanford University
Stanford, California 94305

Email: heenan@stanford.edu

January 2, 2002

Acknowledgement and Disclaimer

This report is intended to review the current state of practice for classifying large volumes of unstructured text. This review has been performed as part of the Regnet Project, which is funded by the National Science Foundation under Grant No. EIA-0085998.

Any opinions, findings, and conclusions or recommendations expressed in this report are those of the author and do not necessarily reflect the views of the National Science Foundation.

Introduction

Statement of Purpose

The purpose of this document is to discuss current methods for structuring and facilitating access to unstructured text. The scope of the evaluation is limited to commercial or open-source technologies. The research for this document was conducted as part of the Regnet Project at Stanford University. The Regnet Project is funded by the National Science Foundation and is focused on the application of information technology to regulation management and regulatory compliance.

Methodology and Organization

The research for this document included telephone interviews with people working in industry, online research into classification technology and information retrieval, analysis of company websites, analysis of company and third-party white papers on information retrieval, as well as evaluation of specific products where possible.

The first section of this paper begins by defining terms and by situating the problem of accessing unstructured text within the larger domain of information management and information retrieval. Next, it explores the concordance and the index as two approaches to the problem of facilitating access to unstructured text. The first section ends with the introduction of classification as a more powerful approach for enabling access to unstructured text than either the concordance or the index.

The second section of this paper focuses on the technology marketplace for classification products used to facilitate access to unstructured text. This marketplace breaks down into three parts:

1. Information Discovery products, used for the design of a classification structure;
2. Text Classification products, used for the population of an existing classification structure; and
3. User Interfaces, used for accessing the contents of a populated classification structure.

Each segment of the marketplace is addressed, including general discussion of that segment as well as specific evaluations of products. These evaluations are not intended to critique or to endorse any specific products. Rather, such evaluations serve to highlight product features for illustrative purposes. A summary of the software products reviewed is given in the Appendix. The list is not meant to be exhaustive.

In the conclusion, the various approaches for the population of an existing classification structure are compared in terms of the benefits and drawbacks for each approach. Each approach involves tradeoffs between the cost of that approach and the quality of its results. Which of these is the "right" approach for a given organization depends upon that organization's specific needs for a text-mining information retrieval solution.

Section 1 – Definition of Terms, and Context

Terms such as taxonomy, ontology, information management, knowledge management, and information retrieval each can carry slightly or significantly different meanings depending upon context. In fact, among terminology purists strong disagreements can arise over what is an appropriate use of one term or another. The present discussion will cover these and related concepts, so it is worth positing here working definitions for the central terms while acknowledging that there is not an industry-wide consensus on their meaning.

Taxonomy and Ontology

In its most general sense, a taxonomy is a hierarchical classification structure in which the child nodes in the structure inherit or share in common some properties held by their ancestor nodes, but the reverse is not true. From one taxonomy to the next, there can be variations in the nature of the relationship between descendent nodes and ancestor nodes and there can be variations in what is being classified. Generally speaking, however, within one taxonomy the nature of this relationship should remain consistent.

The term "ontology" is more slippery. For some, a taxonomy is but one example of an ontology and other examples exist that use different organizing principles. For others, the two terms are synonymous and interchangeable. At its most general, there seems to be some consensus that an "ontology" is a framework of relations between entities in which the relationships are based upon the abstract essences of those entities and not upon the particulars of the entities themselves. For example, a pure ontological approach might treat the English word "car" as if it were essentially the same as the French word *voiture* because in the abstract they refer to the same idea.

In this paper, the general terms "classification hierarchy," "categorization hierarchy," "classification structure," and "categorization structure" will be used in lieu of the terms "taxonomy" or "ontology" in order to avoid confusion or disagreement. Furthermore, although for some there is a meaningful difference between the verb "to classify" and the verb "to categorize," in this document the terms are treated as interchangeable.

Information and Knowledge

In the Information Management industry, it is often said that information is unprocessed data or un-interpreted text, while knowledge is the meaning gleaned from the act of processing or interpreting information. According to this conceptualization, a database of voting results for 50 years' worth of national elections contains information. An interpretation of the trends in voting behavior that are revealed by an analysis of that information represents knowledge. Likewise, an online document repository of ten thousand interview transcripts contains information, while an interpretation of the conceptual themes common to most of those interviews represents knowledge. However, this question of knowledge is an interesting one. Specifically, can "knowledge" as described above exist outside of a human mind? If the interpretation of trends in voting behavior is encapsulated in a chart or a paper document, is it still knowledge? Or does it revert to being "information" again until

someone reviews the chart or reads the document? Even then, the chart and the document do not change in the process of being read (assuming the reader does not have edit privileges). What changes is the level of knowledge within the reader's mind. One could argue that knowledge only exists in the human mind, and it exists as a result of the process of interpreting information, even when the information being interpreted is itself an interpretation of other information. Since the term "knowledge" carries this ambiguity of meaning, in this paper the terms "knowledge" and "knowledge management" are avoided in favor of the broader terms "information" and "information management."

Information Management and Information Retrieval

Information Management is the field concerned with collecting, storing, and facilitating access to information. In broad terms, the information management field ranges from filing cabinets and rolodexes to document management systems, personal computers, Palm-like hand helds, relational databases such as Oracle or Microsoft SQL Server, and data warehousing offerings from companies such as Veritas or Legato Systems.

Information Retrieval is a field within information management that deals with facilitating access to information. The entities being retrieved by an information retrieval system traditionally have been either numbers or text, although current research is focused on facilitating access to video and audio archives as well as databases of images. In this vein, in Summer 2001 Google released on its site the first version of its image-search offering.[1]

For information retrieval of text, structured text and unstructured text are two distinct information types. Structured text is part of a larger data structure, in which the text is defined in terms of its form, its content, or its purpose. For example, in a database of customers, the First Name and Last Name fields will always contain structured text. The possible contents of the First Name cell are highly constrained: the field is unlikely to contain more than 25 to 35 characters, it is unlikely to contain punctuation marks or numeric symbols, it will probably not contain significant conceptual meaning, and its purpose is limited to identifying one customer among many.[2]

Unstructured text is not constrained in this way. The form, content, and purpose of unstructured text is not easily predicted based upon data-type alone. In the same database of customers, unstructured text might be found in a Customer Service Call Center Notes field, in which telephone customer service representatives type notes regarding the subject of each call. Similarly, in an online archive of newspaper articles, the actual text of each article is unstructured text while the contents of the "Title" and "Author" fields are not.

[1] See http://images.google.com/

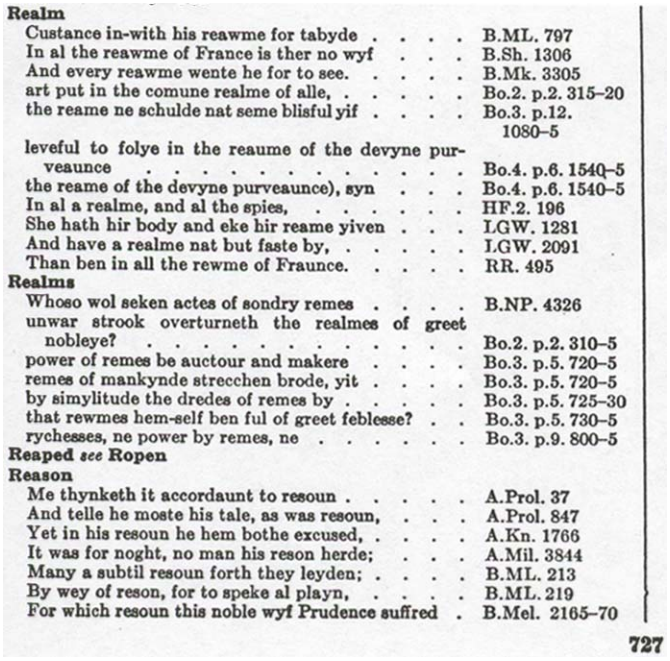
[2] Of course, Neither First Name alone nor First Name with Last Name is sufficient to identify someone uniquely in all cases.

Non-Classification-Based Approaches to the Problem of Unstructured Text

While it is a simple task to search for articles or books by title or author, searching on the conceptual content of unstructured text is a different issue. The problem is how to enable searches not only on titles or author names, but on the subject matter of the text itself.

The Concordance

As early as the 14th century, the goal of enabling searches on the conceptual content of text gave rise to the concordance, an alphabetical arrangement of the principal words contained in a book along with citations of the passages in which they occur. In some cases, an entry would show the search term along with a certain number of words immediately preceding and following. Originally, these tools were developed for the Bible, to show in how many texts of scripture any word occurs. Later, concordances were developed to enable concept-search on non-religious texts, including the complete works of Chaucer (See Figure 1) and of Shakespeare. [3]



Realm
Custance in-with his reawme for tabyde . . . . B.ML. 797
In al the reawme of France is ther no wyf . . . B.Sh. 1306
And every reawme wente he for to see. . . . . B.Mk. 3305
art put in the comune realme of alle, ... Bo.2. p.2. 315–20
the reame ne schulde nat seme blisful yif Bo.3. p.12. 1080–5
leveful to folye in the reaume of the devyne purveaunce Bo.4. p.6. 1540–5
the reame of the devyne purveaunce), syn . . . Bo.4. p.6. 1540–5
In al a realme, and al the spies, ... HF.2. 196
She hath hir body and eke hir reame yiven . . . LGW. 1281
And have a realme nat but faste by, ... LGW. 2091
Than ben in all the rewme of Fraunce. . . . . RR. 495

Realms
Whoso wol seken actes of sondry remes . . . . B.NP. 4326
unwar strook overturneth the realmes of greet nobleye? Bo.2. p.2. 310–5
power of remes be auctour and makere Bo.3. p.5. 720–5
remes of mankynde strecchen brode, yit Bo.3. p.5. 720–5
by simylitude the dredes of remes by Bo.3. p.5. 725–30
that rewmes hem-self ben ful of greet feblesse? . . Bo.3. p.5. 730–5
rychesses, ne power by remes, ne Bo.3. p.9. 800–5

Reaped *see* **Ropen**

Reason
Me thynketh it accordaunt to resoun . . . . . A.Prol. 37
And telle he moste his tale, as was resoun, . . . A.Prol. 847
Yet in his resoun he hem bothe excused, ... A.Kn. 1766
It was for noght, no man his reson herde; . . . A.Mil. 3844
Many a subtil resoun forth they leyden; . . . . B.ML. 213
By wey of reson, for to speke al playn, ... B.ML. 219
For which resoun this noble wyf Prudence suffred . B.Mel. 2165–70

727

Figure 1: Search results for the term "reason" in a manually-developed concordance to the works of Chaucer. Note that the standard spelling for the search term yields two variant spellings, "reson" and "resoun." This concordance was started in 1871 with a dozen volunteers. Each volunteer was assigned a portion of text, and was to write each line of text on a slip of paper. Volunteers had to note variant spellings for each word, the definition of each word, its inflectional form, and the rhyming relationships for the final word in every line. The concordance was not published until 1927, and then only in an incomplete form. Source: A Concordance to the Complete Works of Geoffrey Chaucer, The Carnegie Institute of Washington, 1927.

[3] Source: Oxford English Dictionary Online. See http://www.oed.com

To the extent that a concordance makes available the locations and contexts of a given search term, it is an early example of the full-text keyword searches available with modern computers. Essentially, in early concordances all the likely keyword searches have been conducted manually, in advance. In this sense, a pre-computer concordance can be thought of as a static representation of some fraction of all possible searches—while today's computer-enabled full-text keyword search can be thought of as a dynamic concordance generator.

The blurred distinction in the computer era between concordances and keyword search can be seen at Concordance.com.[4] The site, which has about 150 e-texts available, packages as an online concordance what are essentially computer-enabled variations on keyword search. Figures 2 and 3 show a search at Concordance.com on the term "reason" in George Berkeley's Treatise Concerning the Principles of Human Knowledge.

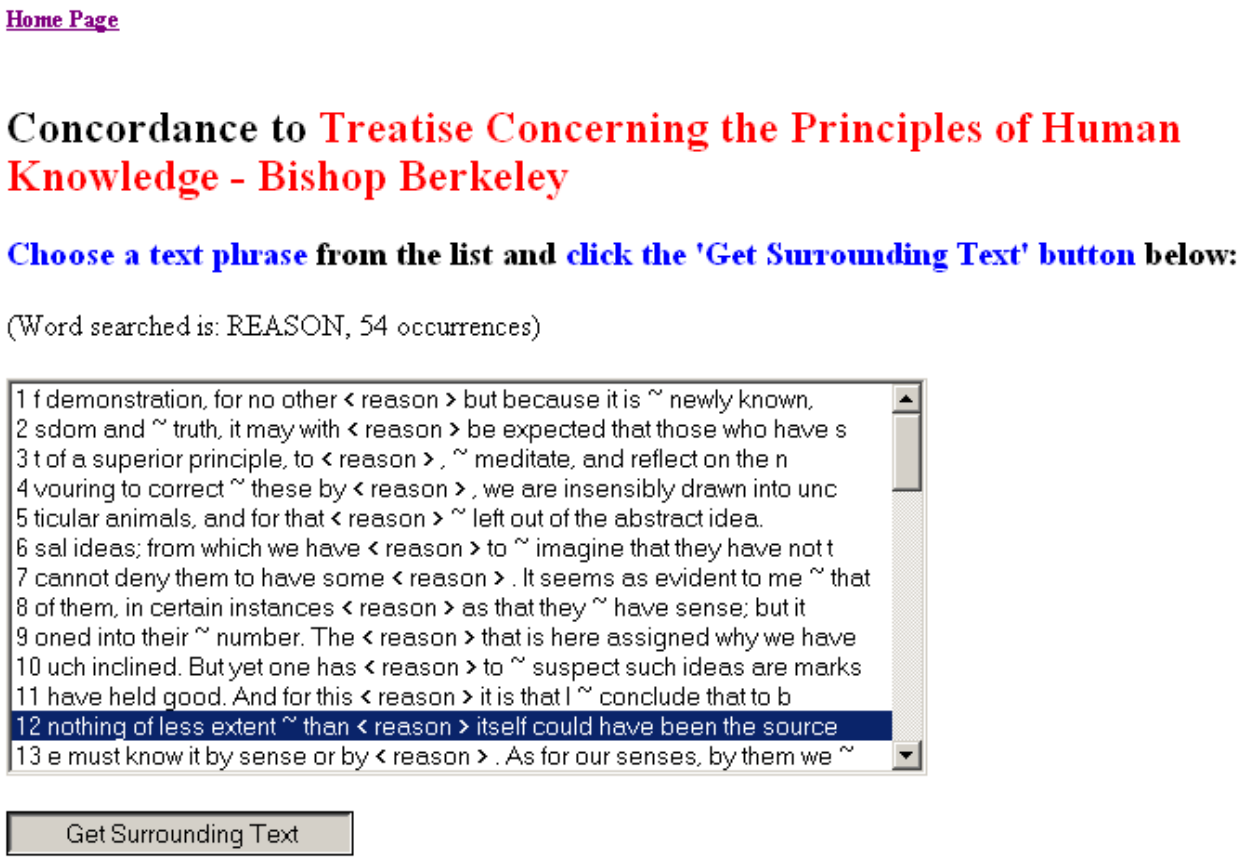


Figure 2: Concordance.com's results page for a search on "reason" in George Berkeley's Treatise Concerning the Principles of Human Knowledge.

[4] See http://www.concordance.com A good description of the searches available at this site is at http://www.concordance.com/instruct.htm

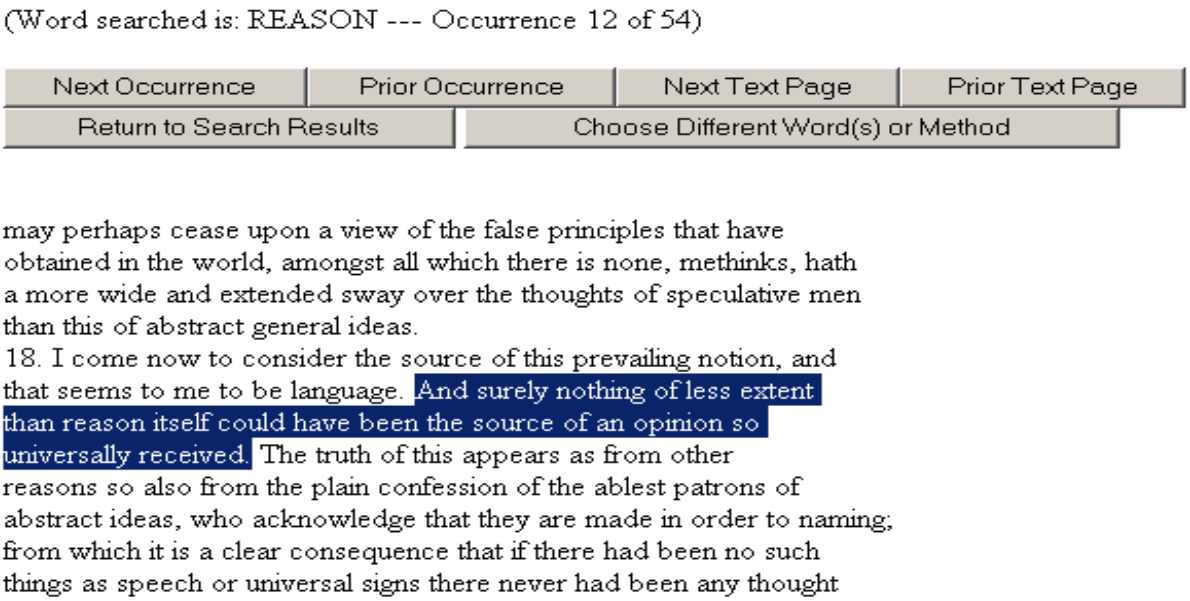
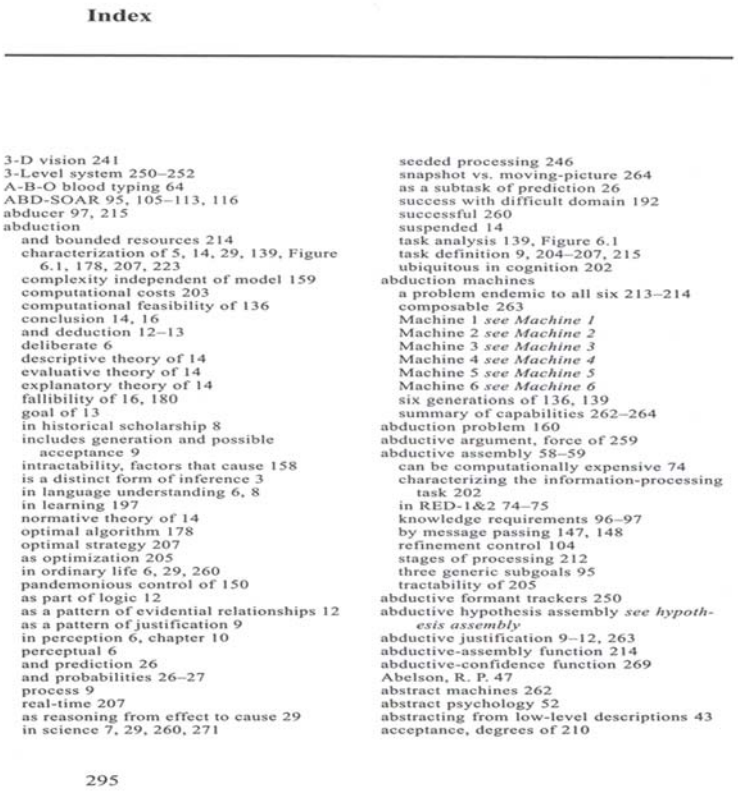


Figure 3: Concordance.com's Surrounding Text page for one of the occurrences of search term "reason," in George Berkeley's Treatise Concerning the Principles of Human Knowledge.

The Index

The index came into use as another means to locate ideas within a work of unstructured text. Found at the end of a book, or as a separate volume or volumes at the end of a series of books, an index is an alphabetized list of the proper names and concepts occurring in a text, along with an indication of the places in which they occur (See Figure 4 for an example of a Back of Book Index). Unlike a concordance, an index does not contain extracts of the text surrounding a search term for context—perhaps because this allows the index to include a greater amount of reference information in a compact space.

Descendants of the early indexes exist in today's world of computing in two forms relevant to the present discussion. Like the concordance, both of these forms serve to enable searches within a body of unstructured electronic text. The first functions in essentially the same manner vis-à-vis the content of a website as does a back-of-book index to the text of that book. This form of index is exemplified by a page on a website with an alphabetical listing of the main names, concepts, and offerings on the site. HTML index pages only differ from back-of-book indexes in that the location information comes in the form of hyperlinks rather than Chapter-Section-Page citations. (N.B.: HTML index pages that are analogous to back-of-book indexes should not be confused with the "home" page of a website, which often has a file name of "index.html." or "default.html.") See Figure 5 for an example of an HTML "Back of Site" index.



Index

Figure 4: Back of Book Index. From Abductive Inference: Computation, Philosophy, and Technology, page 295. John R. and Susan G. Josephson, Editors.

IMDb Index

If you're curious about the wide variety of features here at IMDb or if you're trying to find a particular one, this is the place to come. Though it's not an exhaustive list of every single page here, it's a good road map for finding your way around.

Advertising	Get more information about advertising and promotional opportunities on our site
Awards	Nominees and winners for hundreds of awards (i.e. Oscars®, Emmys, Golden Globes) and festivals (i.e. Sundance, Cannes)
Ballot	Rate a variety of movies in our weekly ballot
Birthdays	Who was born on your birthday?
Bottom 100	The 100 lowest rated movies according to our users
Box Office	Weekly winners at the U.S. and U.K. box office... and more
Browse	Browse and search through various categories of information in our database.

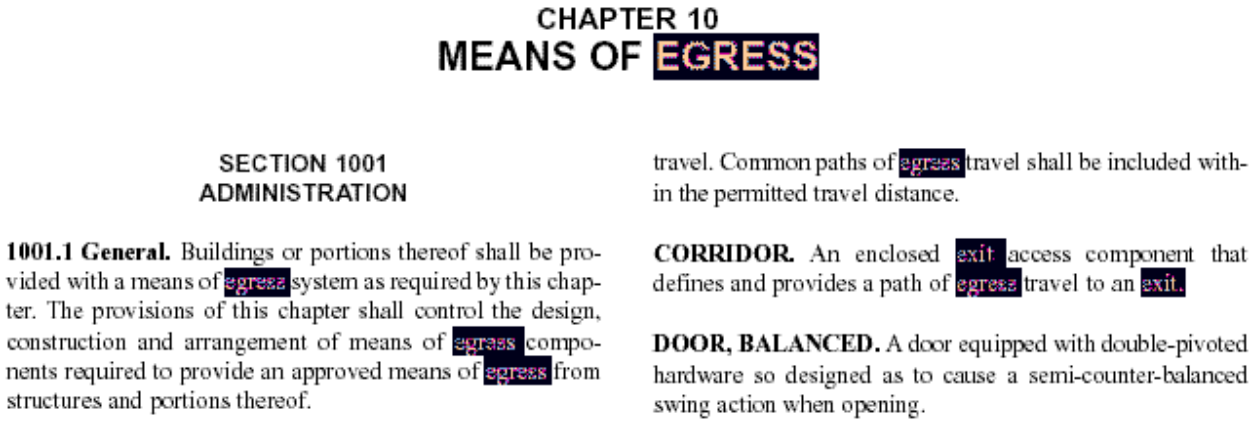
Figure 5: "Back-of-Site" index page for the Internet Movie Database website (http://us.imdb.com/a2z)

Although these HTML indexes improve upon printed back-of-book indexes, in that following references is easier, they do not break significant new ground in information retrieval. Another form of index is particularly interesting because it does just that. While its purpose is the same as other indexes—to enable the location of ideas within unstructured text—its implementation exploits the potential of indexes in ways that were impossible before computers. Often found as components of CDROM-based archives of unstructured text, this form of index is written not for an end user to use directly, but for computers to use in fulfilling the search queries of end users. The user experience is not significantly different than basic full-text keyword search. Nonetheless, the results can be more relevant by virtue of having been filtered through the index.

For example, with both keyword search and index-mediated search, the end user types search terms into a search field. However, for a given string, basic full-text keyword search will compare the string against all characters in the text corpus and return any strings that match. Most keyword search offerings allow the user to restrict the search to full words (as opposed to sub-strings of words). Many keyword search offerings allow stemming as well as wildcard characters. Despite these features, full-text keyword search can miss relevant search results, such as multi-word phrases or single words that are synonymous to the search term.

In contrast, index-mediated keyword search offers the benefits of full-text keyword search with the benefits of editorial knowledge about synonym meanings and other domain-specific information that might be relevant to improving search effectiveness. In this type of search, when the user enters a search term, the computer first compares it to entries in the index. When it finds a match, it looks in the index for synonyms to the term and for any other information impacting how it should execute the search. The computer will then use that information to inform how it executes the search on the corpus of text. See Figures 6 and 7, respectively, for an index-mediated search and the results of that search.

Figure 6: Index-Mediated Keyword Search functionality on the International Building Code CD-ROM. Here, the search term is "exit." Word Stemming and Thesaurus options are turned on.



CHAPTER 10
MEANS OF EGRESS

SECTION 1001
ADMINISTRATION

1001.1 General. Buildings or portions thereof shall be provided with a means of egress system as required by this chapter. The provisions of this chapter shall control the design, construction and arrangement of means of egress components required to provide an approved means of egress from structures and portions thereof.

travel. Common paths of egress travel shall be included within the permitted travel distance.

CORRIDOR. An enclosed exit access component that defines and provides a path of egress travel to an exit.

DOOR, BALANCED. A door equipped with double-pivoted hardware so designed as to cause a semi-counter-balanced swing action when opening.

Figure 7: Results of Index-Mediated Search on the term "exit." Note that the synonym "egress" is also retrieved as a result of the index's thesaurus function.

Index-mediated searches similar to the examples above are increasingly common, and so the question of how a given index was made is of increasing importance to the comprehensiveness and accuracy of searches on unstructured text.

At their simplest, computer-readable indexes will include manually-created non-hierarchical synonym equivalents for a list of terms. However, if there is no a-priori consensus on the meaning of the terms, then searchers can face two problems: First, they can receive incomplete results if the terms in the index are not the search terms people are likely to use. Second, they can receive irrelevant results if the synonym relationships are based upon meanings from a different information domain than that of the searcher.

The development and use of controlled vocabularies is one approach to improving the quality of index-mediated search. A controlled vocabulary establishes within a limited domain a set of standard word meanings and synonym relationships for those words. By using a controlled vocabulary as part of index-mediated search, a system can be responsive to likely search terms and irrelevant results can be reduced.

Classification as an Approach to the Problem of Unstructured Text[5]

Although basic controlled vocabularies can help improve the quality of index-mediated search, more complicated relationships than synonymy exist between terms. Take four terms: A, B, B`, and C. Term B can be at once a synonym of B`, a broader term than C, and a narrower term than A. These relationships are meaningful to the searcher, but enabling a search system to take advantage of such relationships requires that they first be organized into a formal classification structure. The most sophisticated forms of index-mediated search incorporate full-fledged classification structures into

[5] The Dewey Decimal System is perhaps the most widely known use of classification for access to unstructured text. For more information, see http://www.oclc.org/dewey/about/about_the_ddc.htm.

their indexes. The classification structures can often be browsed directly, or queried by a computer in response to a user entering a search term. However, the quality of the search experience will depend upon the quality of the classification structures against which the searches will run. Consequently, when implementing a classification structure as a part of a search system, it is important to have an understanding of the nature of classification and its role in information retrieval.

On the Topic of Classification

The assignment of entities to categories is a central human activity, necessary for higher-level thought. It is also a central component in the more advanced methods for facilitating access to unstructured text. Professor Kenneth Bailey of UCLA is author of a monograph on methods of classification.[6] His introduction sheds light on the issue:

> Classification "is almost the methodological equivalent of electricity—we use it every day, yet often consider it to be rather mysterious. It is one of those things that we all use without knowing very much about how it works."
>
> "In its simplest form, classification is merely defined as the ordering of entities into groups or classes on the basis of their similarity. Statistically speaking, we generally seek to minimize within-group variance, while maximizing between-group variance. This means that we arrange a set of entities into groups, so that each group is as different as possible from all other groups, but each group is internally as homogenous as possible. By maximizing both within-group homogeneity and between-group heterogeneity, we make groups that are as distinct (nonoverlapping) as possible, with all members of a group being as alike as possible. These are general goals that specific classification techniques may alter somewhat."
>
> …
>
> "Almost everything is classified to some degree in everyday life, from chewing gum (bubble and nonbubble), to people (men and women), to animals, to vegetables, to minerals. Grouping objects by similarity, however, is not quite as simple as it sounds. Imagine that we throw a mixture of 30 knives, forks, and spoons into a pile on a table and ask three people to group them by "similarity." Imagine our surprise when three different classifications result. One person classifies into two groups of utensils, the long and the short. Another classifies into three classes,—plastic, wooden, and silver. The third person classifies into three groups,—knives, forks, and spoons. Whose classification is 'best'?"
>
> …
>
> "The lesson here should be obvious—a classification is no better than the dimensions or variables on which it is based. If you follow the rules of classification perfectly but

[6] Bailey, Kenneth D. Typologies and Taxonomies: An Introduction to Classification Techniques, Sage University Papers, Series on Quantitative Applications in the Social Sciences, 07-102, 1994. Thousand Oaks, CA: Sage.

> classify on trivial dimensions, you will produce a trivial classification. As a case in point, a classification that they have four legs or two legs may produce a four-legged group consisting of a giraffe, a dining-room table, and a dancing couple. Is this what we really want?"
>
> "One basic secret to successful classification, then, is the ability to ascertain the key or fundamental characteristics on which the classification is to be based. A person who classifies mixtures of lead and gold on the basis of weight alone will probably be sadder but wiser. It is crucial that the fundamental or defining characteristics of the phenomena be identified. Unfortunately, there is no specific formula for identifying key characteristics, whether the task is theory construction, classification, or statistical analysis. In all of these diverse cases, prior knowledge and theoretical guidance are required in order to make the right decisions."

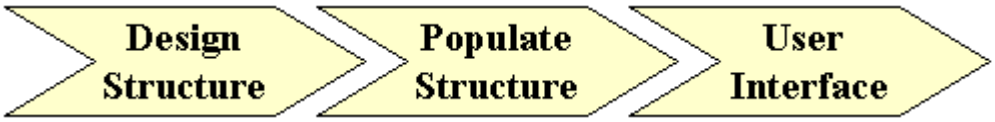
Bailey mentions the ubiquity of classification in our daily lives. Classification is ubiquitous because it adds structure to the variety of objects and ideas that exist. In so doing, classification defines those objects and ideas and enables us to communicate, reason, argue about, or simply reference that which has been categorized. Yet the fact that different people are likely to classify forks, knives, and spoons in different ways is indicative of the need for formal classification structures. The examples Bailey offers—and indeed, much of the on-the-fly classifications that people do—involve simple, flat classification structures (binary, ternary, quaternary, etc. differentiations lacking a hierarchy). No explicit assumptions are made regarding the choice of one grouping over another and there is no attempt at sub-classification or super-classification. Consequently, such casual classifications lead to incomplete "definitions" of the classificand and errors in communication can result. It is far more descriptive to talk of a fork as being at once a serving fork (type of fork), made of silver (type of material), and in need of cleaning (physical state) than it is simply to talk of it on just one of these dimensions. With the added defining information they contain about a given classificand, formal hierarchical classification structures are powerful search facilitators.

The development of an appropriate, stable classification hierarchy is a critical task in the creation of a classification-based unstructured-text information retrieval system. By assigning unstructured text to positions within a hierarchy, the text is converted from an unstructured to a structured state. Once categorized, the text is defined by its position in the structure. As a result, in developing the hierarchy it is important to think about what type of logical relationship should exist between categories and subcategories. "Is-A" relationships are usually best because they are the most stable, although "Part-To-The-Whole" and "Is-A-Process-Of" relationships have been used in corporate information management efforts.

Regardless of the type of logical relationship between categories and subcategories, the relationship should be consistent throughout the structure. Finally, it is important that the designer of the structure make explicit and that the end user be aware of the assumptions under which the structure was created. Just as with non-hierarchical synonym lists, searchers can receive incomplete results if they use search terms that don't appear in the classification hierarchy. Likewise, they can receive irrelevant results if they search in a classification hierarchy designed for a different information domain than that of their search goals. With an awareness of these assumptions, searchers can more accurately conduct their research. Without this awareness, information can disappear in the ether just as surely as if it had been censored.

Section 2 – The Technology Marketplace

Thankfully, concordances no longer require decades spent manually conducting thousands of keyword searches and compiling the results into bound volumes. Today, a number of companies offer technology to help facilitate access to unstructured text.



Different companies target different segments of the process of using classification to facilitate access to unstructured text. The reviews in this document focus on three main tasks:

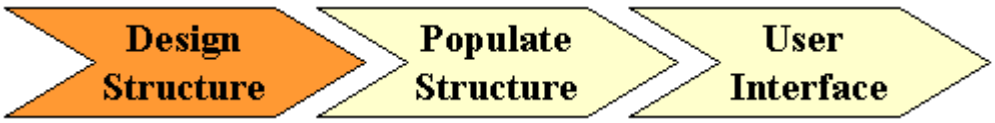
1. The design or selection of an appropriate classification structure is the foundation of any category-based unstructured text retrieval system.

2. Once a hierarchy of categories has been developed or has been selected, the structure must be populated with text. After this point, it is no longer appropriate to refer to the text as unstructured. The act of populating a classification structure with text transforms that text from unstructured to structured form.

3. A suitable user interface allows searchers to access the structured textual information that now resides in the hierarchy.

In practice, the process should be dynamic through time. That is, as new documents on new topics are subject to classification, the structure must evolve if its existing hierarchy is to represent adequately the information in the text collection. The process of populating the classification structure must occur at regular intervals as well, depending upon the rate at which new documents become available. The user interface does not necessarily have to change, although it is always good to strive for improvements in usability and information-display.

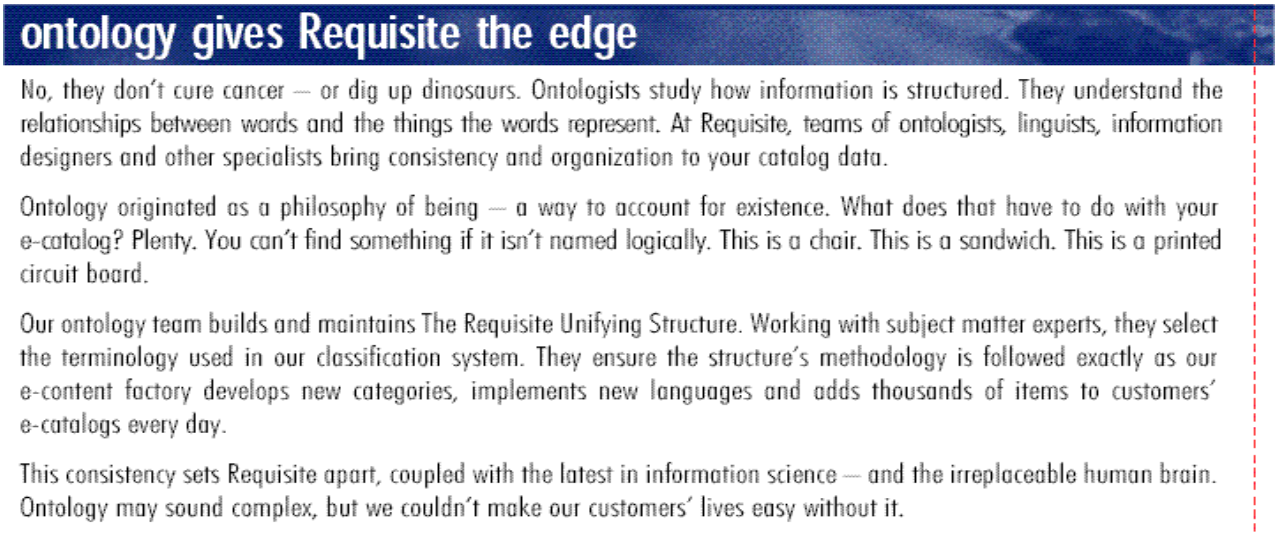
The Design of a Classification Structure



Although some organizations are exploring ways to derive concept hierarchies from text automatically, the process of designing a classification structure is still primarily the realm of human intelligence. As Professor Bailey notes, "it is crucial that the fundamental or defining characteristics of the [classificands] be identified." Unfortunately, "there is no specific formula" for this so "prior knowledge and theoretical guidance are required in order to make the right decisions." As a result, when organizations need to develop an overarching categorization structure they often turn to consultancies, which sell man-hours in addition to technology solutions. Among the consulting

firms that have done work in this space are Accenture (formerly Andersen Consulting), PriceWaterhouseCoopers, and Booz, Allen.

Other companies devote internal resources. Yahoo has had a position of Chief Ontologist since as early as 1996, tasked with overseeing Yahoo's classification structure for their directory of web content. The business-to-business marketplace developer VerticalNet has an ontology group to ensure that their classification hierarchy enables site visitors to find the products or services they need. Another business-to-business company, Requisite Technology, has linguists in its ontology group to manage the classification framework for products and product descriptions in its e-procurement catalog offerings.[7] See Figure 8 for Requisite's description of what their ontology group does.



ontology gives Requisite the edge

No, they don't cure cancer — or dig up dinosaurs. Ontologists study how information is structured. They understand the relationships between words and the things the words represent. At Requisite, teams of ontologists, linguists, information designers and other specialists bring consistency and organization to your catalog data.

Ontology originated as a philosophy of being — a way to account for existence. What does that have to do with your e-catalog? Plenty. You can't find something if it isn't named logically. This is a chair. This is a sandwich. This is a printed circuit board.

Our ontology team builds and maintains The Requisite Unifying Structure. Working with subject matter experts, they select the terminology used in our classification system. They ensure the structure's methodology is followed exactly as our e-content factory develops new categories, implements new languages and adds thousands of items to customers' e-catalogs every day.

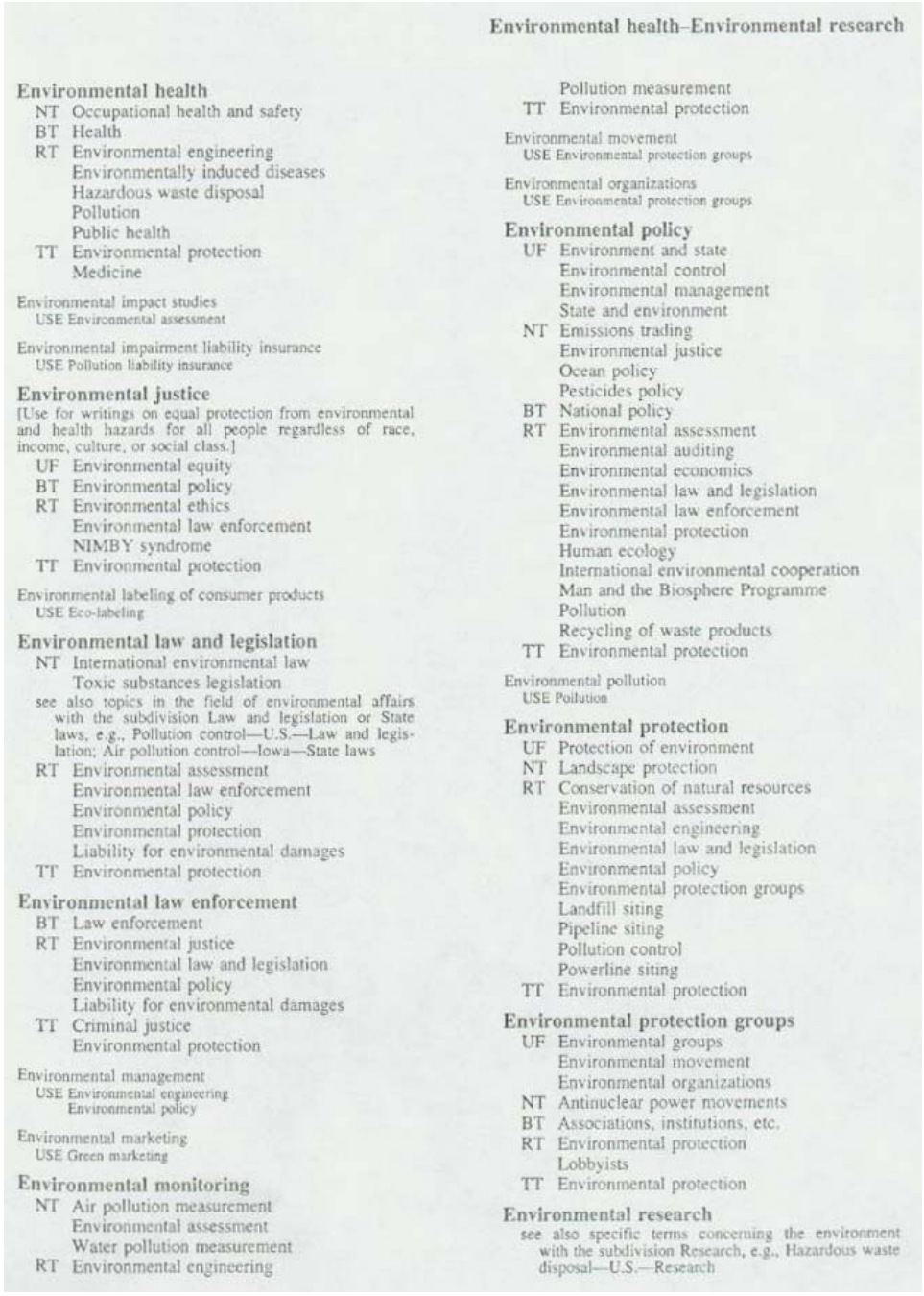
This consistency sets Requisite apart, coupled with the latest in information science — and the irreplaceable human brain. Ontology may sound complex, but we couldn't make our customers' lives easy without it.

Figure 8: Requisite Technology marketing collateral on the role of their ontology group.
Source: http://www.requisite.com/pdf/rus.pdf

Requisite employs a two-level tree of categories and subcategories and its system includes a range of attribute information on classificands—such as manufacturer name and product name. Of note is that Requisite adheres to the "Is-A" relationship between the parent and child categories in their hierarchy.

In some cases, an industry group rather than a specific company takes on the task of defining a categorization structure. In numerous industries, teams of librarians have organized industry-specific vocabulary terms in relation to other terms. The resulting industry-specific thesauri provide a broad, shallow categorization structure that can form the basis for more specific hierarchical categorization work. Some examples of this type of structure are the Medical Subject Headings (MESH) thesaurus, and the Legislative Indexing Vocabulary (LIV). Figure 9 shows several entries on environmental terms from the LIV thesaurus.

[7] See http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2638502,00.html



Environmental health–Environmental research

Environmental health
NT Occupational health and safety
BT Health
RT Environmental engineering
Environmentally induced diseases
Hazardous waste disposal
Pollution
Public health
TT Environmental protection
Medicine

Environmental impact studies
USE Environmental assessment

Environmental impairment liability insurance
USE Pollution liability insurance

Environmental justice
[Use for writings on equal protection from environmental and health hazards for all people regardless of race, income, culture, or social class.]
UF Environmental equity
BT Environmental policy
RT Environmental ethics
Environmental law enforcement
NIMBY syndrome
TT Environmental protection

Environmental labeling of consumer products
USE Eco-labeling

Environmental law and legislation
NT International environmental law
Toxic substances legislation
see also topics in the field of environmental affairs with the subdivision Law and legislation or State laws, e.g., Pollution control—U.S.—Law and legislation; Air pollution control—Iowa—State laws
RT Environmental assessment
Environmental law enforcement
Environmental policy
Environmental protection
Liability for environmental damages
TT Environmental protection

Environmental law enforcement
BT Law enforcement
RT Environmental justice
Environmental law and legislation
Environmental policy
Liability for environmental damages
TT Criminal justice
Environmental protection

Environmental management
USE Environmental engineering
Environmental policy

Environmental marketing
USE Green marketing

Environmental monitoring
NT Air pollution measurement
Environmental assessment
Water pollution measurement
RT Environmental engineering
Pollution measurement
TT Environmental protection

Environmental movement
USE Environmental protection groups

Environmental organizations
USE Environmental protection groups

Environmental policy
UF Environment and state
Environmental control
Environmental management
State and environment
NT Emissions trading
Environmental justice
Ocean policy
Pesticides policy
BT National policy
RT Environmental assessment
Environmental auditing
Environmental economics
Environmental law and legislation
Environmental law enforcement
Environmental protection
Human ecology
International environmental cooperation
Man and the Biosphere Programme
Pollution
Recycling of waste products
TT Environmental protection

Environmental pollution
USE Pollution

Environmental protection
UF Protection of environment
NT Landscape protection
RT Conservation of natural resources
Environmental assessment
Environmental engineering
Environmental law and legislation
Environmental policy
Environmental protection groups
Landfill siting
Pipeline siting
Pollution control
Powerline siting
TT Environmental protection

Environmental protection groups
UF Environmental groups
Environmental movement
Environmental organizations
NT Antinuclear power movements
BT Associations, institutions, etc.
RT Environmental protection
Lobbyists
TT Environmental protection

Environmental research
see also specific terms concerning the environment with the subdivision Research, e.g., Hazardous waste disposal—U.S.—Research

Figure 9: Not your basic Roget's Thesaurus: An industry-specific thesaurus of terms. UF means that the entry is "Used For" the terms following the UF designation. NT (Narrower Term) denotes terms that are narrower in scope than the entry. BT (Broader Term) denotes terms that are broader in scope than the entry. RT denotes "Related Terms." Source: Legislative Indexing Vocabulary, The CRS Thesaurus, Library Services Division, Congressional Research Service of the Library of Congress. 22nd Edition, December 1998.

Top-Down and Bottom-Up Design of Categorization Structures

A classification hierarchy can be developed from a top-down perspective, a bottom-up perspective, or a combination of both. A top-down hierarchy can be created with pencil and paper by a person simply thinking about meaningful ways to break a topic into categories and sub-categories. When applied to a collection of documents, however, the document content may not easily break down into the categories that have been created. Similarly, the documents may include a wide range of topics that are not accounted for in the top-down categorization structure.

A bottom-up categorization hierarchy can be created by looking through the document collection, letting it "speak to you" about which topics are important. In this case, one should pay attention to the dispersion of topics throughout the texts, as well as which topics are central versus of ancillary importance. When applied to the document collection for which it was developed, this custom-fit solution can be extremely productive in facilitating search. However, as new documents are added to the corpus of text the information topography is likely to change. New topics may become prominent, or existing topics may be discussed in different levels of depth relative to the overall body of text. As a result, the document collection can outgrow this sort of classification structure. Likewise, the custom-fit hierarchy may not be portable to documents in other domains.

A hybrid approach is likely to yield the best results. In a hybrid approach, the top-level categorization is informed by—but not driven by—the content of the documents themselves. Hybrid classification structures require less-maintenance and are more portable than custom-fit bottom-up structures, yet they can be more responsive than top-down classifications are to the changing content of the corpus over time.

Unfortunately, both bottom-up and hybrid approaches to the classification of text can be labor-intensive if humans are needed to audit the text collection for conceptual drift. As a result, it is in this area that technology can play a strong role in the design of categorization structures. Technology that allows a categorization-builder efficiently to assess the conceptual topography of a document collection will minimize the labor-intensive aspects of creating/auditing bottom-up and hybrid classification structures.

Technology Assistance for Building Classification Structures – Cartia and SPIRE

Cartia, Inc. was one company which had a product offering in this area.[8] Cartia's natural language algorithms ran against unstructured text, identifying relationships between concepts of central, secondary, and tertiary importance. Cartia used natural language filtering to remove noise words such as "the" and "a" in order to reduce the text to words that carry conceptual content. Next, the statistical frequency of the remaining words would be calculated. Of note was Cartia's claim that it could use context to account for polysemy, as with the word 'bank,' which could refer to the shore alongside a river, an array of telephones, or a place that stores money.

[8] In the technology downturn of 2000/2001 Cartia may have gone out of business or been acquired, as by 8/01 its website was no longer up, no one responded to the DNS email address of record, and the phone line to its corporate headquarters goes unanswered. It is still worth mentioning, however, as its approach was unusual.

After each word had been analyzed in context, the separate "units of meaning" were mapped in relation to one another on a two-dimensional topographic map containing peaks and troughs (See Figure 10). The greater the similarity between any two documents, the closer together they would appear. Concentrations of documents about a similar topic formed peaks, and the distance between peaks represented how closely those topics are related. On the topographic map, clicking on one of many small black circles (each circle represents a document) would allow searchers to access the original document.

The core technology for the Cartia topographic map was developed by the U.S. military as part of the SPIRE program (Spatial Paradigm for Information Retrieval and Visualization) on information visualization.[9] See Figure 11 to view the SPIRE topographic map of themes. SPIRE remains an active research program at the government's Pacific Northwest National Laboratory, of Richland, Washington.[10]

Figure 10: Cartia's Document Topography Interface.

[9] See http://showcase.pnl.gov/show?ENTER_LESSON&tours/it/infoviz
[10] See http://www.pnl.gov

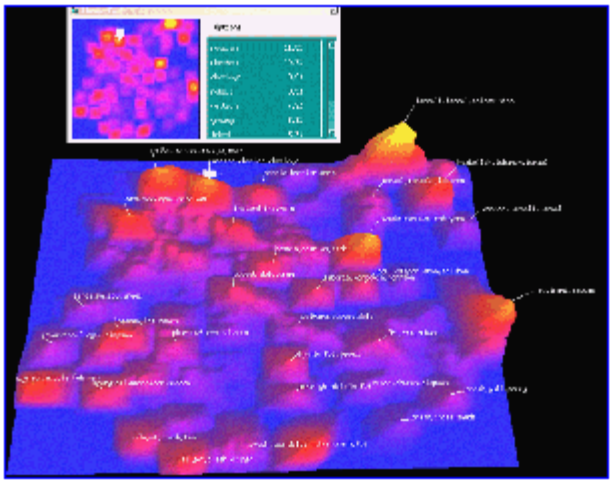


Figure 11: A "ThemeView" concept topography, from the SPIRE site at the Pacific Northwest National Lab. The underlying technology for "ThemeView" and for Cartia is the same.

Technology Assistance for Building Classification Structures – SemioMap Discovery

Semio Corporation of San Mateo, California, also has an information visualization tool that enables information discovery. SemioMap Discovery is a visual interface into phrase co-occurrence relationships in a document collection. Semio crawls all text in a text collection, extracts noun phrases from that text, and creates lexical maps of phrases based upon the frequency of phrase co-occurrence. Figure 12 contains an example of a lexical map created by SemioMap Discovery.

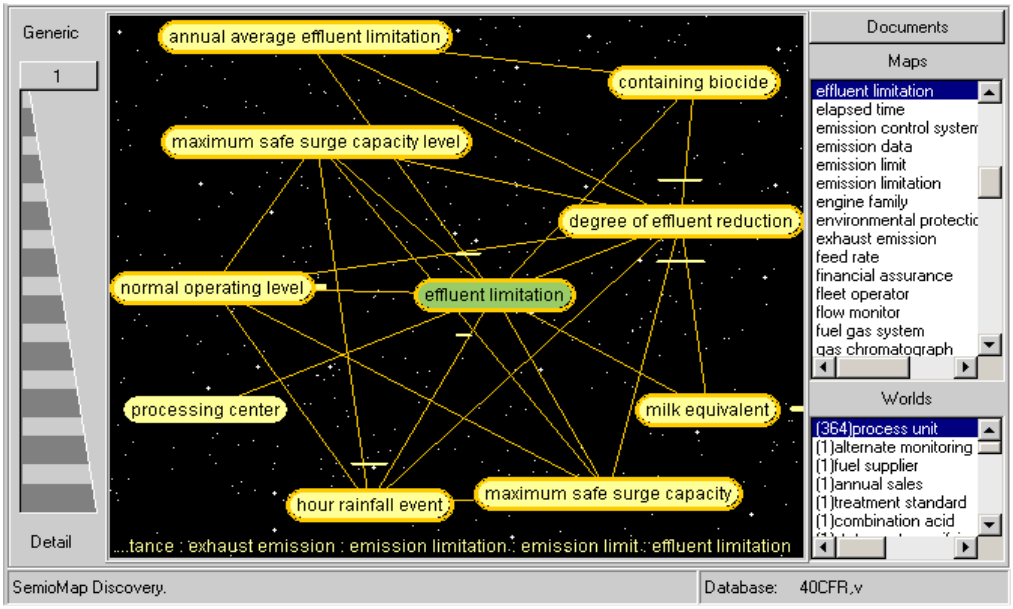


Figure 12: SemioMap Discovery, an information visualization tool from Semio Corporation, of San Mateo, California.

Each node in Figure 12 represents a single noun phrase from the corpus of text under review. In this case, the text is United States environmental regulations from 40 CFR (Code of Federal Regulations). Lines between two nodes indicate that a co-occurrence relationship exists between them. At level 1 (left hand side of image), each line represents a strong co-occurrence relationship. At level 10, each line indicates a weak co-occurrence relationship. Note that the phrase "effluent limitation" is highlighted in the central display, and that the phrase is also the name of the map in which it participates (see "Maps" window on right of image). Maps are named after their most interconnected noun phrase. By looking first at the listing of maps at the right of the interface, and then looking through each of those map displays individually, one can quickly learn—at a general level—what a given document collection is about.

Technology Assistance for Building Classification Structures – Semio Lexicon Builder

Semio has another tool which helps in learning what topics are discussed in a large document collection. Lexicon Builder is a tool which takes the noun phrases that Semio extracts and groups them according to the words they share in common. For example, from Figure 12 the noun phrases "effluent limitation," "degree of effluent reduction," and "annual average effluent limitation" might be grouped beneath the word "effluent." It is possible for phrases to appear in multiple groupings if they share one word with one set of phrases, and another word with a separate set of phrases. For example, "degree of effluent reduction" might also join the phrase "emission reduction analysis" under the heading of "reduction." Figure 13 shows some of the terms extracted from 40 CFR, grouped under the general term "bioaccumulation."

The Cartia/SPIRE technology, Semio's SemioMap Discovery, and Semio's Lexicon Builder offer capabilities that are relatively rare on the marketplace. Few companies offer technologies that provide an unmediated view of the topics in a collection of unstructured text. Not only do these technologies provide an interesting interface to the documents themselves, they offer an invaluable view into the conceptual topography of a document set without requiring the intermediation of search. This ability to overview a text collection facilitates the construction of suitable bottom-up or hybrid classification hierarchies.

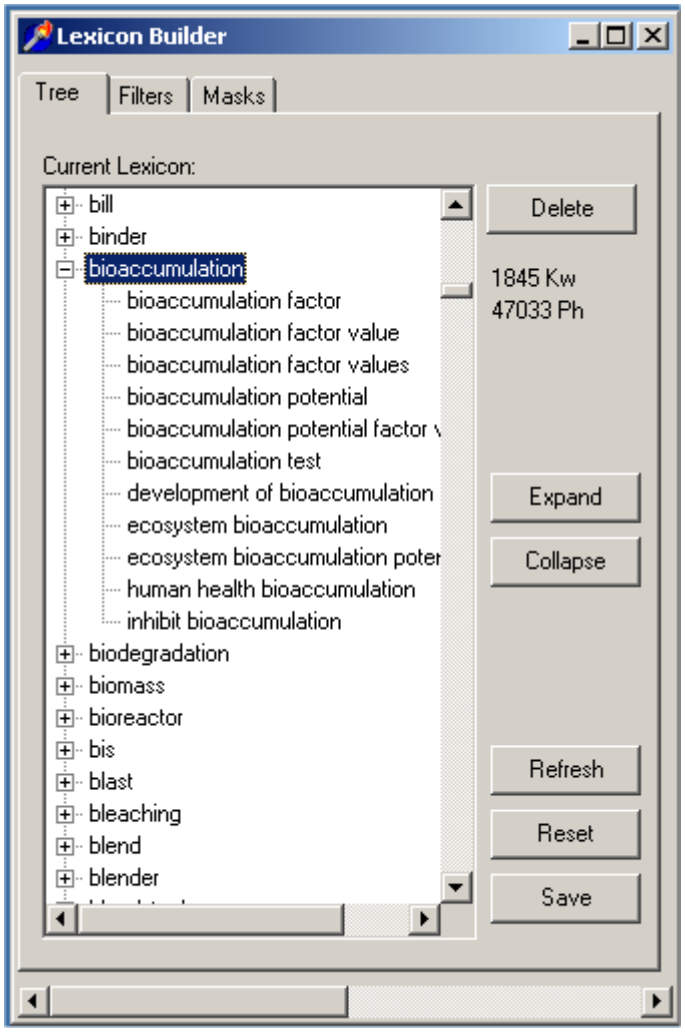


Figure 13: Semio Lexicon Builder, an information discovery tool from Semio Corporation of San Mateo, California.

The Population of a Classification Structure

With an already-built classification structure, the task remains to assign entities to positions within that hierarchy. Traditionally, classification has been a manual process. Today, a number of technology solutions exist for automating or partially-automating the act of categorization, yet there is still debate over whether humans do it better.

While human intelligence is critical to the development of a categorization structure, once such a structure has been built the inconsistency and labor-intense nature of human categorization makes

the case for a technology solution. On the other hand, the risk that a technology will over-assign entities to categories (ie, false positives) or under-assign them (failure to place an entity in a category to which it should belong) makes the case for at least some level of human oversight.

Manual Categorization – Yahoo! Inc.

In the mid 1990s, what would eventually become Yahoo Corporation began when its founders decided to categorize their list of bookmarks to interesting websites. As the entity grew, Filo and Yang continued using manual classification as a way to differentiate Yahoo in the marketplace. Maintaining this type of manual classification structure is labor-intensive. In Yahoo's first years of existence, fully 75% of its workforce held the job-title of "Surfer." A surfer at Yahoo would spend hours surfing the Internet, looking for websites that might need to be categorized in the surfer's specific niche of responsibility. Some focused on banking web sites. Others focused on gaming or travel. The result was a high-quality classification of sites that afforded visitors to Yahoo with highly-relevant search results as compared to competing search offerings at the time.

Figure 14: Home Page of Yahoo, Inc. as of Summer 2001.

In Figure 14, the Yahoo classification structure appears at the bottom-left of the home page. As of Summer 2001, Yahoo still maintains a staff of more than 100 full-time surfers, even after significant job cuts due to market conditions. The search functionality at Yahoo is now handled by Google.

Manual Categorization – The Open Directory Project

Like the Yahoo classification structure, the Open Directory Project (ODP) is a manual categorization of sites on the Internet (see Figure 15). The ODP is a non-commercial enterprise started by Netscape in the spirit of the Open Source community. The categorization work is conducted by the directory's nearly 40,000 volunteer editors worldwide.

Figure 15: The Open Directory Project.

Anyone can submit a site for inclusion in the open directory, but not all sites are accepted. Sites must first pass through a screening process and be approved by the editor for an appropriate category before the site is classified. To help standardize this process, the ODP publishes on its site

a set of guidelines to assist editors in the often-subjective task of classifying new sites, as well as a set of guidelines to advise the public on what types of sites would be good candidates for submission.[11] The Open Directory Project's populated classification structure is made available for licensing to anyone, for free, provided that they adhere to the ODP's Open-Source licensing agreement.[12] The Open Directory is currently in use by AOL, Lycos, and Google (see Figure 16), among others.

Both the Open Directory's and Yahoo's classification structure are impressive endeavors to apply human classification to web content on a large scale. Yet while both projects benefit from human intelligence in the classification process, as a consequence they are also vulnerable to the inherent weaknesses of human categorization. These weaknesses arise from the difficulty for humans to apply categorization rules in a consistent manner.

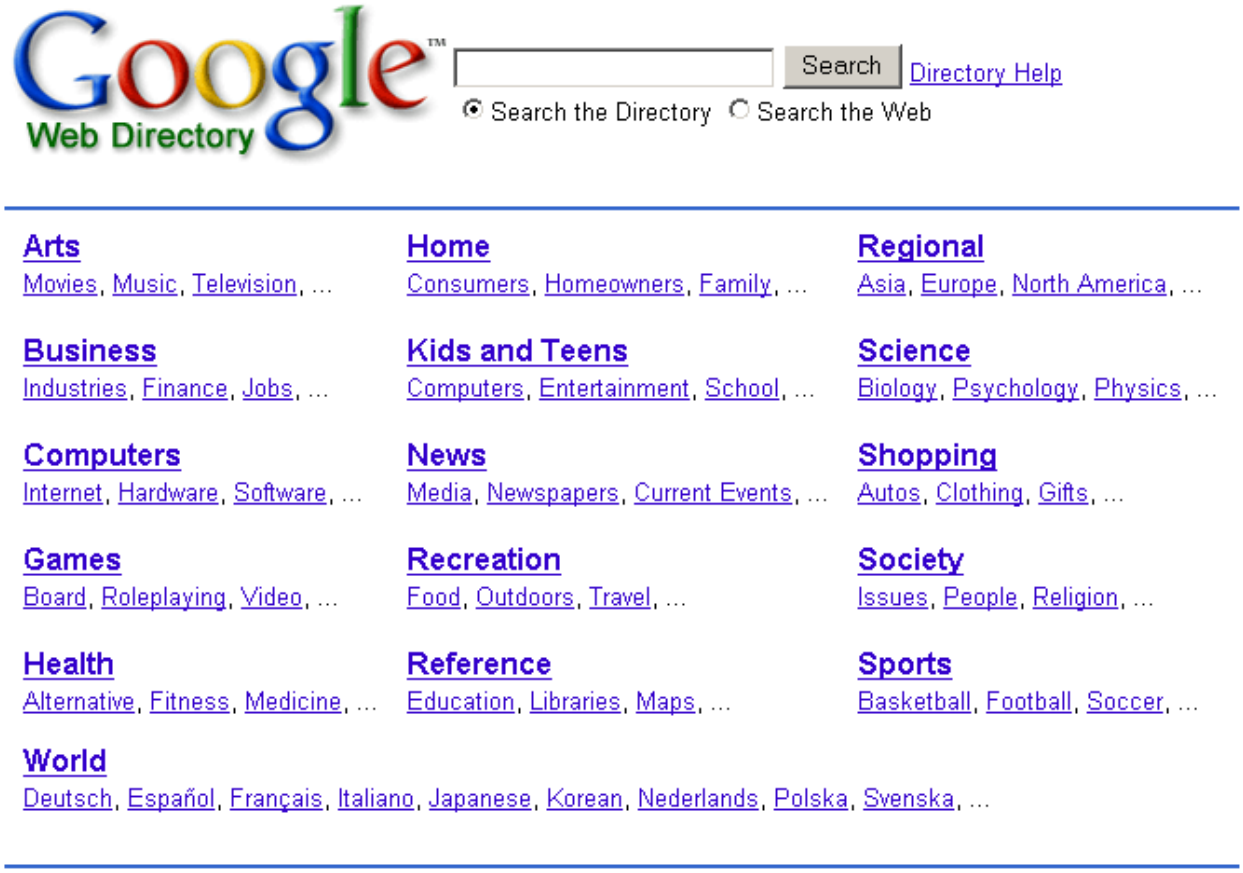


Figure 16: The Google Web Directory is the Open Directory.

This inconsistence comes mainly in two forms: A human classifier using constant classification rules can look at the same classificand at two points in time, yet categorize it differently each time. Likewise, two or more human classifiers using constant classification rules can simultaneously

[11] See http://dmoz.org/guidelines.html and http://dmoz.org/add.html
[12] See http://dmoz.org/license.html

categorize the same classificand in different ways. Both Yahoo and the ODP are aware of the importance of consistent classification, and both entities use categorization workflows and checks-and-balances to increase consistency. Unfortunately, in the end diachronic and synchronic inconsistency seem to be a part of having large teams of classifiers work on a joint effort. Consequently, one cannot have full confidence in manually created classification structures—even when they do add value to search.

Partially Automated Classification – Overview

The dominant approach to ensuring consistent application of classification rules is to take humans out of the picture to some degree. The extent to which human participation remains a part of the process varies from company to company. Usually, human participation takes the form of an oversight role in the categorization process. In these circumstances, an editor is often used to conduct spot-checks on documents that have been categorized by the system. When documents that have been misclassified are found, the editor can tweak the classification rules (in a rules-based system) or manually place the document in the appropriate category before rerunning the training sequence (in a supervised learning system).

This section reviews three products that are used to populate classification structures with documents in a partially-automatic fashion – Plumtree's Directory with Boolean Filters, Semio Tagger, and Interwoven Metatagger.

Partially Automated Classification – Plumtree Software

Plumtree Software's primary product is the Plumtree Corporate Portal. A corporate portal, also known as an Enterprise Information Portal (EIP), gives a company's employees, partners, customers, and suppliers a central access point "for the key information and services they need to do business with [that] organization."[13] While different Enterprise Information Portals may provide a different range of services, all good EIPs need to address the problem of facilitating access to unstructured text.

Plumtree's initial approach to categorizing unstructured text is particularly interesting in the extent to which it relied upon humans without actually having humans do the categorizing. In Plumtree's early product releases, humans were required to create Boolean keyword filters the computer would then use to populate an existing categorization hierarchy.

For example, for a category called "Mobile Communication," a human would need to create a filter for that category along the lines of:

> (wireless OR cellular OR "mobile phone" OR "mobile communication" OR "mobile telephone" OR (cell AND NOT (prison OR battery OR human)))

[13] See http://www.plumtree.com

Similar keyword filters were needed for every category in a Plumtree hierarchy (see Figure 17). With the filters in place, human involvement stops. Next, the computer begins the process of populating the categories by relentless application of the filters against the corpus of unstructured text. Any documents containing the right combination of keywords for a given category automatically populate that category.

Since computers replace humans for the actual process of categorizing, the Boolean Filter approach eliminates diachronic and synchronic variation in the application of categorization rules. Still, the creation of Boolean filters is an onerous task. The more specific the sub-category, the longer and more intricate the keyword filter must be. Lengthy and intricate filters make it difficult for a human to audit the categorization rules for overly-broad, overly-specific, or completely off-target filters. If a filter is missing an "AND NOT" parenthetical, that category may contain content that should not be there (overpopulation of a category). Likewise, if a filter is missing the full range of keywords that imply discussion of that category, then not all documents that should appear will appear (underpopulation of a category). Though computers enable rigorously consistent application of filtering rules, they do not lessen the need to develop rules that are both comprehensive and precise. As a result, while Boolean filters improve quality, they remain time-intensive to build and difficult to audit.

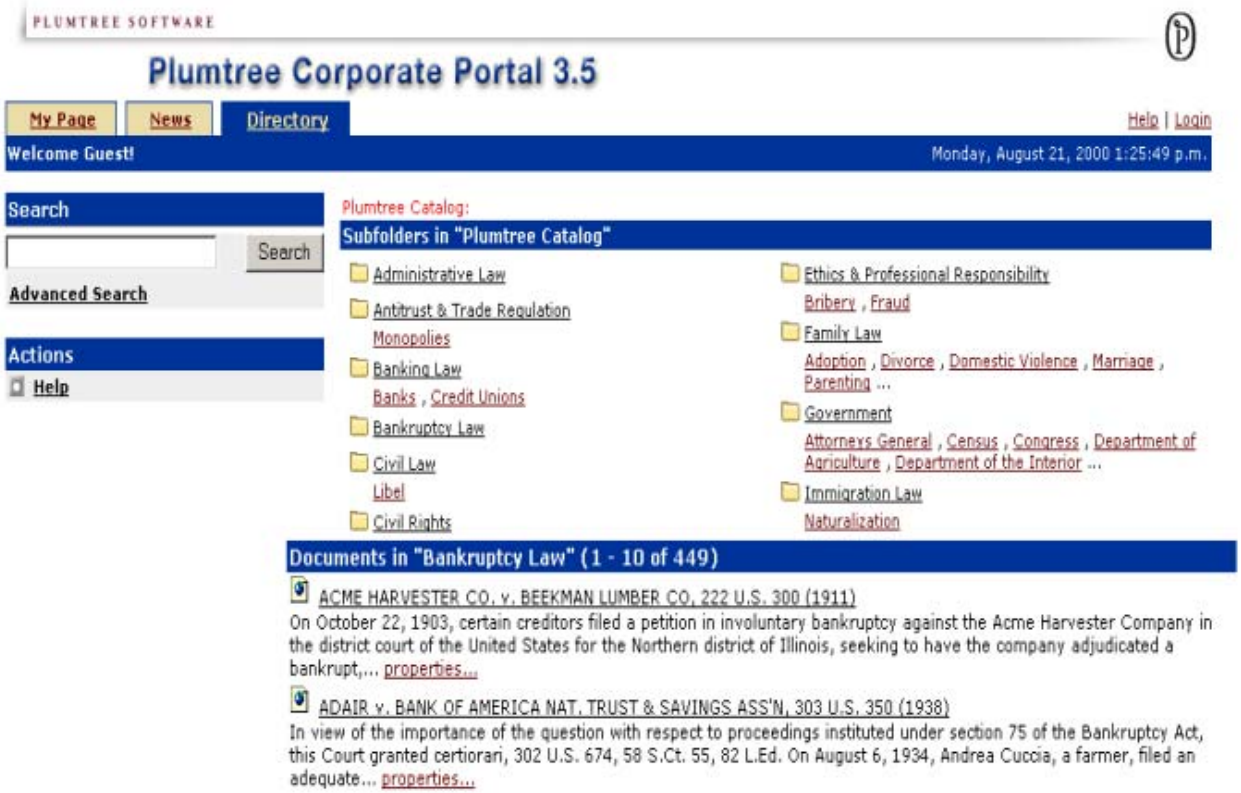


Figure 17: An early version of the Plumtree Portal interface to a classification hierarchy. In Plumtree's early releases, Boolean categorization filters would be set manually for each folder and subfolder in the hierarchy. Plumtree customers are now able to use third-party categorization engines to populate a Plumtree classification structure without the need for Boolean filters on each folder.

Partially Automated Classification – Semio Corporation

Semio Corporation is another company that relies upon humans to specify categorization rules for an existing categorization structure. Semio's product is called Semio Tagger. Instead of Boolean keyword filters, with Semio Tagger one creates latching rules and exclusion rules for each category. A latching rule will pull into a category pre-qualified noun phrases that include the word or phrase that makes up the rule. An exclusion rule does the reverse. Documents are categorized according to where the noun phrases they contain latch into the conceptual hierarchy: the more phrases a document contains that latch into a given category, the more that document is "about" that category. A benefit of this approach is that this process of creating latching and exclusion rules is straightforward. Part of the reason for this is that one need not worry about nesting appropriate "and," "and not," and "or" clauses into a filter. With Semio's latching rules, this effect is achieved by the combination of latches and exclusions at various levels of the hierarchy (see Figure 18).

```
!Pollution
      +pollution
      +pollutant
      -air pollution
      -air pollutant
      -water pollution
      -wastewater pollution
      -sea pollution
      !Air Pollution
            +air pollution
            +air pollutant
            -air pollution control
            -control of air pollution
            -control of ozone air pollution
            !Air Pollution Control
                  +air pollution control
                  +control of air pollution
                  +control of ozone air pollution
      !Water Pollution
            +water pollution
            +water pollutant
            +wastewater pollutant
            +sea pollution
            -water pollution control
            -control of water pollution
            !Water Pollution Control
                  +water pollution control
                  +control of water pollution
```

Figure 18: Latching and Exclusion rules from part of a Semio Category on the topic of "Pollution." The categories are capitalized and indicated by an exclamation mark (!). The latching and exclusion rules that implement the category appear beneath the category to which they apply. Latching rules are preceded by a plus sign (+), exclusion rules by a minus sign (-).

Perhaps more importantly, the process of auditing the quality of Semio's classification rules is easy: since Semio's rules categorize documents by means of noun phrases, one can audit a Semio

classification by skimming the lists of noun phrases that have latched. If the noun phrases in a given category all belong in that category, then the documents from which they came will be appropriately classified as belonging to that category as well.

Finally, the quality of a Semio-populated classification structure tends to be high because Semio's categorization rules run against a set of pre-qualified noun phrases that Semio has extracted from the corpus of text. To qualify for latching, a noun phrase must occur within a cycle of co-occurrence including at least two other noun phrases, each of which also has a co-occurrence relationship with the other. Semio uses phrases that participate in cycles of co-occurrence because they are less likely to signify tangential concepts than those that don't. The result is that a given document tends to be classified on the basis of concepts that are central to its content rather than tangential. In the next release, an alternative phrase extraction approach will obviate the need for this feature.

One of the strengths of Semio is its ability to deal with synonyms and acronyms. For example in Figure 18, under the category "Water Pollution Control," there are two latching rules. Although each is a distinct noun phrase, they signify essentially the same concept. It is possible within Semio to create a list of equivalencies between synonymous terms, such that Semio would treat "control of water pollution" as if it were the phrase "water pollution control." Thus, the latching rule "+control of water pollution" would be unnecessary. The single latching rule "+water pollution control" would pull to that category all phrases in which either variant participated. Similarly, one can create a list linking acronyms with their expanded forms, such that a latching rule for FEMA ("+fema") would yield all phrases in which either "FEMA" or "Federal Emergency Management Agency" participated.

One of the weaknesses of Semio is the polysemy problem, or when the same word conveys multiple, unrelated meanings. For example, the United States Agency for International Development is often reduced to the acronym "AID." The acronym AID could cause problems in implementing a category on non-governmental organizations. A latching rule "+aid" would pull in phrases in which the English word "aid" participates, but which are unrelated to the Agency for International Development. To avoid this, one could implement a latching rule "+agency for international development," which would not yield extraneous phrases. However, that rule would miss all the instances in which the agency is referred to as simply AID. Clearly, developing an ability to infer context for polysemous terms is an important classification problem that needs to be addressed. In the next release, Semio's lexical resources will support regular expressions as latching and exclusion rules. This feature should Semio deal with the polysemy problem.

Overall, Semio's categorization approach strikes a good balance between human oversight and the efficiencies of computerized classification. In fact, this is underscored by the fact that Plumtree has partnered with Semio to use Semio's categorization engine in some of its corporate portal implementations. Eli Lilly Corporation is a marquee example of this, running a Semio-enabled Plumtree Portal that serves 30,000 Lilly employees.

Partially Automated Classification – Interwoven Metatagger

Interwoven Inc.'s flagship product is called TeamSite. Interwoven claims that the "TeamSite software offers a flexible, scalable, standards-based platform for creating, managing and deploying…enterprise-class, business-critical Web content…." [14] Though TeamSite could be used for basic website design and publishing, that would be technology overkill; for all intents and purposes the product is an Enterprise Information Portal platform. And as is the case with any EIP vendor, Interwoven has to deal with streamlining the process of classifying content for the portal. As of Fall 2000, Interwoven uses intellectual property it acquired from Metacode, Inc. to enable classification within TeamSite. Interwoven has renamed the Metacode software—it is now called Metatagger—and it is an add-on feature to TeamSite (see Figure 19).

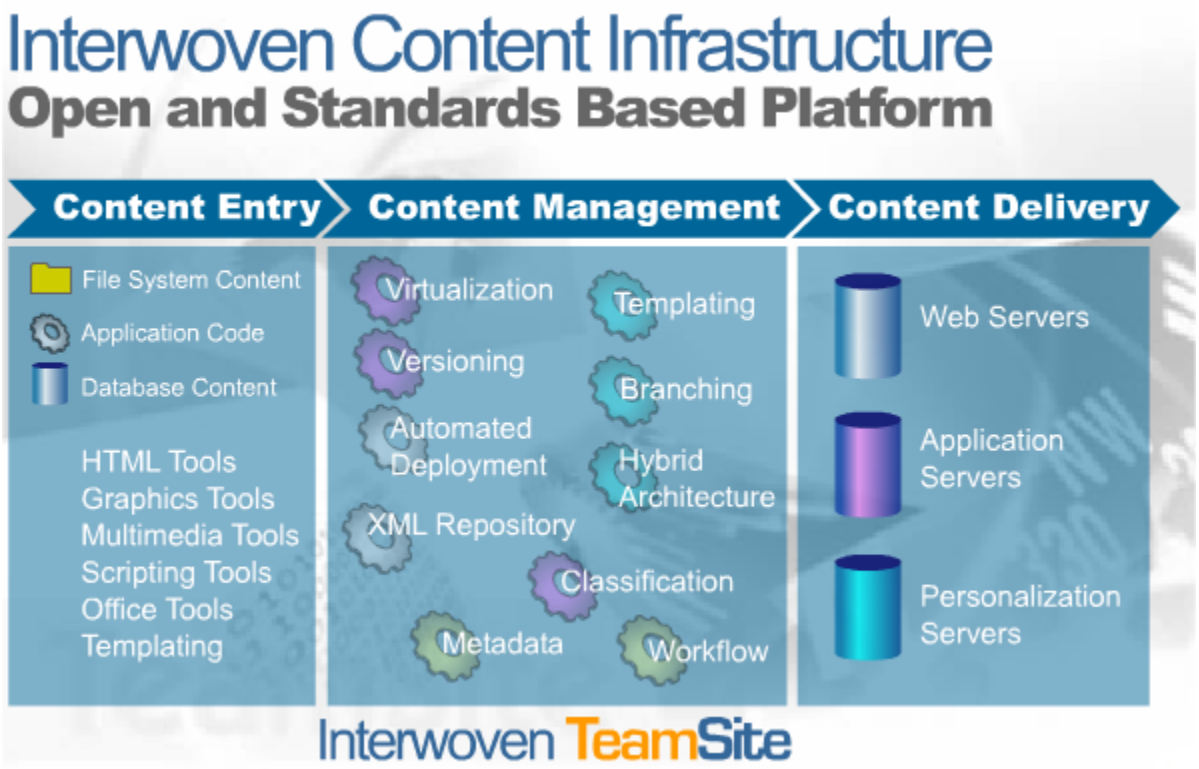


Figure 19: Chart of features available with Interwoven's TeamSite Corporate Portal product. Note the "Metadata" and "Classification" features.

Metatagger scans through unstructured text, comparing terms against human-generated text files that contained product catalogs, taxonomies, industry-standard controlled vocabularies, or custom vocabularies. Text gets classified according to where it latches into these classification hierarchies or lexicons. Like Semio, Metatagger can handle synonyms but can not deal well with polysemy.

The MetaTagger/TeamSite pairing allows a valuable level of human oversight on the classification of text. Behind the scenes, humans still develop the conceptual hierarchies, controlled vocabularies, and lexicons that MetaTagger will use to categorize text. However, since MetaTagger is part of a

[14] See http://www.interwoven.com

corporate portal, Interwoven has made it possible for some of the users of that portal to give editorial input on the first round of automatic classification.

For example, consider a subject matter expert (SME) who drafts a document and submits it for publication on a TeamSite-enabled corporate portal. MetaTagger will automatically scan the document and label it according to where its text latched into any number of controlled vocabularies/taxonomies. The TeamSite system will then show the subject matter expert how his document has been categorized. The SME will then be required to "sign-off" on this categorization or make changes to it according to his domain specific knowledge. Only after the SME has reviewed the automatically generated categorization tags will the document be accepted into the system.

While this feature is compelling, it goes against the idea of limiting the opportunities humans have to introduce inconsistency in the categorization process. Consequently, it appears that in an implementation of TeamSite with MetaTagger an organization should have a thorough review process rather than allowing SMEs to tweak MetaTagger's initial categorization results. If MetaTagger is classifying documents in ways that it shouldn't, a review process will enable the SMEs' feedback to be incorporated into the underlying controlled vocabularies and taxonomies. This is far preferable to having each SME modify a flawed default categorization in different ways. If properly implemented, this new opportunity for human review can improve the categorization process. Overall, Interwoven offers a powerful approach for streamlining the process of classifying unstructured text while maintaining human oversight on the classification process.

Largely Automated Classification – Overview

As with the manual and partially-automated approaches to categorization, the largely-automated approach assumes a pre-existing categorization structure—regardless of how that structure was created. However, in the largely-automated approach the categorization engine requires that the structure already be populated with a representative set of training documents for each category. These categorization engines scan the text of all the documents in the training set, modeling the documents to calibrate features for each category and subcategory. Once a system like this has been trained administrators can use it to process large volumes of documents. As new documents are processed, the system will compare the features of those documents with those of the category models and assign documents to categories accordingly.

There exists a wide range of algorithm-families an organization can use to implement this form of supervised-learning-based categorization. Among others, these families include Bayesian inference, neural networks, decision trees, k-nearest neighbor techniques, maximum entropy models, vector space models, and hidden Markov models. Each can be used for this problem—with varying degrees of success. Some of the companies using supervised learning as an approach to text categorization state clearly which approach(es) they employ. Other companies speak only in general terms without revealing any information about their specific implementation method.

This section reviews three products that are used to categorize unstructured text in a largely automatic fashion: Inxight Categorizer, Autonomy Categorizer, and the Hummingbird EIP's classification engine.

Largely Automated Classification – Inxight Software

Inxight Software of Santa Clara, California has a number of "knowledge management" related products, one of which is the Inxight Categorizer. Categorizer is a categorization engine that Inxight sells to companies that want to "[add] horsepower to [their] knowledge management portal." Inxight does not sell a portal solution themselves, preferring to be a vendor of components that facilitate information retrieval in existing portal or intranet environments.

The company does not identify which type of approach they use to classify text, although from the company's collateral it seems that their method uses statistical inference, most likely a Bayesian algorithm: "new documents are compared with a large collection of [already classified] documents (the training set). …[Then] the Categorizer selects similar documents from the training set and infers the probable [classification] for the new document from these examples."[15]

Inxight is aware of the risk of completely turning over to computers the process of populating a classification structure with text. As a result, while Categorizer relies upon the training set of documents to infer categorization rules for new documents, an administrator can set a ranking function so Categorizer will route certain documents to a human for review. Essentially, any document that has a likelihood of belonging in a given category that is below the ranking threshold will not get classified automatically. Once such documents have been manually reviewed and classified, an administrator can then add them to the training set—enabling the Inxight Categorizer to improve its accuracy for subsequent rounds of classification.

With a statistical inference-based classification system such as Inxight Categorizer, the advantage of speed comes at the possible sacrifice of accuracy and coverage. Without aggressive auditing, it is possible that the statistical models which define the categories could be flawed. The result can be either overpopulation or underpopulation of the categories. Yet with aggressive auditing one loses the speed of a largely-automated system. For some, an inference-based system like Inxight Categorizer can be a valuable, time-saving solution. For others, the classification risks inherent to a supervised-learning system will not be worth the savings in speed.

Largely Automated Classification – Autonomy Corporation

Like Inxight, Autonomy Corporation of San Francisco, California sells a suite of "knowledge management" products. One of these products is named the Autonomy Categorizer. In assigning text to categories, Autonomy's Categorizer relies upon a combination of approaches, including Bayesian inference as well as maximum entropy.

As with Inxight, Autonomy requires first that a training set of documents be categorized into an existing classification hierarchy. Subsequently, new documents are classified according to how their

[15] See Inxight Categorizer white paper at http://www.inxight.com/pdfs/whitepapers/km_categorizer.pdf

features match the models for each category, as determined by the training set. Administrators have the capacity to re-train the system by manually-classifying additional documents into the training set.

Autonomy's Categorizer has been used to power existing portals at organizations such as Brio Technology, Novartis, and FranceTelecom. Autonomy's own product line also includes a portal offering, called "Portal-in-a-Box." It enables customers to implement an Autonomy-powered portal using Autonomy's categorization engine to facilitate access to unstructured text.

Autonomy maintains a library of 700 pre-built classification hierarchies on different topics to serve as starting points for classifying text at client sites. It is unclear whether Autonomy also maintains a library of pre-classified training texts to accelerate the customization of each hierarchy's feature models to specific clients. Such a library would be very useful in streamlining the application of Categorizer to new document sets.

Autonomy includes a Windows Explorer-style interface for administrators to use in modifying the relationships between categories (see Figure 20). Of note is the fact that the rules by which documents will be pulled into one category or another are not apparent – all that is shown are parent and child categories. This is because the rules are probability formulae developed from the training set, as opposed to lexical latching rules as with Semio (see Figure 18) or MetaTagger. The category models in lexicon-based systems are directly accessible precisely because they are non-formulaic. While inference-based systems can be highly effective, it is difficult to know exactly why a given document has populated a given category.

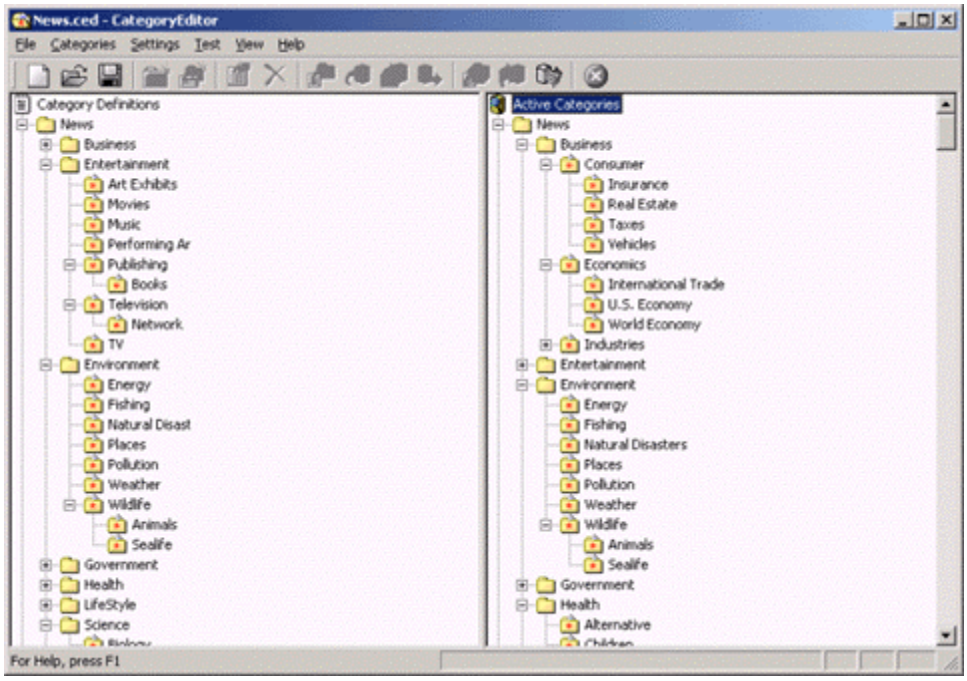


Figure 20: Autonomy Corporation's Category Editor.

Largely Automated Classification – Hummingbird

Hummingbird, LTD. is another company with a largely-automated text categorization solution. The Hummingbird Enterprise Information Portal is sold with their categorization engine built-in.

Hummingbird uses a different approach than either Inxight or Autonomy: Neural Networks. Like inference-based systems, neural nets require pre-categorized training documents in order to "learn" how to populate a categorization hierarchy with new documents.

Neural nets are so called because they attempt to mimic a portion of the functionality of the human brain. Computer science professor Daphne Koller, of Stanford University, explains:

> "How does the brain work? The brain is composed of billions of neurons (about 10^{11}). Each one looks like this:
>
> The dendrites act like input wires. They get chemical messages from other neurons (via synapses), which raise or lower the electrical potential of the cell. When the electrical potential passes a certain threshold, the neuron "fires", sending a chemical signal on its output wire, the axon.
>
> The way in which one neuron can affect another depends on the type and strength of the connection between them. The brain learns by

> modifying the strength of these connections between the different neurons, in response to experience (mostly sensory inputs). When the brain gets some input, some neurons fire, causing certain connections to be strengthened. As experiences pile up, some connections develop and others die, representing our long-term memory and experience.
>
> Pieces of software can be written to act as a single neuron, with a "bunch of input wires, each with its own weights." The software "takes its inputs, performs a very simple computation, and outputs the result on its output wire." This can be combined to form a two-layer process, whereby the outputs from multiple software neurons can be treated collectively as an input by another single software neuron.

Professor Koller adds that the implications for classification are such that, "with enough hidden units, a two-layer neural network can approximate any decision boundary arbitrarily well…."

One problem with neural nets as applied to text classification is that the reasoning used to classify individual documents is not available for review. If an administrator wants to revise the way a neural net is classifying a given document, it requires re-training of the network—and even then the retraining may not yield the precise categorization wanted by the administrator.

A large segment of the text classification marketplace is composed of companies using some form of supervised machine learning. Inxight, Autonomy, and Hummingbird are only part of the market, but they are representative of the three major business models in text classification: selling a categorization engine; selling a categorization engine, but also selling a portal that uses that engine; and selling an enterprise information portal that includes a built-in categorization engine. Semio falls into the first group. Interwoven falls into the third. Other text classification companies likewise will fall into one of these three groups.

That said, even within one of these groups companies can choose from a range of algorithms. With the largely-automated set of companies, the most common at this point seem to be Bayesian inference and neural nets, but they all use pre-categorized training documents to prime the categorization pump. Unfortunately, although using a training set of text can allow these largely-automated systems to save time later in the classification process, the quality of the classification can suffer because of inherent flaws at the source: the training set.

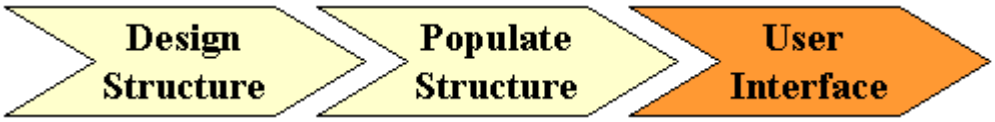
The largely-automated systems that require training sets rely upon the categorization relationships in the training data in order to classify new documents. It is unclear whether this reliance is well-founded: if – as is often the case – the training documents are populated into a classification hierarchy by hand, then all the weaknesses of manual classification translate directly to weaknesses in the training hierarchy itself. If the documents in a training hierarchy derive their positions in that hierarchy from a process that is un-comprehensive and inconsistent, then neither a Bayesian infererence engine nor a neural network-based system, nor other such systems will likely be able to compensate. More likely, such supervised learning systems will simply apply the training set's

questionable classification relationships to new documents on a broader scale than could a purely manual endeavor. And while in this case diachronic inconsistency might be controlled—a computer is consistently applying the rules, after all—any inconsistency at the time of training acts as an original flaw that casts subsequent categorization relationships into question.

In the end, to choose among the different software approaches an organization must first determine what sort of an error-rate is acceptable—both in terms of wrongly classified documents as well as correctly, but incompletely classified texts. If a company is only classifying internal marketing documents, a certain error-rate may be acceptable so long as the classification process is fast and efficient. In such a case, Bayesian inference or neural networks may be appropriate. If an organization is classifying documents that someday may be subject to disclosure under the Freedom of Information Act, an easily-auditable lexicon-based approach like Semio or Interwoven might be more advisable.

For organizations committed to Bayesian inference or neural nets, perhaps a combination of technologies provides a solution: a lexicon-based technology like Semio or Interwoven could be used for populating a structure with training texts. Then an organization can turn to any of the largely-automated systems to take it from there, secure in the knowledge that the initial training relationships were not manually created.

User Interfaces to Populated Classification Structures and Classified Text



After a corpus of unstructured text has been assigned to positions within one or more categorization structures, it is no longer appropriate to refer to the text as "unstructured." The act of populating a classification structure with text transforms that text from unstructured to structured. The task remains to make this newly-structured text available to searchers. In the marketplace there exists a range of user-interface approaches, from non-linear visualization interfaces to standard hierarchies.

Non-Linear Visualization Interfaces – Antarti.ca

One of the more interesting—though not necessarily always useful—approaches to accessing a populated classification structure is that of an information visualization company called Antarcti.ca Systems.[16] They have licensed the populated hierarchy of the Open Directory Project and used their Visual Net software to create a 2D and 3D cartographic interface to the directory.[17] Figures 21 through 25 represent a full drill-down to the US EPA's Superfund website in the Open Directory using the Antarcti.ca interface.

[16] See http://antarcti.ca
[17] The Antarcti.ca Interface to the Open Directory is at http://www.map.net

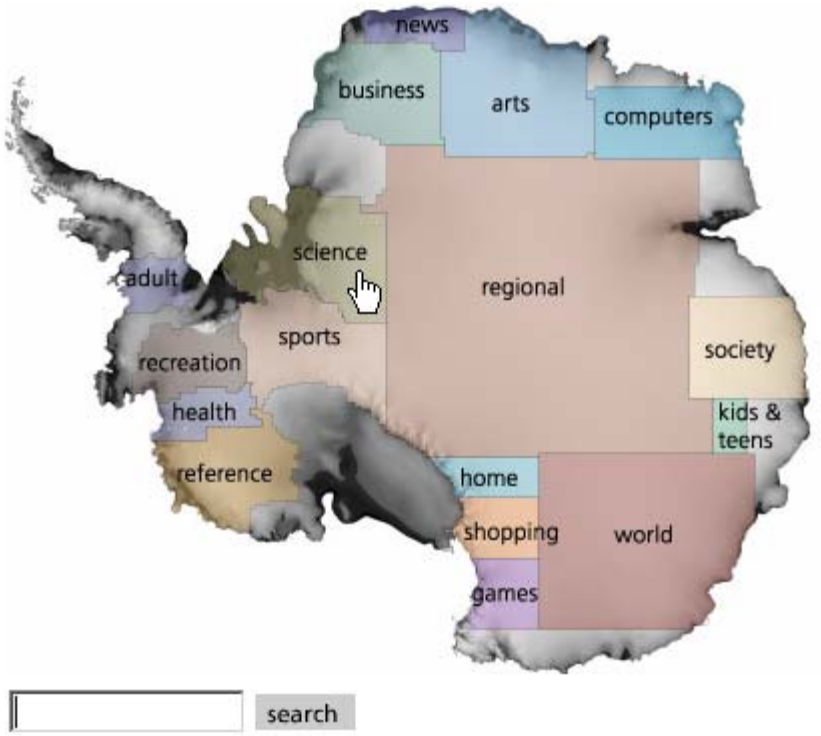


Figure 21: Antarcti.ca Visual Net interface to the top level of the Open Directory Project's categorization structure. Mouse icon indicates a click into the category "Science."

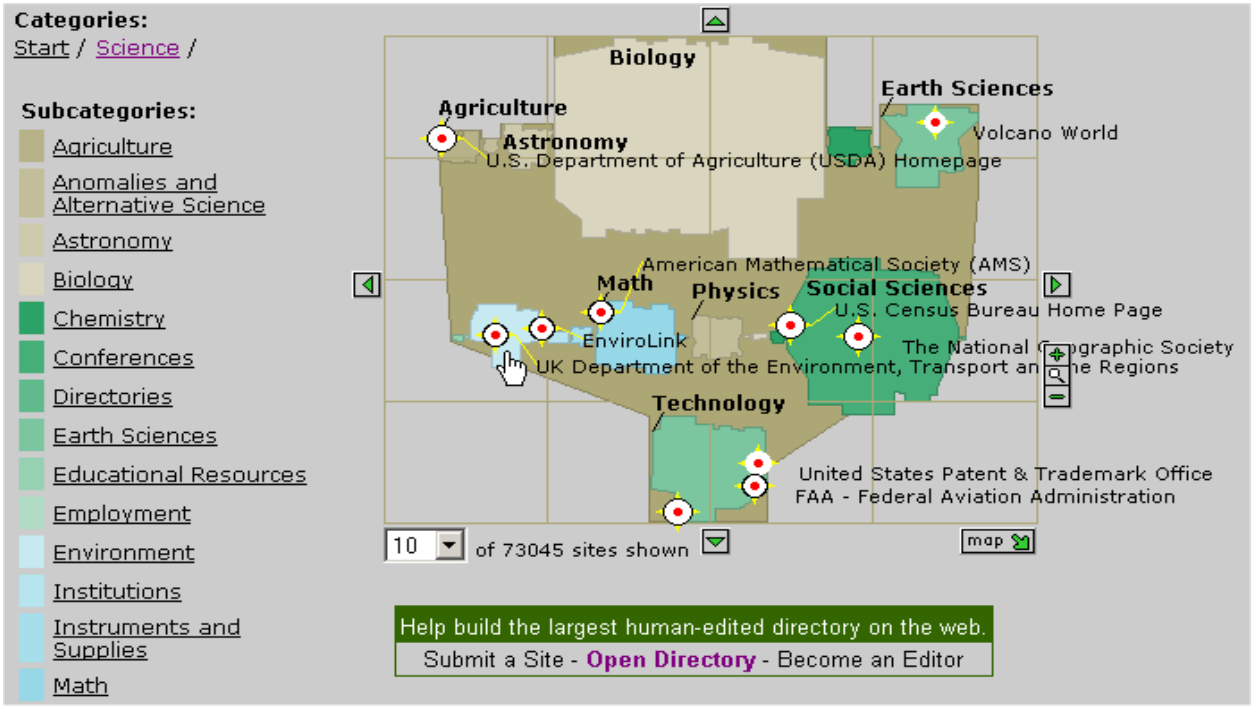


Figure 22: The category "Science." Mouse icon indicates a click into the subcategory "Environment."

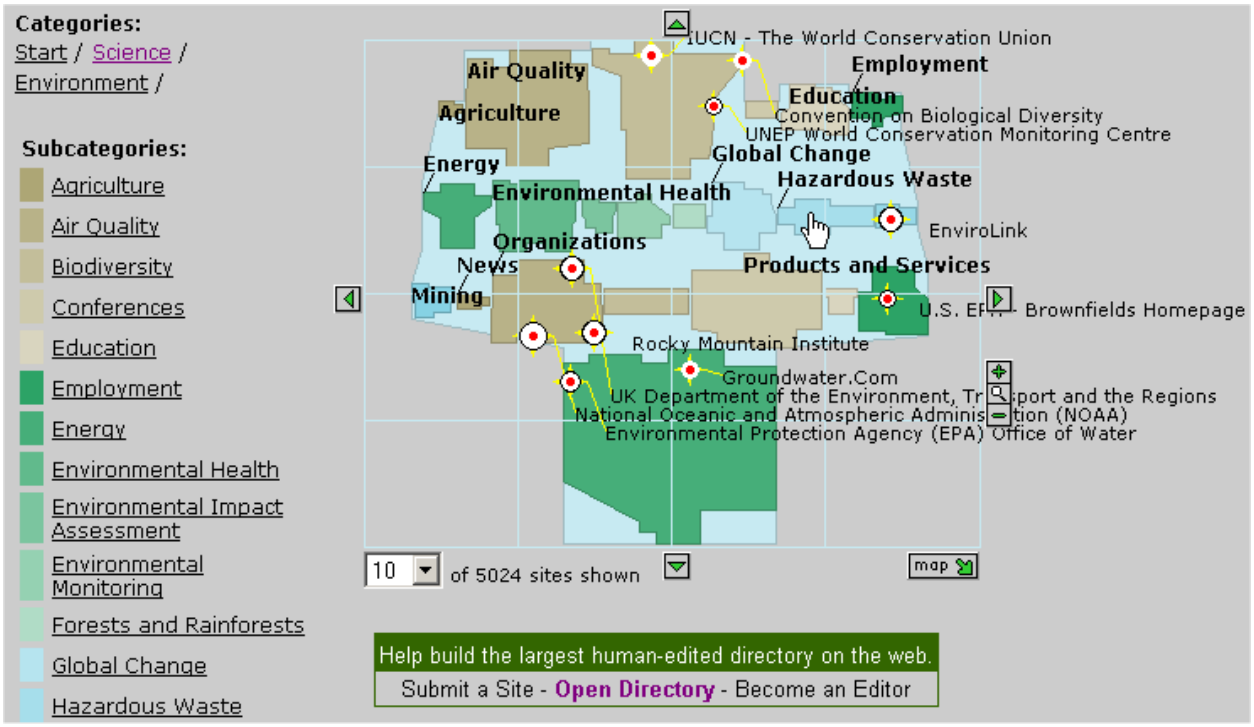


Figure 23: The subcategory "Environment." Mouse icon indicates a click into the sub-subcategory "Hazardous Waste."

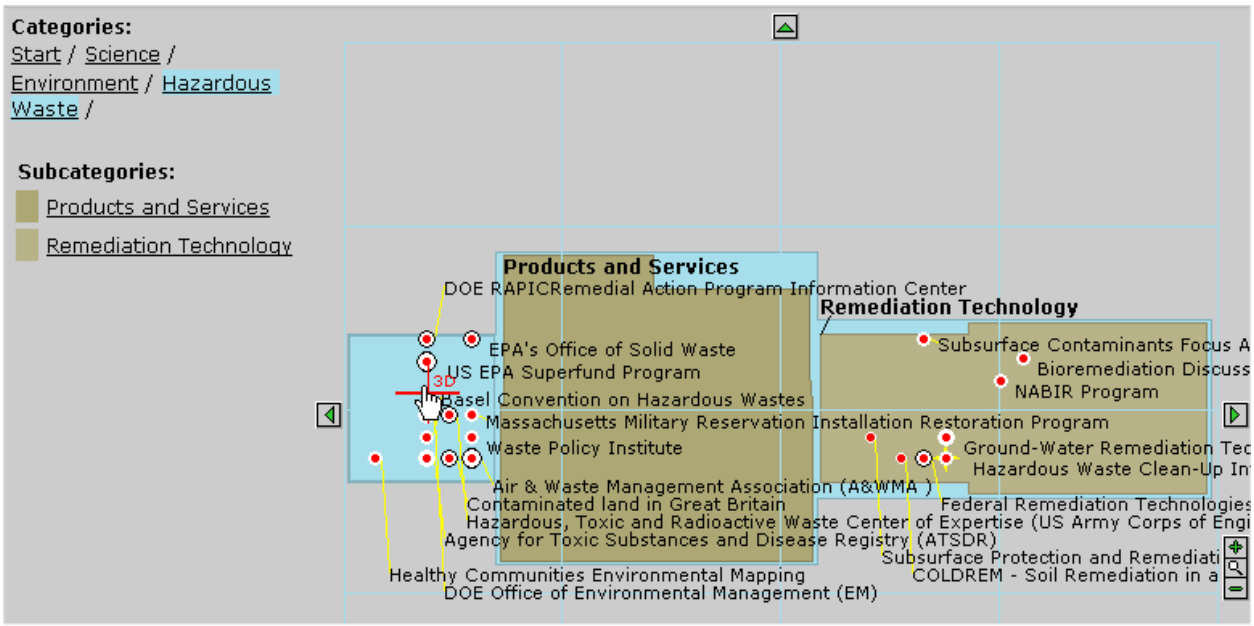


Figure 24: The sub-subcategory "Hazardous Waste." Red circles indicate websites. The white bands indicate the relative quantity of links to that site. The black bands indicate the relative quantity of links from that site to other sites. One could go directly to any site from the 2-dimensional version of this interface by clicking on the appropriate red circle. In this case, the mouse icon indicates a click through to the 3-dimensional version of the interface. Note that the mouse is just below the US EPA Superfund Program site.

Figure 25: Visual Net's 3-Dimensional interface to the populated hierarchy of the Open Directory Project. Larger buildings represent larger sites. The mouse icon is on the building representing the US EPA Superfund Program's site. Mousing over a site brings up a site profile at left. Referring back to Figure 24, note the position of the EPA's Office of Solid Waste, above and to the right of the Superfund site. In this interface, the US EPA's Office of Solid Waste is represented by the large building at the right side of the screen.

Among the valuable features of this interface to the Open Directory is the opportunity to use surface area to gauge the relative number of sites that inhabit a given category. In Figure 21, for example, one can see immediately that the "Regional" category contains roughly five times as many websites as does the "Science" category. Yet, while Antarcti.ca's software provides visually arresting images and a fun interface, it may not prove as useful as a traditional UI to a categorization hierarchy. The 3-dimensional interface can be slow and sometimes is difficult to navigate. For those just looking around, such drawbacks may not seem too significant. However, when one is searching for information under time constraints, it may be preferable to have a faster interface even if it is less visually appealing.

The 2-dimensional interface provides a more efficient experience than the 3D version. At first, it seems to obscure more than it reveals, but over time one grows more accustomed to the navigation and the look-and-feel. One can use the white and black bands around the website symbols to see the relationship between inbound and outbound links to a site. Presumably, a site with many inbound links is more worth visiting than one with few such links.

Google already incorporates this type of information when it calculates a site's relevance to a query, and traversing a list of highly-relevant search results from Google may be more appealing to time-

conscious users than combing through site names spread across the landscape of a Visual Net category, as in Figure 24. To be sure, comparing a list of Google's search results to Visual Net's interface to the Open Directory is not entirely fair: The one is generated in response to a query, the other is a browsable hierarchy. That said, even the traditional browsable interfaces chosen by Google for its Web Directory and by the Open Directory itself (Figures 16 and 15, respectively) are more clear and easy to use for a first-time user than is the Visual Net interface. It remains to be seen whether this is due to lack of familiarity with the UI (witness the relative scarcity of non-traditional interfaces in high-profile installations) or whether it is inherent to the interface itself.

Non-Linear Visualization Interfaces – Inxight Software

In addition to the Inxight Categorizer, Inxight Software offers an interface product called Star Tree Studio. With Star Tree Studio, one can create Star Trees, like the one in Figure 26. A Star Tree administrator can point the Star Tree Studio at a website, crawl all the links on the site, and publish what amounts to a map of that site. Alternatively, an administrator can produce a site map manually, creating child nodes in the Star Tree and typing in the URL for the appropriate page. Interwoven is one company that has used Star Tree Studio to create their site map. Others include Porsche, BestBuy.com, and Cigna Healthcare.[18]

Though the implementations on the Inxight website utilize Star Tree Studio as a site-map builder, Star Tree Studio can be used to build an interface to hierarchies like the Open Directory. While no one seems to have applied Star Tree Studio to the Open Directory yet, Figure 27 is an application of the product to a concept hierarchy developed with the aid of Semio's Lexicon Builder (Figure 13) and the Semio categorization engine. Figure 18 shows a portion of this concept hierarchy dealing with the topic of Pollution. In Figure 27, this "Pollution" category can be seen at the far left, above the "Waste Taxonomy Root Node."

The Star Tree interface takes some getting-used-to, but its advantages soon become clear: not only are you able to see down to multiple levels of subcategories within one branch of the hierarchy, even after you have drilled down in one area you are still able to see the categories and subcategories of other branches of the tree. With the Star Tree interface, it is easy to remember where you are in the overall hierarchy. Moreover, it is easy to jump from deep within one branch directly to a subcategory in another branch without backtracking to a common parent node. While the Star Tree interface may not merit use as a primary interface to a concept hierarchy, it contains enough valuable features that an organization would be well-advised to consider using it alongside a more traditional Yahoo! style structure.

One weakness of the Star Tree interface is that it does not allow for lateral relationships. That is, it does not allow the creation of links between the leaves at the bottom of one branch and the leaves at the bottom of another branch, even when there may be a real conceptual connection between them. Still, lateral linkages are more web-like than tree-like, and Star Tree Studio does not purport to create anything other than (Star) trees.

[18] A list of Inxight Star Tree implementations is at: http://www.inxight.com/products/star_tree/demos.html

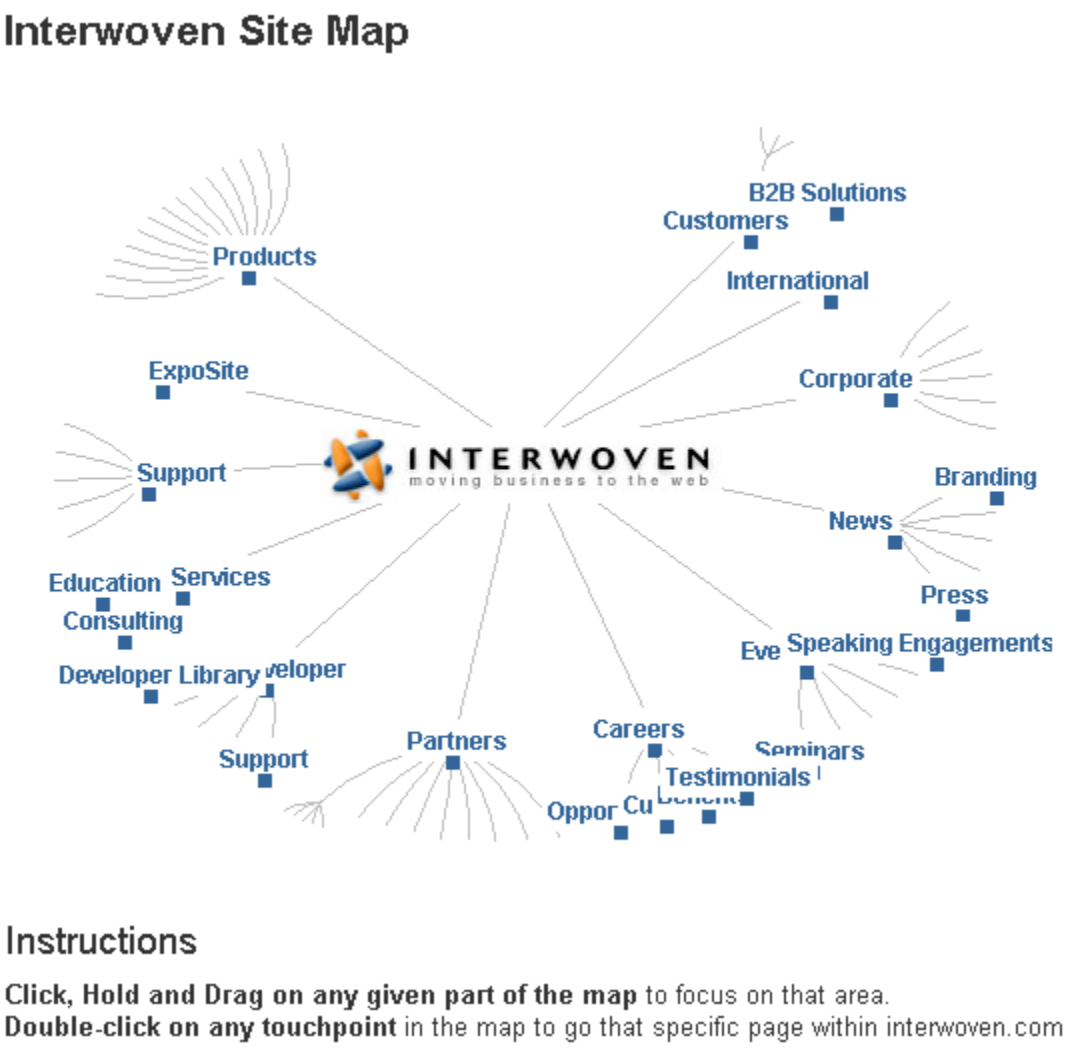


Figure 26: Inxight Star Tree site map of Interwoven Inc.'s website.

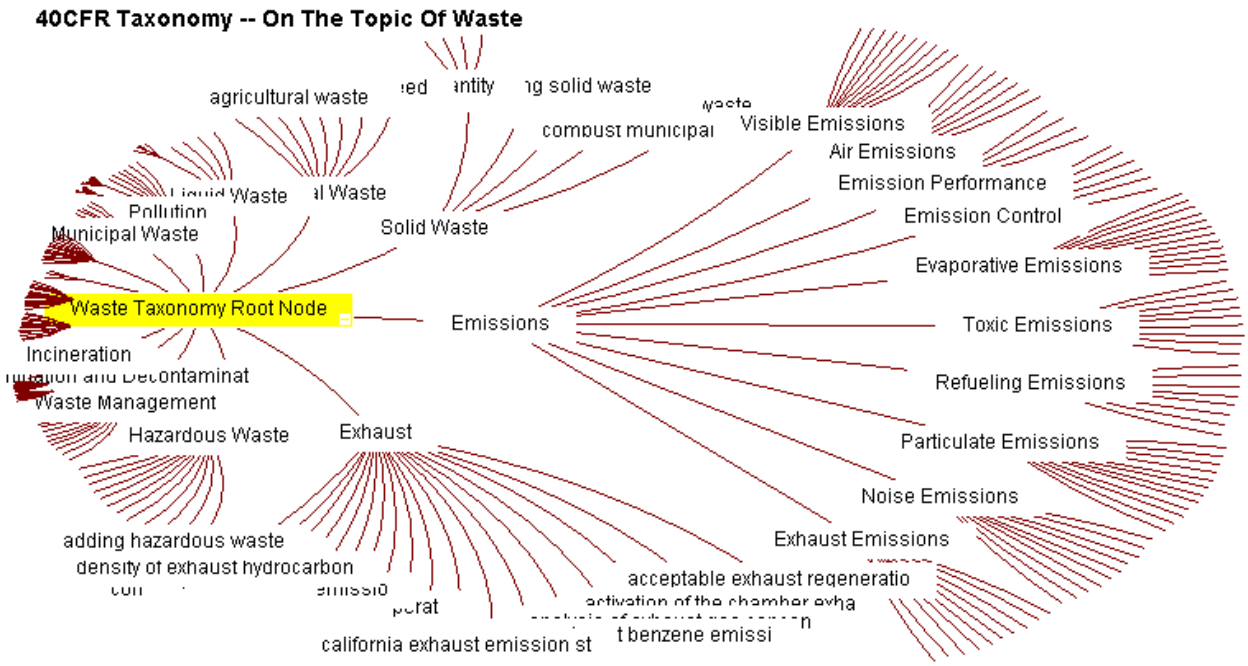


Figure 27: Inxight Star Tree interface to a partial categorization hierarchy on the topic of "Waste."

Non-Linear Visualization Interfaces – The Brain

The Brain Technologies Corporation, of Santa Monica, CA, has an interface tool that allows for lateral connections of the sort that are not possible in Star Tree Studio. As an example, Figure 28 shows The Brain highlighting both the parent-child relationship between "Artificial Intelligence" and "Automatic Speech Recognition," but also the lateral relationships between the topic of "Automatic Speech Recognition" and the terms "Computer Language" and "AARON."

This is possible because the architecture of The Brain does not assume a strict hierarchy of superordinate and subordinate relationships between entities, as the Star Tree interface does. Rather, with The Brain, the existence of a link between two entities can convey just the idea that a relationship exists—without specifying whether that relationship is hierarchical.

The Brain has been used on a number of sites besides KurzweilAI, including the website for the World Economic Forum.[19] More relevant to the present discussion is the fact that The Brain, like Google and Antarcti.ca, has licensed the Open Directory Project's classification hierarchy. Figures 29 through 32 repeat with The Brain user interface the same drill-down to the US EPA's Superfund website that was shown in Figures 21 through 25.

[19] See http://www.weforum.org/knowledgeforum.nsf/Main?OpenFrameset

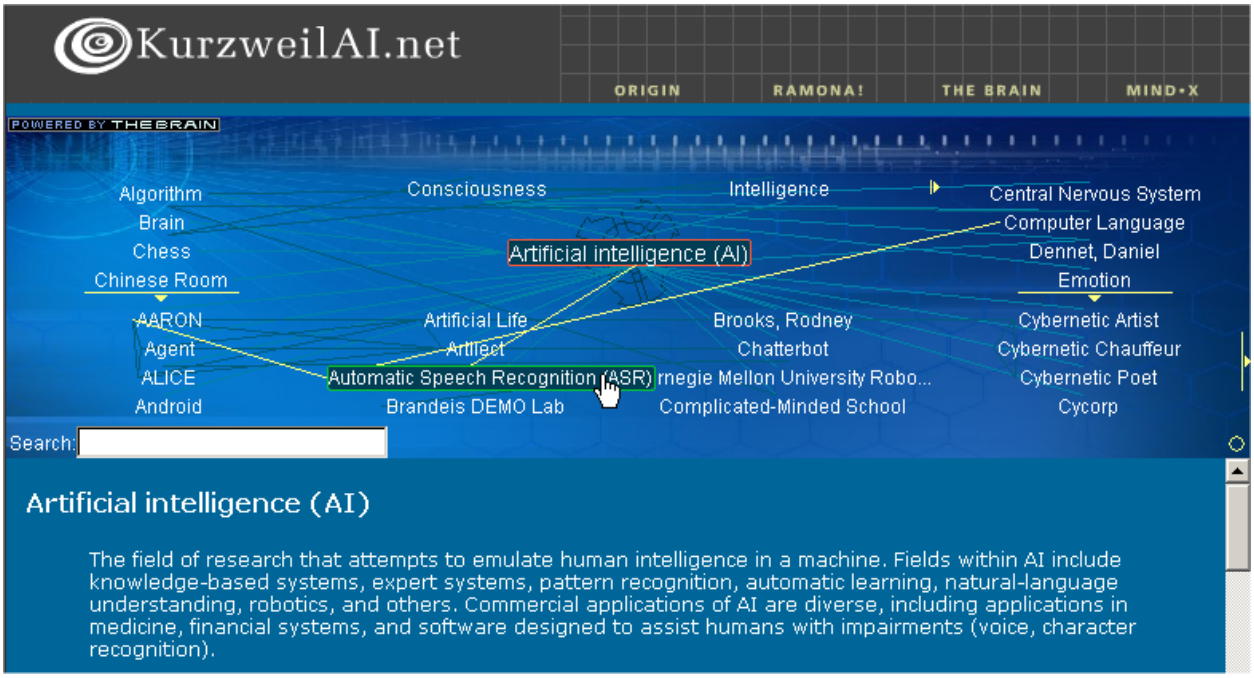


Figure 28: An implementation of The Brain at the website of technologist Ray Kurzweil. Source: http://www.kurzweilai.net.

Figure 29: The Brain's interface to the top level of the Open Directory Project's categorization structure. Mouse icon indicates a click into the category "Science."

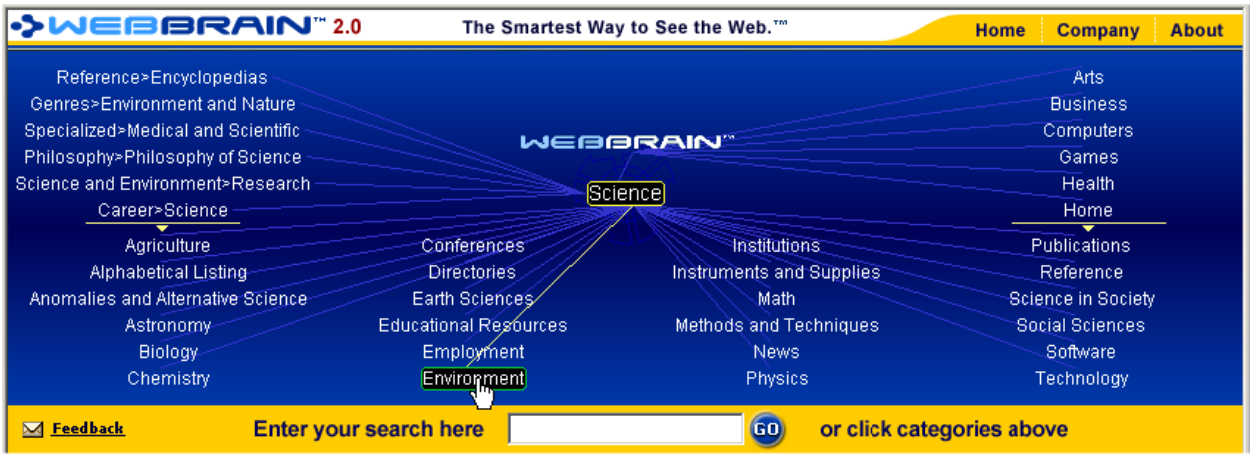


Figure 30: The category "Science." Mouse icon indicates a click into the subcategory "Environment."

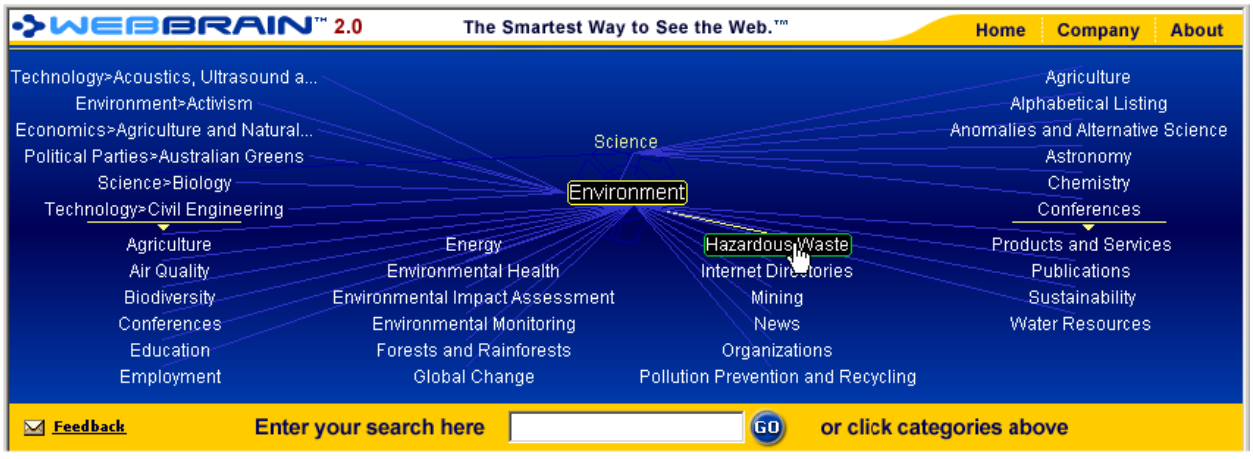


Figure 31: The subcategory "Environment." Mouse icon indicates a click into the sub-subcategory of "Hazardous Waste."

Figure 32: The sub-subcategory "Hazardous Waste." Item 22 is a link to the EPA's Superfund Home Page. This is identical to the link represented by the building in the center of Figure 25, as well as to the red circle representing the US EPA Superfund website in Figure 24.

As compared to Antarcti.ca's Visual Net software, The Brain is significantly faster-loading, faster to navigate, and more visually organized. Perhaps this is because The Brain does not stray quite as far from the traditional Yahoo style interface to hierarchies seen in Figure 14.

Traditional User Interfaces to Classification Hierarchies – Autonomy and Semio

Most interfaces to populated classification hierarchies do not stray much, if at all, from the traditional user interface model of Yahoo's web directory. Google's Web Directory and the Open Directory Project itself are virtual clones of the Yahoo directory interface (Figures 15 and 16). Plumtree Corporation makes an allusion to the Windows Explorer file system with its folder icons (Figure 17), but the underlying structure is still traditional. Autonomy's Portal-in-a-Box includes a pared down interface for populated classification hierarchies that is in the vein of Yahoo. Semio Corporation's user interface for classification hierarchies also uses this standard directory-style look-and-feel. Figures 33 and 34 represent a drill-down to a document record in the Autonomy interface. Figures 35 through 37 represent a drill-down to a document record in the Semio interface.

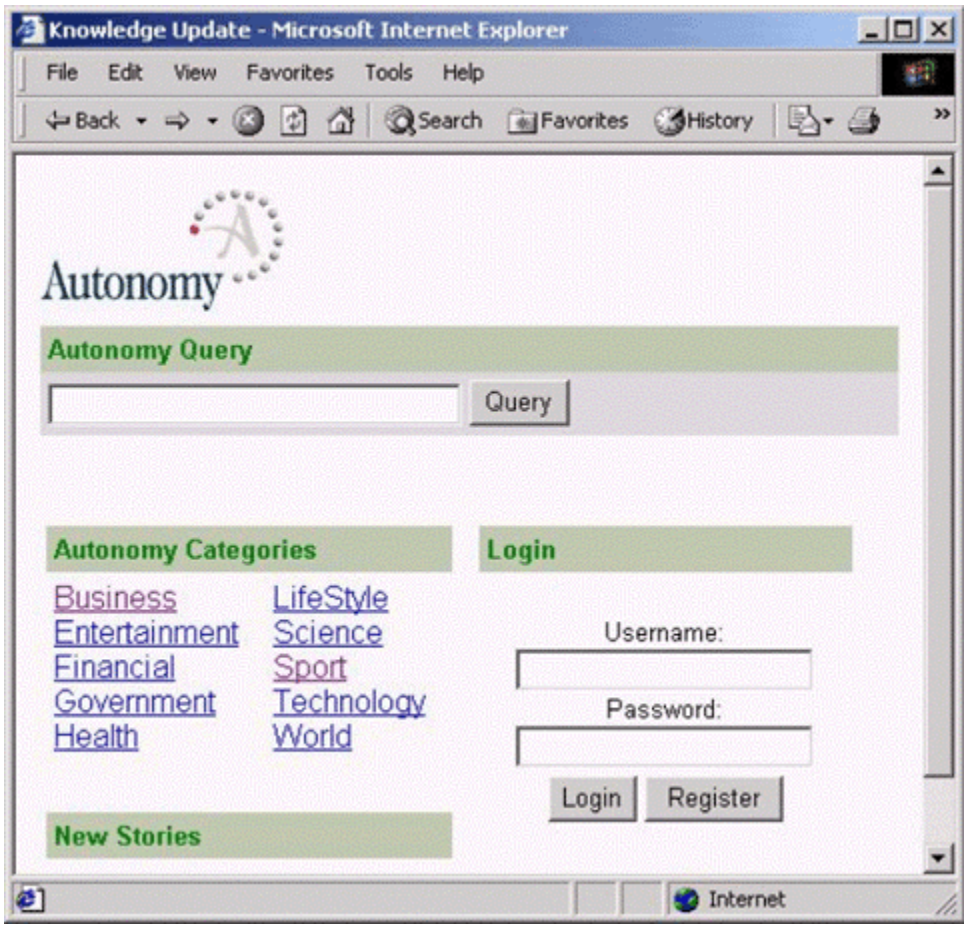


Figure 33: A categorization hierarchy viewed through the Autonomy user interface.

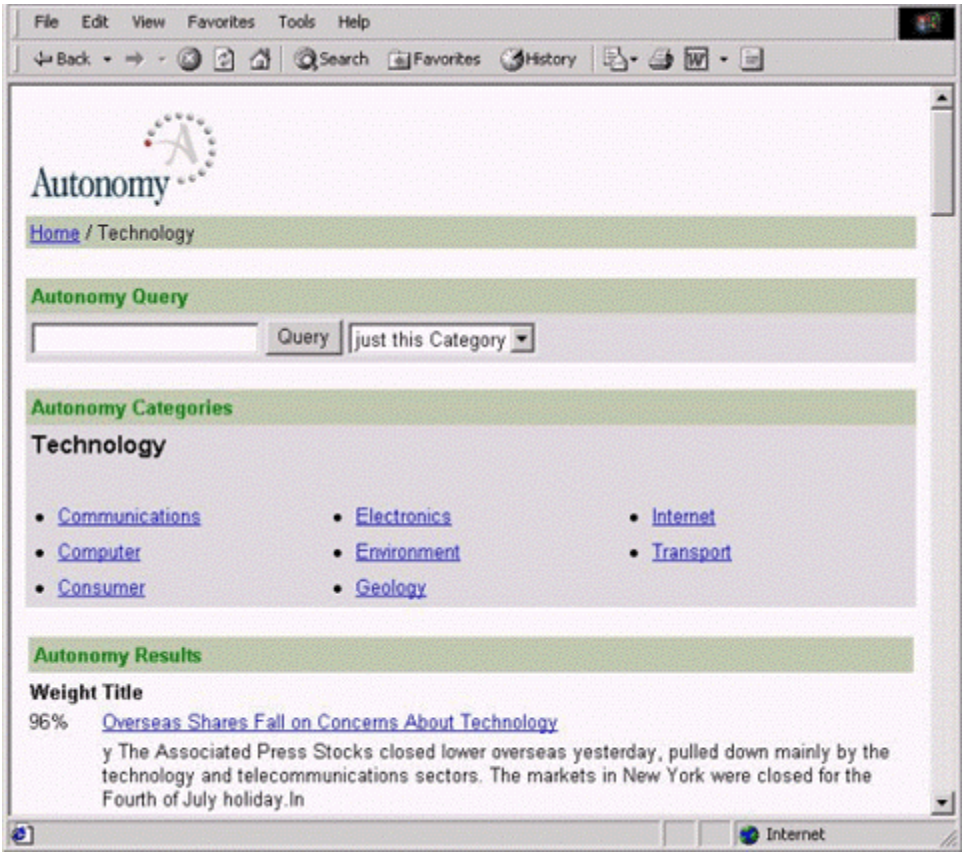


Figure 34: The category "Technology," its subcategories, and one document record listing in the Autonomy user interface.

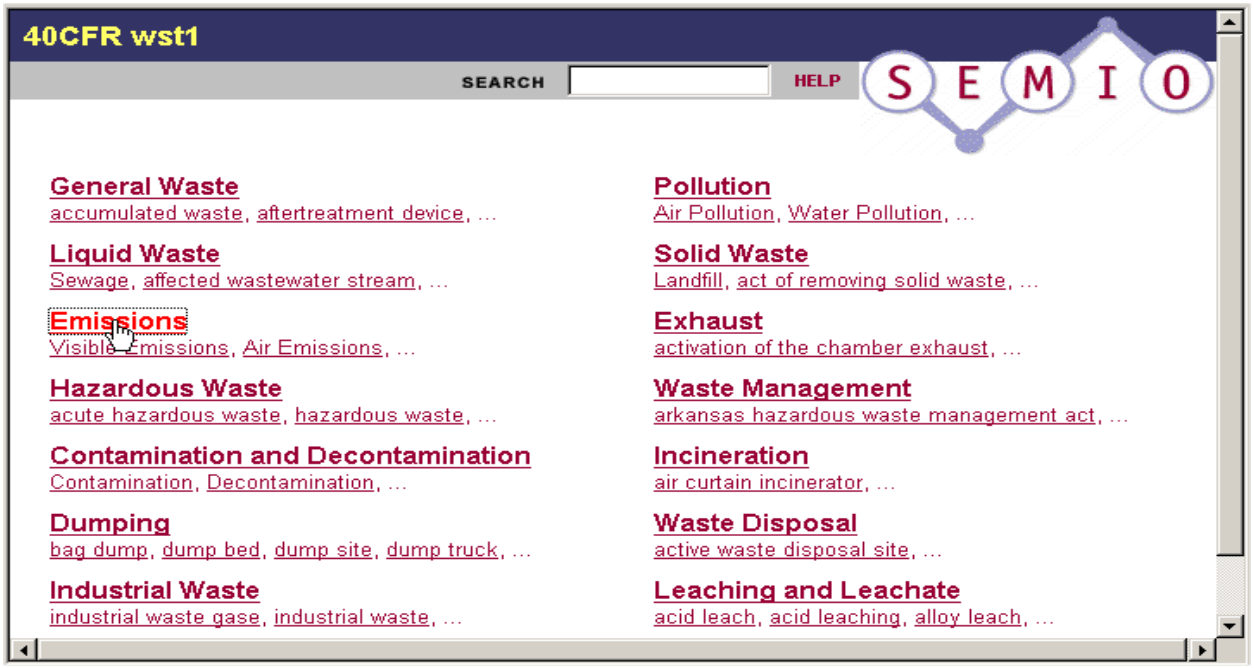


Figure 35: A categorization hierarchy viewed through the Semio user interface. The mouse icon indicates a click into the category "Emissions." Note that the categorization structure is the same as that in Figure 27. The underlying document content is the same as that in Figure 12.

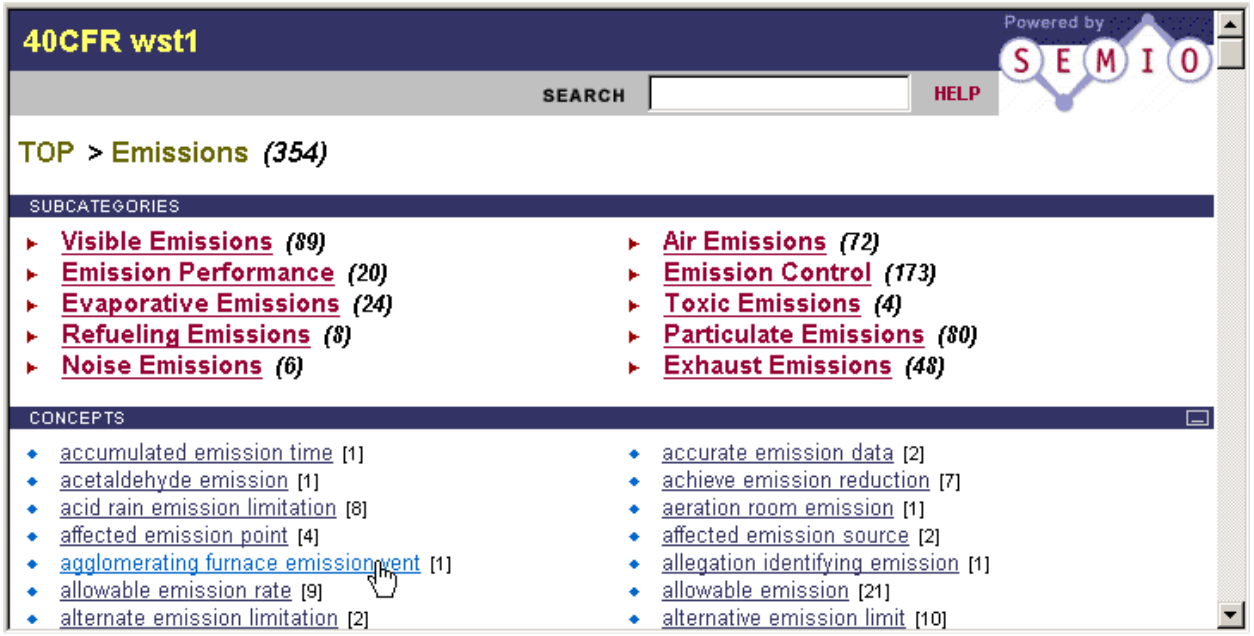


Figure 36: The Semio user interface, showing the category "Emissions," its subcategories, and concepts that have latched into the Emissions category at this level. Mouse icon indicates a click through to the documents in which the concept "agglomerating furnace emission vent" occurs.

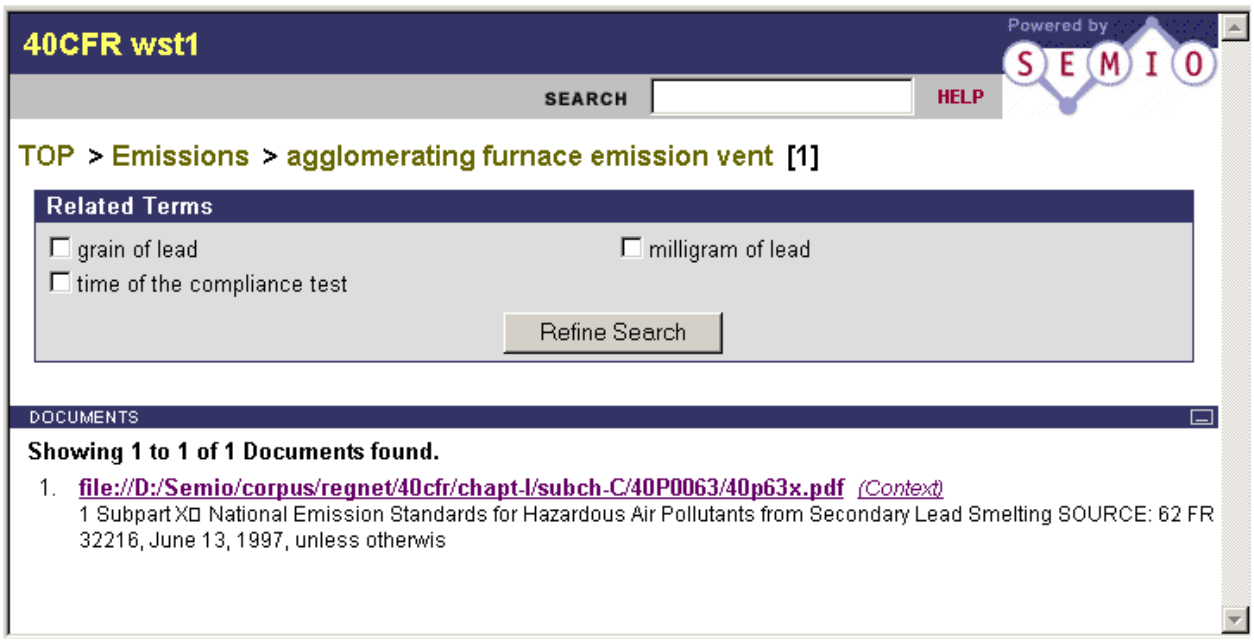


Figure 37: A document record in the Semio user interface.

Whether these traditional Yahoo-style interfaces to classification hierarchies really enable more efficient browsing than the non-traditional UIs is an area for further research. Qualitatively, it seems to be the case. Yet regardless of the type of UI, all successful interfaces to a classification hierarchy must enable a user to move easily between focus and breadth, to keep track of one's place in the hierarchy even while homing in on a piece of information, to facilitate access to content without neglecting its context.

Conclusion

This paper has explored three general approaches to classifying unstructured text: manual, partially automated, and largely automated categorization. Manual classification efforts require tremendous input of effort and rigorous quality-control processes in order to compensate for humans' limited ability to apply classification rules in a consistent fashion. In practice, the costs of a high quality manual classification can be prohibitive. As a result, usually either quality suffers or organizations look to technology for assistance.

Partially automated classification involves human specification of classification rules, but automated (and therefore consistent) application of those rules. The weak link in partially automated classification is in the creation of the rules. Without a quality-control process, poorly-thought-out classification rules can result in the misclassification or non-classification of very large numbers of documents. Consistent application of rules means little if the rules themselves are flawed. Yet, with well-thought-out rules a partially automated classification system can provide a good blend of human judgment and computational efficiency/consistency.

In largely automated classification, a computer learns classification rules by evaluating a training set of already-categorized documents. The system then classifies new documents according to the rules it has learned from the training set. Human auditing is necessary in order to verify that the rules the system has learned are leading to accurate classifications decisions. A human editor can manually override classification decisions by moving wrongly categorized documents to different categories. By re-running the training process periodically, the system can update its classification rules to incorporate the human editor's decisions. There are two main weaknesses with such systems. The first involves the training set of documents. Usually, the training set of already-categorized documents is the result of a manual effort. If the training set is large, the inconsistency of manual categorization can create an original flaw in a largely automated system's categorization rules. Conversely, if the training set is small it is unlikely to provide a sufficient basis for the system to categorize an inflow of new documents. The second main weakness involves a lack of transparency for the categorization rules themselves. In largely-automated systems, the rules that the system has learned are either difficult to comprehend or altogether inaccessible. As a result, such systems are hard to audit for quality. Nonetheless, in situations where there is some allowance for error, largely automated classification systems can be a good solution for efficiently categorizing a large quantity of text.

Regardless of the type of approach to classification, in evaluating a text classification engine the ability to access the engine's classification data in a raw format should be an important factor. Not only should one assess how well the engine will classify unstructured text, one should also determine how easily the resulting classification data can be used as an input to further computing operations. The reason Visual Net and TheBrain are able to apply their novel user interfaces to the same category information as one finds in the Open Directory is because the Open Directory Project makes its data available for post processing (Figures 21, 30, and 15, respectively).

Most prominent vendors of classification engines do enable post-processing, and most use XML to do so. Interwoven and Semio both can export XML tags that describe at the document level the categories in which the document participates. Similarly, Inxight's Categorizer comes with an API in Java and C that developers can use to extract XML tags for use in other applications. On the other hand, IBM has a text categorization product[20] that exports categorization information not in XML but in flat text files.

Overall, the categorization of text based upon its conceptual content and the availability of the raw categorization data in computable format opens exciting possibilities for dynamic document and content management. In the publishing industry, the "book" or "magazine" as a unit could give rise to dynamically generated custom publications in which text is automatically pulled from databases depending on how it has been categorized and tagged. Online libraries will be able to permit searches not only on author names, document titles, but also on category information, thereby becoming more than just electronic versions of the old card catalogs. Decision-support tools will arise that leverage categorized text to enable more efficient compliance with regulations, faster responses to masses of email, or real-time routing of local news information to business travelers overseas. Moreover, the maturing of Automatic Speech Recognition technology means categorization engines may soon be able to process unstructured audio archives in addition to text.

[20] See http://www-3.ibm.com/software/data/iminer/fortext/index.html

As the industry moves towards this future, developers of classification hierarchies must ensure that the relationships between parent and child categories are consistent throughout the structure, and that the features on which a classification is based are central rather than tangential. Developers of text categorization engines must focus on minimizing the rate of false classifications while comprehensively classifying text into all of the appropriate categories. The question "If a tree falls in the woods, and no one is there to hear it, does it make a sound?" can be applied to text categorization: "If a user enters a query, and a certain relevant document is not returned, does that document exist?" Increasingly, a user's answer might be, "no it does not."

Users are beginning to treat the set of documents that have been indexed by categorization engines as if it were both accurate and exhaustive. While accurate, comprehensive classification is the goal, the state-of-the-art is not there yet. Consequently, on the one hand users must not naïvely rely too much upon back-end categorization and indexing processes when searching for text-based digital information. On the other hand, the developers of categorization hierarchies and categorization engines must realize their immense responsibility to end users: as users assume increasing levels of accuracy and comprehensiveness, the risk increases that documents that fall through the cracks will seem not to exist.

Appendix A – Product Matrix

Product Matrix

Organization	Information Discovery for Classification Structures	Categorization Engines	User Interfaces to Classification Structures
Autonomy	Autonomy Clusterizer	Autonomy Categorizer	Autonomy Portal-in-a-Box
ClearForest	ClearForest ClearResearch	ClearForest ClearTags	NA
Hummingbird	Hummingbird Fulcrum Knowledge Server-Automatic Taxonomy Builder	NA	Hummingbird Enterprise Information Portal
IBM	IBM Intelligent Miner for Text, Clustering Tool	IBM Intelligent Miner for Text, Categorizer Tool	IBM Enterprise Information Portal
Interwoven	NA	Interwoven Metatagger	Interwoven TeamSite
Inxight	Inxight Murax	Inxight Categorizer	Inxight Star Tree
Open Directory Project	NA	Volunteer Human Editors (Open Source Manual Classification)	Directory Mozilla (www.dmoz.org)
Pacific Northwest National Lab	SPIRE (Spatial Paradigm for Information Retrieval and Visualization) - ThemeView	NA	NA
Quiver	Quiver QKS Classifier-Automatic Taxonomy Builder	Quiver QKS Classifier	Quiver QKS Output and Display Interface
Semio	SemioMap Discovery	Semio Taxonomy	Semio Taxonomy
Verity	NA	Verity Intelligent Classification	NA
VisualNet/Antartic.ca	NA	NA	VisualNet Geographic Metaphor Interface

Appendix B – Information Discovery Products

Autonomy Clusterizer

Feature	Product
Name of Product	Autonomy Clusterizer
Type of Interface	Visual Display of Conceptually Related Clusters in a Text Collection
Direct Access to Information Space?	Yes
Stand-Alone Product or Integrated Product?	Stand Alone or Integrated with Autonomy's Categorizer Product.
Overall Review	The Autonomy Clusterizer reveals the main clusters of topics in a corpus of text. This can be used to analyze changes in a text collection over time, as new topics become current and other topics are no longer "hot." More importantly for the purpose of developing classification structures, this direct view into the conceptual landscape of a corpus of text provides an invaluable overview. In addition to providing a direct view into this landscape, the Autonomy Clusterizer can automatically generate a taxonomy for use by the Autonomy Categorizer. By reviewing the category and subcategories of an automatically-generated taxonomy, one can also gain insight into a document collection. These two information discovery tools can help in the process of building classification hierarchies.

ClearForest ClearResearch

Feature	Product
Name of Product	ClearResearch
Type of Interface	Visual Display of Relationships Between Terms
Direct Access to Information Space?	Yes and No
Stand-Alone Product or Integrated Product?	Stand Alone
Overall Review	relationships between central terms, people, company names, or even places within a body of text. Users can browse these relationships through a web-like visual interface. In this sense, the tool is useful for information discovery. ClearForest allows--or requires--the input of "Rulebooks" to define the relationships between terms.

Hummingbird Fulcrum Knowledge Server

Feature	Product
Name of Product	Hummingbird Fulcrum Knowledge Server
Type of Interface	Administrator's Tool for Clustering - Used with Automatic Taxonomy Building Tool
Direct Access to Information Space?	No
Stand-Alone Product or Integrated Product?	Integrated with Hummingbird's Automatic Taxonomy Product
Overall Review	The Hummingbird Fulcrum Knowledge Server has the capability to generate a taxonomy of terms automatically from a corpus of text. This initial categorization of text is then used as a training set for Hummingbird's neural net text classification system. Using the automatic taxonomy function and then observing the relationships between the categories and subcategories, one can indirectly discover the topics that are central in the text collection. Although this approach can work for information discovery, it would be better to use a product specifically designed for that purpose.

IBM Intelligent Miner for Text – Clustering Tool

Feature	Product
Name of Product	IBM Intelligent Miner for Text, Clustering Tool
Type of Interface	Browsable Tree of Documents
Direct Access to Information Space?	Yes
Stand-Alone Product or Integrated Product?	Stand Alone or Integrated with IBM's Text Classification Product.
Overall Review	into clusters according to the terms--and the relationships between those terms--that it finds in the documents. Each node of the cluster tree indicates the top three terms which, together, define that node. Documents beneath that node deal with those terms (among others). IBM's Clustering tool provides a way to survey the overall conceptual space in a document collection. The product can be used on its own as a way to access documents, or it could be used to help in the building of a classification structure. For example, the fully-automated feature of IBM's text categorization tool (Categorizer) uses this clustering tool to produce an initial taxonomy.

Inxight Murax

Feature	Product
Name of Product	Inxight Murax
Type of Interface	Search-Mediated List of Documents
Direct Access to Information Space?	No
Stand-Alone Product or Integrated Product?	Stand-Alone
Overall Review	documents returned. Murax analyzes relationships between concepts in the corpus of text and returns documents based on topic or content similarities. Users can specify whether the results should be tied to a specific subject, a specific phrase or phrases, or an overall concept. A search through Murax will return a broader set of relevant documents than would a traditional key-word search. However, Murax reveals information about a document collection only through the mediation of a search term.

Quiver QKS Classifier – Automatic Taxonomy Builder

Feature	Product
Name of Product	Quiver QKS Classification Engine
Type of Interface	Administrator's Tool for QKS Classifier
Direct Access to Information Space?	No
Stand-Alone Product or Integrated Product?	Integrated part of Quiver's QKS Classifier Product
Overall Review	automatic development of a categorization structure and the automatic classification of documents into that structure. When the tool is used for categorization, an administrator can step in to confirm or override specific classification decisions. This automatic taxonomy-building feature could be used as an information discovery tool. By letting the software develop an initial classification based upon its analysis of the text collection, a user can learn what major topics are discussed in that collection. This knowledge can help in the development of an evidence-driven classification structure.

SemioMap Discovery

Feature	Product
Name of Product	SemioMap Discovery
Type of Interface	Visual Network of Co-occurrence Relationships Among Terms
Direct Access to Information Space?	Yes
Stand-Alone Product or Integrated Product?	Both
Overall Review	SemioMap Discovery uses paragraph-level phrase co-occurrence to develop a lexical network of relationships among terms in a corpus of text. It can provide a visual overview of the conceptual space of very large text collections. Users can navigate the network of phrase co-occurrences and drill down to documents in which those phrases occur. The tool can also be used for creating classification structures that are custom-fit to the topics in a corpus of text. Rather than--or in addition to--creating a classification structure in a top-down manner, an organization can use SemioMap Discovery to build such a structure from the textual evidence.

Pacific Northwest National Lab – SPIRE

Feature	Product
Name of Product	Spatial Paradigm for Information Retrieval and Visualization ThemeView
Type of Interface	Visual Map of Topically-Related Clusters of Documents
Direct Access to Information Space?	Yes
Stand-Alone Product or Integrated Product?	Stand-Alone
Overall Review	product of the Pacific Northwest National Lab, located in Washington state. The tool takes as input a body of text, identifies central and tangential topics as well as the relationships betweent them, and presents this information to the user in a topographical map-style interface. "Mountain peaks" on the map represent a cluster of documents about a given topic. The larger and taller the peak, the more documents are about that topic. The distance between two peaks represents how closely related their respective topics are to one another. This product can be used as a tool to help in the construction of a content-driven classification structure.

Appendix C – Categorization Engines

Autonomy Categorizer

Feature	Product
Name of Product	Autonomy Categorizer
Automatic generation of taxonomy	Yes, using Autonomy Clusterizer
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	No
Training data (supervised learning)	Yes
Lexical rules	No
Linguistic rules	No
False positives/ false negatives	Risk of false classifications depends upon the quality of the training data.
Auditing capabilities	Autonomy Categorizer uses Bayesian Inference to assign new documents to categories. The rules "learned" by the system are not readily apparent to the administrator, although it is possible to "teach" the system on an ongoing basis by manually reassigning wrongly classified documents into the appropriate categories and then re-running the training process.
Overall Review	Autonomy Categorizer is an inference-based supervised learning system. Such systems are most suitable when you need coarse-to-medium granularity and for situations where there is a moderate tolerance for error. Autonomy offers a broad range of information retrieval, information management, and information discovery tools.

ClearForest ClearTags

Feature	Product
Name of Product	ClearForest ClearTags
Automatic generation of taxonomy	No
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	No
Training data (supervised learning)	No
Lexical rules	Yes
Linguistic rules	Yes
False positives/ false negatives	Depends upon the lexical and linguistic rules (rulebooks) applied during the tagging process.
Auditing capabilities	documents by defining lexical relationships (relationships between terms) as well as linguistic patterns that signify a given document should be classified in a given way.
Overall Review	In ClearForest's methodology, entities are extracted and documents are clustered according to co-occurrence, and extracts more than just noun phrases. Entities such as zip codes, addresses, company names, or personal names can be identified for extraction. In terms of classification, ClearForest can apply lexical rules (eg: if a document contains this term, categorize it this way), as well as linguistic rules (eg: if a document contains this linguistic pattern, categorize it this way).

IBM Intelligent Miner for Text – Categorizer

Feature	Product
Name of Product	IBM Intelligent Miner for Text, Categorizer
Automatic generation of taxonomy	Yes, optional
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	Yes, IBM claims
Training data (supervised learning)	Yes
Lexical rules	No
Linguistic rules	No
False positives/ false negatives	Risk of false classifications depends upon the quality of the training data.
Auditing capabilities	requires training data.
Overall Review	learning categorization system that can run in a partially automatic or fully automatic mode. Such systems are in which there is a moderate tolerance for error. IBM's Intelligent Miner family of products includes additional components beyond the Categorizer. These products range from a topical clustering tool (Clusterizer) as well as a Summarizer. The IBM Enterprise Information Portal product serves as a framework within which one can implement a full text-mining solution.

Interwoven Metatagger

Feature	Product
Name of Product	Interwoven Metatagger
Automatic generation of taxonomy	No
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	No
Training data (supervised learning)	No
Lexical rules	Yes
Linguistic rules	No
False positives/false negatives	The lack of contextual awareness increases the risk of false positives.
Auditing capabilities	custom-built or industry-standard controlled vocabularies as the basis for categorizing documents. One can view the whether they serve as strong signifiers for a particular topic. For example: If a document containing term XYZ must be about the topic T by the very fact of its containing that term, XYZ is a strong signifier for that topic. Consequently, documents containing the term XYZ can reliably be tagged as dealing with topic T.
Overall Review	product to Interwoven's portal, TeamSite. Within TeamSite, the product can be employed to automate the classification of text.

Inxight Categorizer

Feature	Product
Name of Product	Inxight Categorizer
Automatic generation of taxonomy	No, administrators must provide a pre-categorized training set
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	Yes, Inxight claims
Training data (supervised learning)	Yes
Lexical rules	No
Linguistic rules	No
False positives/ false negatives	Risk of false classifications depends upon the quality of the training data.
Auditing capabilities	Inxight Categorizer uses probabilistic Inference to assign new documents to categories. Inxight Categorizer has an iterative, interactive process for creating a training set. User-feedback in the training process can help improve Categorizer's accuracy when it is applied to new text. It is possible to "teach" the system on an ongoing basis by manually reassigning wrongly classified documents into the appropriate categories and then re-running the training process
Overall Review	learning system. It is most suitable when you need coarse-to-medium granularity, and when you have a moderate tolerance for error. Inxight is a component vendor, selling multiple knowledge management products that can help "supercharge" an existing knowledge management infrastructure. Presently, Inxight does not sell a package that pulls all these components together into one framework.

Quiver QKS Classifier

Feature	Product
Name of Product	Quiver QKS Classifier
Automatic generation of taxonomy	Yes, optional
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	No
Training data (supervised learning)	Yes
Lexical rules	No
Linguistic rules	No
False positives/ false negatives	Risk of false classifications depends upon the quality of the training data.
Auditing capabilities	inference-based classification tool. QKS Classifier can run in a fully-automatic or partially automatic mode. In the partially-automatic mode, the system requires human confirmation of the system's classification decisions and this provides a real-time auditing environment.
Overall Review	judgement in the classification process. When QKS Classifier runs in partially-automated mode, document solution as well as from real-time human oversight. This hybrid approach provides the scalability of inference-based supervised learning systems joined with the reliability of manual oversight.

Semio Tagger

Feature	Product
Name of Product	Semio Tagger
Automatic generation of taxonomy	No
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	No
Training data (supervised learning)	No
Lexical rules	Yes
Linguistic rules	No
False positives/ false negatives	Risk of false classifications depends upon the lexical rules supplied to the classification system.
Auditing capabilities	With Semio Tagger, the rules by which a given document gets categorized are transparent because they rely upon noun phrases rather than linguistic rules or probability models. To audit a classification, the specific lexical rules that applied to a document are reviewed.
Overall Review	that make the classification logic transparent. Semio classifications can be highly accurate as a result.

Verity Intelligent Classification

Feature	Product
Name of Product	Verity Intelligent Classification
Automatic generation of taxonomy	Yes, Optional
Allows synonym rules	Yes
Allows acronym rules	Yes
Polysemy (contextual awareness)	Unclear
Training data (supervised learning)	Yes, Optional
Lexical rules	Yes
Linguistic rules	No
False positives/ false negatives	configure the system in a number of ways, each of which has different implications for the accuracy of classification tends to give more errors than a highly-supervised system.
Auditing capabilities	Classification system allow rigorous auditing and quality control. The option for strong involvement of human capabilities make it easy to determine just why a given document was classified in a given way.
Overall Review	Verity's Intelligent Classification has an unsupervised automated mode as well as a supervised mode that involves human judgement. The unsupervised automatic mode works for cases when speed is critical. Verity can take existing classification structures and refine them, either with or without human editorial input. Verity can involve human judgement at a granular level when accuracy is critical.

Appendix D – User Interfaces to Classification Structures

Autonomy Portal-in-a-Box

Feature	Product
Name of Interface	Autonomy Portal-In-A-Box
Type of Interface	Yahoo Directory Style *or* Search Term with List of Results
Browse Classification Structure Directly?	Yes
Overall Review	Autonomy's User Interface to categorized text is a Yahoo-style hierarchy of categories. In any given category, the contents are documents that have been placed there by some form of categorization process. Presumably, this would have been Autonomy's Categorizer tool. Yet there is no reason one could not use an alternative categorization engine and then use Autonomy's portal product as a framework to access already-categorized text. The reason for a given document's categorization is not apparent through the interface itself--users must take it on faith that documents have been categorized correctly. As an alternative to browsing the hierarchy directly, users can enter search terms into a search field. The query yields documents that exist in categories related to the search term. Autonomy's user interface offers users an effective way to search categorized text either by browsing or by entering search terms.

Hummingbird Enterprise Information Portal

Feature	Product
Name of Interface	Hummingbird Enterprise Information Portal (EIP)
Type of Interface	Corporate Portal
Browse Classification Structure Directly?	
Overall Review	management. With respect to the Hummingbird Enterprise Information Portal, users can enter search terms directly to receive a standard ranked list of results. Hummingbird also allows users to drill down within a classification hierarchy directly. Users can filter results within the hierarchy by topic, by source, by author, or a number of other criteria.

IBM Enterprise Information Portal

Feature	Product
Name of Interface	
Type of Interface	Corporate Portal Framework
Browse Classification Structure	No
Overall Review	ells a document management and content management product called the IBM Enterprise Information Portal. Through categorized by the IBM Categorizer product. The user interface to access documents is a search feature on the portal. Search results appear in a standard ranked list format.

Interwoven TeamSite

Feature	Product
Name of Interface	Interwoven TeamSite
Type of Interface	Corporate Portal Framework
Browse Classification Structure Directly?	No
Overall Review	Interwoven focuses on document management and content management. Interwoven sells an Enterprise Information Portal called TeamSite. Most of Interwoven's products, including the Metatagger categorization tool, work in conjunction with the TeamSite product. After documents have been tagged by Metatagger, the user interface to access those documents is a search feature on TeamSite. Search results appear in a standard ranked list format.

Inxight Star Tree

Feature	Product
Name of Interface	Inxight Star Tree
Type of Interface	Hyperbolic Tree
Browse Classification Structure Directly?	Yes
Overall Review	traditional approach to navigating a hierarchy structure. The Star Tree interface displays the root node at the center of a screen, with categories radiating outward. The first "ring" of categories around the root node resembles spokes of a wheel. The subcategories beneath each category radiate out beneath the parent in a fan shape. The pattern repeats for sub-subcategories ad infinitem. A user can drag any portion of the star tree to bring sections of the tree into or out of center screen. The interface provides the user a greater sense of navigational control and it lessens the likelihood of missing a relevant category when drilling into the hierarchy.

The Open Directory Project – Directory Mozilla

Feature	Product
Name of Interface	Open Directory Project (DMOZ)
Type of Interface	Yahoo-style hierarchy or search terms
Browse Classification Structure Directly?	Yes
Overall Review	into categories. The result of this categorization can be viewed online at www.dmoz.org. The user interface to the Open Directory is a traditional hierarchy. Users can drill down to documents or they can enter search terms.

Quiver QKS Output and Display Interface

Feature	Product
Name of Interface	Quiver QKS Output and Display Interface
Type of Interface	Yahoo-style hierarchy or search terms
Browse Classification Structure Directly?	Yes
Overall Review	or browse the hierarchy directly. In addition, users can conduct searches that are constrained to a certain topic within the hierarchy. This ability to limit the search domain provides a specific search capacity that is of benefit to the user. Users can also find related content for further searching through suggestions made by Quiver.

Semio Taxonomy (Viewer)

Feature	Product
Name of Interface	Semio Taxonomy
Type of Interface	Yahoo-style directory or search terms
Browse Classification Structure Directly?	Yes
Overall Review	hierarchy. In addition to showing which documents populate a given category, the interface also shows the specific terms contained in those documents that led to that classification. Users can refine the search to reveal documents that contain a specific set of terms in co-occurrence relationships. A context feature shows the paragraph(s) in which the search term(s) occur.

Antarcti.ca VisualNet Geographic Metaphor Interface

Feature	Product
Name of Interface	VisualNet
Type of Interface	Geographical Metaphor with Drill-Down Capacity, or Search Term Option
Browse Classification Structure Directly?	Yes
Overall Review	Antarcti.ca (http://www.antarcti.ca) offers a visualization interface that can be applied to categorized text. The interface is called Visual Net. Through this interface, top-level categories are represented as regions on a geographical "map." The relative area of each category indicates the relative quantity of documents in that part of the categorization hierarchy. Users drill down by clicking into a portion of the map, revealing a new map representing the subcategories within that category.