Alternatives for Implementing the National Institutes of Health's Research, Condition, and Disease Categorization System

FINAL REPORT

September 2010

Prepared for:
Luci Roberts, Ph.D.
Office of Extramural Research
National Institutes of Health (NIH)

Prepared by:
Edward Kenschaft
Brian Zuckerman, Ph.D.

Science and Technology Policy Institute
1899 Pennsylvania Avenue NW, Suite 520
Washington, DC 20006

Preface

This work was conducted by IDA's Science and Technology Policy Institute under contract OIA-0408601, Task OSTP-20-0002.34, "Testing, Evaluation, and Use Prospectus of the Research, Condition, and Disease Categorization (RCDC) System," for the National Institutes of Health (NIH) Office of Extramural Research (OER).

The RCDC system is currently implemented using Collexis, a document categorization platform. OER tasked IDA to: (1) identify and compare commercial alternatives to the Collexis implementation; (2) propose and discuss alternative technical approaches for implementing RCDC; and (3) recommend ways the RCDC system (or alternatives) can and should be leveraged further within NIH and its affiliated Institutes and Centers (1). This report provides the results of these tasks.

Table of Contents

List of Equations

List of Tables

List of Figures

1 Executive Summary

RCDC categorizes grant proposals and documents regarding other NIH activities, currently on the order of 80,000 per year, for reports to Congress (2) and for public access (3). The current RCDC implementation, unveiled in 2009, uses the Collexis (§8.1.1) *document categorization* platform. When RCDC received a mixed response, OER tasked IDA to: (1) identify and compare commercial alternatives to the Collexis implementation; (2) propose and discuss alternative technical approaches for implementing RCDC; and (3) recommend ways the RCDC system (or alternatives) can and should be leveraged further within NIH and its affiliated *Institutes and Centers* (ICs) (1).

For this study, the incumbent Collexis system was compared to various alternative platforms, of which five were reviewed in detail:

1. Recommind Decisiv Categorization (§8.1.3)
2. InfoSphere Classification Module (§8.1.4)
3. Autonomy IDOL (§8.1.6)
4. SAS Enterprise Content Categorization (§B.1.1)
5. SmartLogic Semaphore (§B.1.2)

Evaluations were based on 50 weighted criteria, considered in four weighted groups: Business (20%), Technical (20%), Functionality (40%), and Usability (20%) (§7.1). Of the five alternatives to Collexis considered in greatest detail, three scored higher than did the incumbent system. These three alternatives all score substantially higher than does Collexis on the Business criteria, and scored equal to or higher than Collexis on the Functionality criteria. Two of the three also scored higher on the Technical criteria. Collexis scored highest on the usability criteria.

There is no indication that any recent innovations in document categorization research produce dramatically better results than the technology used in the most sophisticated commercial tool. There is therefore no compelling incentive to develop and maintain a custom solution (§8.3).

In addition to the platform comparison, a methodology is also suggested for evaluating and refining the category structure, within the constraints of Congressional reporting needs, in order to facilitate more consistent and defensible categorization (§10.1).

Finally, suggestions are provided regarding additional applications where the RCDC technology could be leveraged further by NIH and its affiliated ICs (§11).

2 Explanation of Formatting

Throughout this report, the following fonts are used:

- *italics*: term listed in the Index of Technical Terms (§D).
- `evenly spaced`: literal text
- SMALL CAPS: the name of a *concept* (§5.2.3)
- LARGE CAPS: the name of a *category* (§5.2.4)

3 Background

In 2003, a committee of the National Research Council recommended that NIH "improve the quality and analysis of its data on the allocation of NIH funds by disease for planning and priority-setting purposes" (4). Section 402B of the National Institutes of Health Reform Act of 2006 directed NIH to "establish an electronic system to uniformly code research grants and activities" of NIH and affiliates, producing results that would be searchable by a variety of criteria (5).

In anticipation of these requirements, work began in 2002 on a pilot for a document categorization system implemented using the Collexis (§8.1.1) platform (6). NIH acquired a Collexis site license in January 2004 (7). RCDC began beta testing in 2005, and was formally launched in February 2009 (8), categorizing NIH-funded research from FY 2007 and FY 2008. For FY 2007, categorization reports were generated both manually and using RCDC, allowing results to be compared. Both sets of results are reported on the NIH website (2). Starting with FY 2008 data, NIH produced only automated RCDC results.

Reception of RCDC among IC *domain experts* was mixed, in part because system maintenance required substantially more of their time than anticipated, and in part because categorization results often differed from expert judgments. Accordingly, in 2008 NIH tasked IDA to evaluate RCDC's *correctness* on beta testing data from FY 2005 and FY 2006. The report, submitted in March 2009, concluded, "In general, the RCDC system showed good agreement with subject matter experts for only a small fraction of the roughly 200 categories" (9). The largest, hence most financially significant, categories showed the poorest results. Changes intended to improve correctness were made in later years, but later results have not been formally evaluated.

Part of the problem is confusion over category definitions, independent of the choice of platform (§10.1).

4 Information Sources Used to Conduct the Study

Preliminary information for the study was garnered from published sources, including vendor documentation and industry reviews by groups such as Gartner and Forrester. Five contenders appeared to satisfy most of the key evaluation criteria (§6.1):

1. Recommind Decisiv Categorization (§8.1.3)
2. IBM InfoSphere Classification Module (§8.1.4)
3. Autonomy IDOL (§8.1.6)
4. SAS Enterprise Content Categorization (§B.1.1)
5. SmartLogic Semaphore (§B.1.2)

For each of these five top contenders the study team arranged for a demonstration, followed in most cases by a more detailed technical presentation and discussion. This narrowed the list to three, presented along with Collexis in §8.1. Several rejected solutions are mentioned in Appendix §B.

5 Current RCDC Methodology

This section summarizes the existing RCDC/Collexis categorization methodology.

5.1 Contributors

The key contributors to the RCDC system are medical domain experts and RCDC staff. The latter include Scientific Information Analysts (SIAs) and the Thesaurus Curator.

5.1.1 IC Experts

NIH consists of 27 *Institutes and Centers* (ICs), as established by Congress (5). IC medical experts provide the knowledge at the core of RCDC, through collaborative decision-making sessions and independent validity checks of the categorization results (§5.3.7).

5.1.2 Scientific Information Analyst (SIA)

Each new category (§5.2.2) is assigned a *Scientific Information Analyst* (SIA) from among RCDC staff. The SIAs moderate collaborative sessions with IC experts and incorporate their decisions into the knowledge base (10).

5.1.3 Thesaurus Curator

The *Thesaurus Curator*, also an RCDC staff member, vets every proposed change to the concept thesaurus (§5.2.3). The Curator also writes the rules that introduce *concepts* from *terms* (§5.2.3).

5.2 Data Definitions

5.2.1 Document

A *document*, also known as a *project*, is (in this context) a digital representation of text related to a grant application or other NIH activity. It typically differentiates key sections (§A.3.3.3) including title and abstract.

5.2.2 Term

A *term* (or *keyword*) is a meaningful word or word cluster, often a noun phrase, which appears in a document. Terms are typically *normalized* so that only the root forms are considered, e.g., `grow bean` instead of `growing beans` (§5.3.1). Only terms of interest in defining concepts and categories are retained.

5.2.3 Concept

A *concept* is a unit of meaning recorded in a *thesaurus*. The RCDC thesaurus contains tens of thousands of concepts. It is based on a harmonization of several medical thesauri, amended with supplemental concepts deemed to be of use in defining categories.

Synonymous terms are distinct terms that map to the same thesaurus concept. For example, in a particular thesaurus, the terms `cancer`, `carcinoma`, and `malignant cyst` may all map to the concept `CANCER`.

A *polysemous* term is a single term that maps to multiple concepts in the thesaurus. These can be full homonyms, e.g., `bank`, the side of a riverbed or a financial institution; or they may involve more subtle distinctions of related meanings, e.g., `rest`, to remain stationary or to relax after labor.

The RCDC thesaurus is updated by the Curator roughly every three weeks, based on suggestions from IC experts. For each new proposed concept, the Curator subjectively evaluates its usefulness and potential impact on other categories (10).

5.2.4 Category

A *category* describes a class of related documents. The list of categories is defined by Congress and updated annually. In FY 2010, twenty new categories were added. Each year the process of adding new categories takes six to eight months elapsed time.

Categories are divided among RESEARCH, CONDITION, and DISEASE. With a few specified exceptions, categories may be infinitely overlapping. A typical document is assigned 4-7 categories. A small category may apply to only a few documents, while a large category may include 25,000 documents.

A category can have subcategories. In some cases, the subcategories exhaustively define the parent category, while in other cases they do not (10). Categories thus form a hierarchical *taxonomy*, but the status of parent categories is ill-defined.

Congress provides the canonical definition of each category, based on recommendations from NIH and other groups of stakeholders within the biomedical research community. However, in many cases this definition differs from the most common usage among practitioners. Furthermore, not uncommonly, practitioners in different ICs use conflicting definitions, based on the differing mission and focus of each IC. When a new category is introduced, one of the first steps is for IC experts to settle on a common definition that reconciles the various points of view, through collaborative sessions moderated by the SIA. In some cases, the result is markedly different from the congressional definition. Also, IC experts involved in validity checking (§5.3.7) may not always be aware of these decisions, instead categorizing documents according to the definitions with which they are familiar.

5.2.5 Category Concept Weight

For each category, a collaborative session of IC experts determines which concepts contribute to that category. Each concept is assigned a *concept weight* between 0 and 1, reflecting its relative contribution to the category. The category concept weight is subjective, and often needs to be revised repeatedly based on unexpected categorization results (§5.3.8).

Concept weights are defined separately for documents (§5.3.4).

5.2.6 Fingerprint

A *fingerprint* (or *concept profile*) is defined by the concept weights of the various concepts that contribute to the category or document. If the fingerprint for a document matches the fingerprint for a category within a specified threshold, the document is assigned to that category (§5.3.5).

5.3 Categorization Process

The overall process of categorizing a new document involves the following steps:

1. Normalize the terms in the title and text (§5.3.1).
2. Map terms to concepts, disambiguating as needed (§5.3.2).
3. From the terms, count the occurrences of recognized concepts within the title and text.
4. Assign a weight to each concept in the text (§5.3.4).
5. Assign categories by comparing concept weights for the document to those for each category (§5.3.5).
6. Apply additional business rules as appropriate (§5.3.6).
7. Perform validity checking (§5.3.7).
8. Revise knowledge base as necessary (§5.3.8).

5.3.1 Normalize Terms

Term *normalization* converts words and word clusters into a smaller number of canonical terms using *natural language processing* (NLP) heuristic processes. For example:

1. morphological analysis:

   `treating cancers` → `treat cancer`

2. stop word removal:

   `efficacy of the drug` → `efficacy drug`

3. word permutation:

   `efficacy [of the] drug` → `drug efficacy`

4. abbreviation expansion:

   `Frog virus (FV) is infectious … FV affects frogs.` →

   `Frog virus is infectious … Frog virus affects frogs.`

5. concept expansion:

   `heart, lung and liver transplants` →

   `heart transplants, lung transplants, liver transplants`

The Thesaurus Curator can configure which heuristics apply to particular words or clusters. For example, the Curator can set flags to prevent either `research vision` or `vision of this research` from being interpreted as `vision research` (10).

5.3.2 Map Terms to Concepts

Each of the normalized terms from the previous step is compared to the thesaurus. If a matching thesaurus entry is found, the matching concept is used; otherwise the term is ignored. Only the most specific matched concept is recognized. For example, `brain cancer` matches the concept `BRAIN CANCER` but not the concepts `BRAIN` or `CANCER`.

Both *synonymy* and *polysemy* are common in the thesaurus (§5.2.3). There is a feature in Collexis, currently not used in RCDC, by which the Thesaurus Curator can write context rules to *disambiguate* the various senses of a polysemous term based on other nearby terms. For example, if the word `organ` occurs in close proximity to the word `transplant`, a rule might conclude that the term is being used in a medical sense. However, if `organ` occurs in close proximity to the word `music`, the system might conclude it is not the medical sense, even though a more appropriate sense is not available in the thesaurus.

The version of RCDC using Collexis v7.0 (§8.1.3.3) will also support disambiguation using part-of-speech tags (1).

5.3.3 Count Concepts

The system counts the number of times each concept occurs in the body of each document, accounting for synonymous and polysemous terms. For example, if the word `carcinoma` and the phrase `malignant cyst` appear at different places in the document, the concept `CANCER` may receive a count of 2. Concepts that occur in the document title are treated specially (§5.3.4).

In the sections below, the number of times concept *i* occurs in the body of document *p* is represented as $k_{p,i}$.

5.3.4 Determine Document Concept Weights

For each concept occurring in a document, a *concept weight* is computed empirically.

Any concept that occurs in a document title is assigned a weight of 1.

Any other concept occurring in the document is assigned a weight between 0 and 1, calculated as the square root of its count divided by the square root of the count of the concept occurring most frequently in that document.

Equation 1: Document Concept Weight

$$weight\ of\ concept\ i\ in\ project\ p = w_{p,i} = \frac{sqrt(k_{p,i})}{sqrt(k_{p,max})}$$

For example, if the only two concepts of interest in the document are `BRAIN` and `FOOT`; and `BRAIN` occurs twice and `FOOT` occurs once; then the document concept weight is calculated as in Equation 2.

Equation 2: Document Concept Weight Example

$$w_{p,brain} = \frac{sqrt(k_{p,brain})}{sqrt(k_{p,brain})} = \frac{sqrt(2)}{sqrt(2)} = 1$$

$$w_{p,foot} = \frac{sqrt(k_{p,foot})}{sqrt(k_{p,brain})} = \frac{sqrt(1)}{sqrt(2)} = 0.71$$

Note that $w_{p,i}$ depends neither on the number of concepts in the document nor on the counts of any other concept other than the highest.[1]

5.3.5 Assign Categories

As described above, each concept is assigned a weight within each document or category to which it contributes. In the case of documents this weight is determined empirically (§5.3.4). In the case of categories, it is determined subjectively, by consensus of IC experts (§5.2.5).

For each document/category pair, the *similarity score* is calculated as the sum of the products of the respective concept weights over all concepts[2].

Equation 3: Similarity Score

$$similarity\ score\ for\ project\ p\ and\ category\ c = s_{p,c} = \sum_{\substack{i\ in \\ concepts}} w_{p,i} w_{c,i}$$

If the similarity score is above a configured threshold, the document is said to belong to that category (1-2).

---

[1] Collexis also supports other weighting heuristics, but these are not used in RCDC.

[2] In the implementation, the sum is actually performed over just those concepts that occur in both the document and the category, rather than all concepts in the thesaurus. However, the result is the same as that described here, given a default weight of 0.

5.3.6 Apply Business Rules

Various *business rules* preempt the process described above. For example, a business rule might indicate that every grant awarded by the National Institute on Aging should be assigned the category AGING, regardless of what terms or concepts occur (or do not occur) in the document. These business rules are identified and coded manually.

5.3.7 Perform Validity Checking

Sample RCDC results are sent to IC experts for *validity checking*. Experts mark each categorization as either *defensible* or *indefensible*. If most experts[3] consider the categorization defensible, it is accepted. If most consider the categorization indefensible, a panel is convened to revise the knowledge base (§5.3.8). If there is significant disagreement among experts, the SIA may schedule a collaborative session to resolve the differences (2). Although disagreement among experts has not been measured, it is thought to be high (§10.1).

Each year the validity checking process takes roughly three to four months elapsed time.

5.3.8 Revise Knowledge Base

Whenever validity checking determines that a document has been categorized incorrectly, a panel of experts attempts to determine where the system went wrong, in order to improve the *knowledge base* for the future. The diagnosis can be challenging, as the error may occur in any step of the categorization process, and may involve multiple factors.

Once the panel comes up with a diagnosis and the appropriate changes(s) are introduced, the classification is run again to see if the error is corrected. This may involve several iterations of

[3] This is the ideal model. More typically, only one expert reviews each result.

trial-and-error. No mechanism exists to determine how the change affects categorizations other than the one being examined.

6 Statistical Methodology

This section summarizes a statistical categorization methodology, particularly as would be involved in re-implementing RCDC. This discussion considers generic aspects of the methodology rather than assuming any specific tool.

6.1 Contributors

The contributors to a statistical methodology are the same as for the current implementation (§5.1), with the exception that the Thesaurus Curator is not needed.

6.2 Data Definitions

Documents, *terms*, and *categories* are the same as for the current implementation (§5.2). While it is possible to model intermediate *concepts*, none of the statistical systems under consideration do so, instead categorizing directly from terms in context (§6.3.5).

While the correlation between each term and category has a *correlation weight* (§6.3.5.2) analogous to a *concept weight*, these weights are calculated empirically rather than assigned by humans. The correlation weights can be displayed or analyzed as needed.

6.3 Categorization Process

Statistical document categorization systems such as Recommind Decisiv Categorization (§8.1.3) and IBM's InfoSphere Classification Module (§8.1.4) generally employ some variation of the following methodology:

1. Identify categories (§6.3.1)
2. Identify business rules (optional) (§6.3.2)
3. Elicit gold-standard documents (§6.3.3).
4. Divide categorized documents into training and test groups (§6.3.4).
5. Train system using training documents (§6.3.5).

6. Evaluate system using test documents (§6.3.6).
7. Elicit human expert review (§6.3.7).
8. Retrain using results of expert review (§6.3.8). Return to step 6.

This process cycles continuously, returning to step 1 whenever the list of categories changes.

6.3.1 Identify categories

Identifying an appropriate hierarchy of categories for a document categorization system (statistical or otherwise) is often a significant challenge. In many cases, categories are determined by business or functional needs. In other cases, a statistical system infers categories from a set of documents by a technique known as *clustering*. The clustering algorithm examines the documents for patterns of similarities and differences, and infers categories based on these patterns. Typically a user specifies the desired number of categories, although there are also heuristics for inferring this.

Having a canonical list of categories will greatly reduce the implementation time for any new system. However, the canonical list mandated by Congress may not be optimum for an automated system. For example, the list may include categories that overlap, or a category defined as the concatenation of two distinct subcategories, e.g., `GLAUCOMA AND OPTIC NEUROPATHY`. Techniques such as clustering can help identify categories that generate more reliable results, which are then mapped to the canonical list for reporting to Congress (§10.1).

6.3.2 Identify business rules (optional)

Although not integral to the methodology, many statistical systems support *inference rules* for categorizing documents according to business-related criteria (§17). These rules are generally applied in advance of statistical analysis. If useful, the results may be included in the *gold-standard* pool (§6.3.3).

6.3.3 Elicit gold-standard documents

Statistical systems require *gold-standard* expert-categorized documents from which to learn. With most systems, generating these documents is the most time-consuming and onerous part of the process. With RCDC, ample human-categorized documents already exist, so this step is straightforward. It will only become a challenge as new categories are added or the existing category structure is modified.

Note that in the real world, there is rarely such a thing as perfect data. Even gold-standard categorizations are likely to contain *noise*, where the human expert assigned the wrong category. A good automated system can identify and flag potential noise for human review (§6.3.7).

Furthermore, human experts often disagree with one another. Some systems include a mechanism to help measure *inter-annotator agreement*, but more often this is a separate process (§10.1). In general, categorizations on which experts differ should not be treated as gold-standard.

6.3.4 Divide categorized documents into training and test

Some of the gold-standard documents for each category are set aside for testing (§6.3.6), while the remaining are used for training (§6.3.5). Usually the documents are divided randomly, although researchers may choose some other method to produce more representative samples.

6.3.5 Train system

The heart of a statistical system is its ability to observe statistical correlations in the training documents. Generally, this will involve correlations between terms occurring in the text and the document's labeled category (or categories). Any other available feature can also be used, e.g., document length or submitting institution.

6.3.5.1 Term discovery

A statistical system generally discovers far more terms than a human would identify, partly because it considers words that a human might overlook, and partly because it considers all word sequences up to a specified length. (A sequence length of four is typical.) Polysemous words are naturally disambiguated (§5.3.2), since their context is always considered.

Most statistical systems incorporate *stemming*, a.k.a. *morphological analysis* (§5.3.1), by default. Most do not routinely eliminate *stop words*, since this often generates worse results by discarding useful information. However, stop word removal is sometimes offered as a configurable option.

For example, after normalization, the phrase `efficacy of the drug` might end up being represented as the term `efficac~ of the drug~`.

6.3.5.2 Correlation weight

Each term-category correlation is inherently assigned a *correlation weight* between 0 and 1 (or 0% and 100%). A high correlation weight indicates that the term occurs quite often within that category, and rarely within any other category. A low correlation weight indicates that the term occurs less often within that category than within other categories.

6.3.5.3 Confidence threshold

Each category is assigned a *confidence threshold* between 0 and 1 (or 0% and 100%), reflecting how lenient the system is when assigning documents to that category. At the extremes, a confidence threshold of 0 indicates that all documents are assigned to that category, while a confidence threshold of 1 indicates that only perfect matches are accepted.

Typically, the confidence threshold for each category is determined automatically, using the value that produces the most correct results on the training data. However, many systems also allow confidence thresholds to be configured manually, or based on user-defined heuristics.

6.3.6 Evaluate system

The trained system is now evaluated using the gold-standard test documents (§6.3.4). This step is crucial, as without it there is no way to identify areas for improvement.

6.3.6.1 Categorize test data

The system is used to generate categorizations for the test documents. Every categorization includes a *confidence level* based on correlation weights of the relevant terms. If the confidence level is higher than the *confidence threshold* for that category, the document is assigned to the category.

A high confidence level indicates that the result is strongly supported by training data, while a low confidence level indicates that the result is only weakly supported by training data. Confidence level should not be confused with probability. Just because a system returns a confidence level of 95% does not mean the result is 95% likely to be correct.

6.3.6.2 Evaluate correctness

The categorizations of the test documents are compared with the gold-standard categorizations to evaluate *correctness*[4]. In general, correctness indicates the degree to which the results of an automated system agree with the "correct" results, however that is determined. In the case of document categorization, correctness reflects the degree to which the automated category assignments match the categorizations produced by IC experts.

---

[4] This quality is more often called *accuracy*, but that term has another meaning in the medical community.

Some common measures of correctness are: *precision* (or *specificity*), the percentage of returned results that are correct; *recall* (or *sensitivity*), the percentage of possible results that are correctly returned; and *F-score*, a weighted average of precision and recall. All measures can be viewed for the system as a whole and for each specific category.

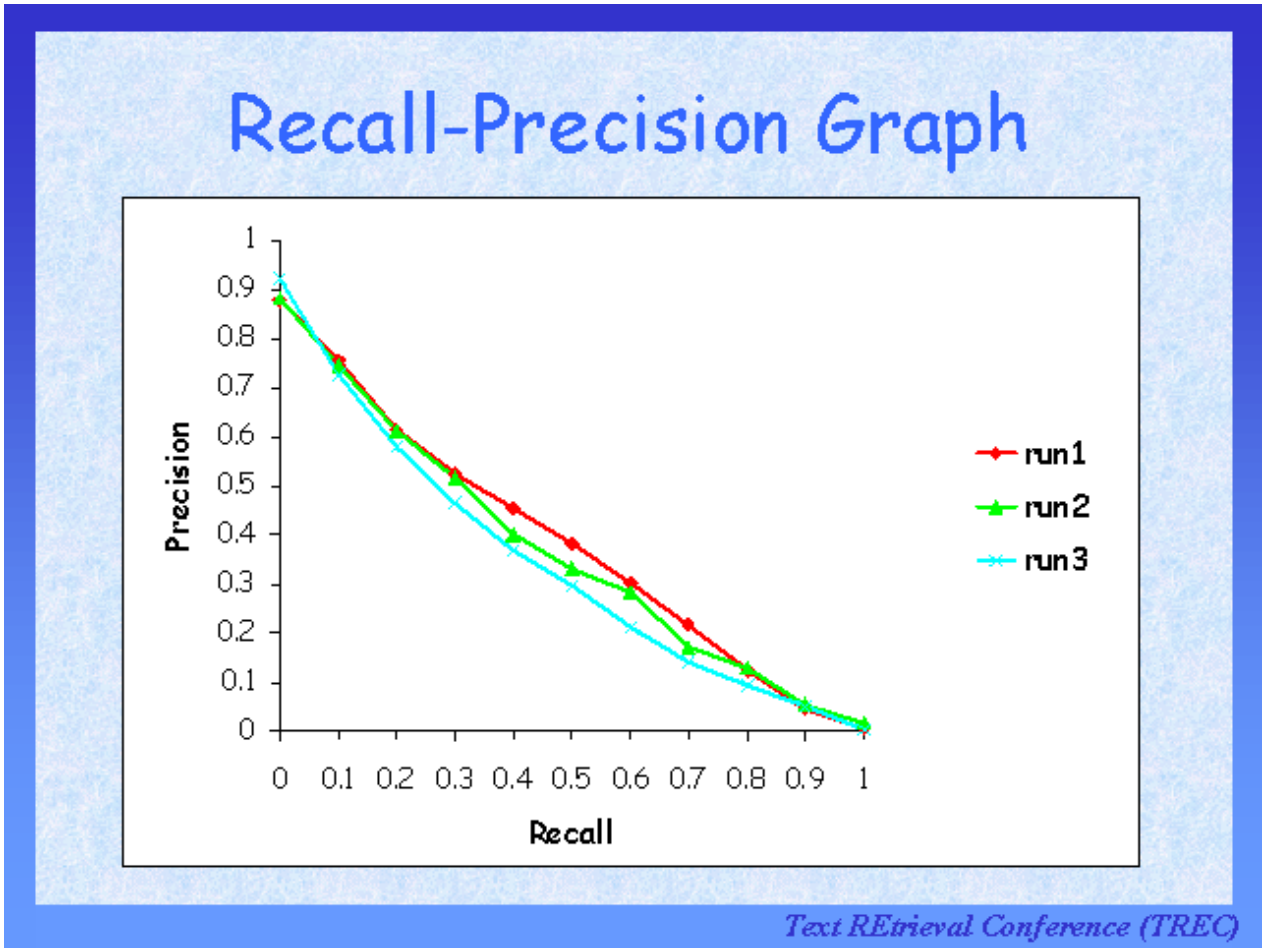
Precision can also be understood as a reflection of *false positives*, i.e., documents that are incorrectly assigned to a given category. Recall likewise reflects *false negatives*, i.e., documents that should have been assigned to a given category, but were not.

There is generally a trade-off between precision and recall, i.e., an operator can readily improve one at the expense of the other. The most common means for doing this is by tweaking confidence thresholds. In the extreme cases, a confidence threshold of 0 provides perfect recall but (typically) extremely poor precision, while a confidence threshold of 1 provides perfect precision but (typically) extremely poor recall.

A common way to display correctness is a *precision-recall graph*, depicting the recall that is achieved for a given precision and vice-versa. In document categorization, a separate precision-recall graph is typically provided for each category.

A sample precision-recall (or recall-precision) graph, taken from the online proceedings of an information retrieval conference (11), is shown in Figure 1. Note that the graphs for all three trials show a precision close to 0 when recall is 1, and a precision close to 1 when recall is 0.

Figure 1: Precision-Recall Graph Example



6.3.6.3 Evaluate confusion

A *confusion matrix* illustrates the degree to which various pairs of categories are confused with one another, i.e., category A is returned when category B should have been or vice versa. High confusion between two categories is often a clue that their definitions can be improved for better results (§10.1).

A sample confusion matrix is shown in Table 1. In this example, the system returned DOG nine times. Of these nine results, 3 should have been identified as CAT, 5 were truly DOG, 1 was RABBIT, and none were CHICKEN.

Correct results are shown on the diagonal. The overall precision (same as recall in this example) can be calculated as the sum over the diagonal (Correct) divided by the sum over all results (Total), or: *19 / 37 = 51.4%*.

Table 1: Confusion Matrix

		Actual			
		CAT	DOG	RABBIT	CHICKEN
Predicted	CAT	**5**	4	2	1
	DOG	*3*	***5***	*1*	*0*
	RABBIT	1	3	**4**	1
	CHICKEN	1	0	1	**5**

Correct	*19*
Total	*37*
Precision	**51.4%**

There is a high confusion rate between CAT and DOG in this example. If this distinction is crucial to the business application, the analysis reveals a need for more work in distinguishing these categories. However, if business considerations permit, a simpler solution is simply to merge the two categories. Assuming this does not change any of the system's results, this would produce the confusion matrix in

Table 2, with a precision of: *26 / 37 = 70.3%*.

Table 2: Confusion Matrix – Merged

		Actual		
		CAT/DOG	RABBIT	CHICKEN
Predicted	CAT/DOG	**17**	3	1
	RABBIT	4	**4**	1
	CHICKEN	1	1	**5**

Correct	*26*
Total	*37*
Precision	**70.3%**

In practice, however, merging similar categories often leads to better results with the combined category. A moderate improvement might produce the revised outcome shown in Table 3, with a precision of: *28 / 37 = 75.7%*.

Table 3: Confusion Matrix – Improved

		Actual		
		CAT/DOG	RABBIT	CHICKEN
Predicted	CAT/DOG	**19**	3	1
	RABBIT	3	**4**	1
	CHICKEN	0	1	**5**

Correct	*28*
Total	*37*
Precision	**75.7%**

Of course, there is no way to know for sure whether merging categories will be advantageous until it is tried and the results are evaluated. Merging CAT and DOG might instead have led to the results in Table 4, with a precision of: *18 / 37 = 48.6%*.

Table 4: Confusion Matrix – Worsened

		Actual		
		CAT/DOG	RABBIT	CHICKEN
Predicted	CAT/DOG	**16**	7	5
	RABBIT	4	**1**	1
	CHICKEN	2	0	**1**

Correct	*18*
Total	*37*
Precision	**48.6%**

In this case, we would know to go back to the original categorization and try something else.

6.3.6.4 Monitoring

A well-designed statistical system continually monitors correctness of test results, providing instant feedback any time a change is made.

6.3.7 Elicit expert review

It is often useful for the system to flag certain categorizations for review by a human expert. Obvious candidates are test documents that the system got wrong. The expert can review these to see whether the system was truly wrong, or whether the documents were originally miscategorized.

Other candidates for review are novel documents where the system returned a low confidence level. The expert can review these results, and then feed the correct results back to the system as new gold-standard training data.

Generally it is helpful to have more than one expert independently review each flagged document. This provides an additional level of assurance in the results, and provides the information needed to measure *inter-annotator agreement*, the degree to which experts concur with each other's results (§10.1).

The current validity process (§5.3.7) is an excellent starting point. The process could be enhanced by using statistical criteria in selecting documents for review, and by having multiple experts categorize each document (§10.1).

6.3.8 Retrain using results of expert review

Novel documents that have been reviewed by an expert (and for which there is full inter-annotator agreement) are fed back into the system for further training. The system thus improves continually through learning, becoming ever more reliable as time goes on. For this feedback cycle to be most effective, the system must be capable of *negative training*, i.e., learning from examples of documents that do *not* belong in a specified category.

Many statistical systems use *freshness* as a weighting criterion. Training examples that are added recently are given a higher weighting than examples that were added a long time ago. This allows the system to adapt automatically to changing conditions, as newer evidence supersedes the old.

7 Evaluation Methodology

This section briefly summarizes the methodology followed in this study for evaluating *document categorization* platforms.

7.1 Evaluation Criteria

Initial evaluation criteria (§A) were derived from the formal work plan, which included lists of mandatory and desired characteristics (1). These were supplemented based on the reviewers' knowledge and experience, and then refined through interactions with NIH representatives, including a formal proposal review. Based on these interactions, each criterion was assigned an associated weight between 1 (optional) and 10 (vital).

Criteria were clustered into related groups and subgroups in order to assure that the overall evaluation did not overemphasize one area. The criteria groups were weighted for the overall evaluation as shown in Table 5. The criteria making up each group are listed in Table 6 through Table 9. Descriptions of key criteria are in Appendix §A.

Table 5: Criteria Groups

	Weight
Total	5
Business	1
Technical	1
Functionality	2
Usability	1

Table 6: Business Criteria (§A.1)

Group	Subgroup	Criterion	Weight
Business	Vendor	Market Presence	3
Business	Vendor	Corporate Strategy	3
Business	Vendor	Tech Leader	2
Business	Vendor	US-Based	1
Business	Vendor	Customer Service	3
Business	Control	Non-proprietary	1
Business	Control	Internal Control	4
Business	Initial Cost	Initial License	3
Business	Initial Cost	Development	2
Business	Initial Cost	Initial Knowledge Acquisition	3
Business	Initial Cost	Status Quo	5
Business	Maintenance Cost	License Maintenance	5
Business	Maintenance Cost	Domain Expert Time	10
Business	Maintenance Cost	Internal Support	2
Business	Maintenance Cost	Consulting	2

Table 7: Technical Criteria (§A.2)

Group	Subgroup	Criterion	Wt
Technical	Interoperability	Oracle	10
Technical	Interoperability	Java	10
Technical	Interoperability	Standards-based	2
Technical	Capacity	Scalability	10

Table 8: Functionality Criteria (§A.3)

Group	Subgroup	Criterion	Wt
Functionality	Correctness	Correctness	10
Functionality	Correctness	Smoothing	2
Functionality	Correctness	Evaluation Mechanism	5
Functionality	Correctness	Inter-Annotator Agreement	1
Functionality	Concept Extraction	Text Normalization	5
Functionality	Concept Extraction	Thesaurus-Based	5
Functionality	Concept Extraction	Noun Phrases	3
Functionality	Concept Extraction	Synonyms	3
Functionality	Concept Extraction	Synonyms, Automated	2
Functionality	Concept Extraction	Disambiguation	3
Functionality	Concept Extraction	Disambiguation, POS Tagging	1
Functionality	Concept Extraction	Disambiguation, Automated	2
Functionality	Concept Extraction	Abbreviation Expansion	1
Functionality	Concept Extraction	Negative Detection	1
Functionality	Concept Extraction	Complex Noun Phrase Expansion	1
Functionality	Concept Extraction	Entity Extraction	1
Functionality	Categorization	Custom Categories	10
Functionality	Categorization	Category Hierarchy	3
Functionality	Categorization	Distinguish Text Fields	5
Functionality	Categorization	Novel Tags	1
Functionality	Categorization	Category Discovery	1
Functionality	Categorization	Custom Business Rules	1
Functionality	Categorization	Adjustable Parameters	1
Functionality	Categorization	Negative Training	3
Functionality	Related Functions	Enterprise Search	2

Table 9: Usability Criteria (§A.4)

Group	Subgroup	Criterion	Wt
Usability	Usability	Comprehensibility	10
Usability	Usability	Transparency	5
Usability	Usability	Graphics	1
Usability	Usability	Collaboration Tool	2
Usability	Usability	Documentation/Tutorial	1

7.2 Evaluation

Based on documentation, industry reviews, and technical presentations (§4), each candidate platform was evaluated for each criterion, receiving a score of –1 (bad), 0 (mixed or unknown), or 1 (good). The criteria within each group were combined according to their respective weights to generate a composite score for the group between 0 (bad) and 5 (good). The groups were then combined according to their weights to generate an overall candidate score, again ranging between 0 (bad) and 5 (good) (Table 10, §9).

8 Document Categorization Solutions

The core of RCDC is a *document categorization* platform, currently one produced by Collexis (§8.1.1). This type of software takes documents and assigns them automatically or semi-automatically into *categories*. Categories are generally taken from one or more existing lists, possibly serving different purposes. For example, an email document might be assigned the category JOB INQUIRY or the category SPAM. A grant proposal document might be assigned the category HYPERTENSION or the category CLINICAL TRIALS.

In the case of RCDC, the software takes grant proposal submissions and like documents, on the order of hundreds a day, and categorizes them into the 300+ canonical categories authorized by Congress.

The original plan for this study was to compare several commercial document categorization platforms (§8.1) and one or two *open-source* alternatives (§8.2), and to consider the possibility of developing a custom solution in-house (§8.3). As it turns out, there are no open-source candidates to compare, and no compelling reason to prefer a custom solution to the best commercial option.

Each of the top candidates is listed here, followed by a list of some significant pros and cons. Key criteria for all candidates are compared graphically in Table 11 (§9).

8.1 Commercial Categorization Platforms

This section describes results of reviewing the incumbent Collexis system. Several other candidates that were not among the top contenders are described briefly in Appendix B.

8.1.1 Collexis

8.1.1.1 Overview

Collexis is a niche software vendor in the area of information discovery and retrieval. Its customers are traditionally medical schools and research institutes, although it has begun to branch out to other public- and private-sector clients (12). Collexis was featured in Gartner's *Cool Vendors in Content Management, 2008* (13).

Collexis was acquired on June 10, 2010, by Elsevier, "the leading global publisher of scientific, technical, and medical information products and services" (14). This is not expected to have any direct impact on the RCDC system.

Annual maintenance costs for Collexis are an order of magnitude higher than for the most expensive of the top competitors.

8.1.1.2 Customer Relations

Most of Collexis' customers are in the medical community (12). All four of its advertized products are geared toward that community. One of them, Grant Management Solution, is a repackaging of RCDC (15). A personal inquiry into their product line elicited the response that they are in the business of providing consulting services, not products, citing RCDC as an example (16).

Information on customer satisfaction is not readily available. Collexis does not provide customer recommendations on their website.

8.1.1.3 Technology

The Collexis implementation categorizes *projects*, e.g., grant proposal documents, based on a *fingerprint*, i.e., a content profile determined by the occurrence of keywords (a.k.a. *terms*) related

to *concepts* in a medical thesaurus. This fingerprint is compared to the fingerprints of 300+ canonical *categories*, defined by Congress and updated annually (§5).

Collexis simulates the behavior of a statistical system using *if-then* rules, each rule carrying a weight assigned by IC experts (§5.2.5). For each document, the weights of the applicable rules are combined to produce weighted results, e.g., suggesting category CANCER with a weight of 47 and category AGING with a weight of 32 (§5.3.5).

The Collexis methodology places high demands on the time of IC experts and RCDC staff. Some of these activities are unavoidable, and will remain part of the methodology regardless of what platform is adopted. These inevitable activities include:

1. *Provide a workable definition for each category* (§5.2.4).

   Consensus on category definitions is a prerequisite for effective categorization (§10.1).

2. *Define business rules* (§5.3.6).

   The system can infer business rules to a large extent, but there will always be policy-driven rules that take precedence over any automated inference.

3. *Perform validity checking on categorization results* (§5.3.7).

   The medical community would be unlikely to accept the validity of an automated system without human checking. Furthermore, categorizations provided by humans generate the information needed to continually improve a statistical system (§6.3.8).

Other activities would be performed automatically if a statistically-based system were used. Such activities include:

1. *Subjectively propose new concepts* (IC experts) *and evaluate the usefulness and impact of each proposed addition* (RCDC staff) (§5.2.3).

   A statistical system identifies meaningful terms automatically and reliably by observing

statistical correlations (§6.3.5.1). The manual approach can easily overlook a concept that could have provided useful information, or introduce a concept that skews results in unforeseen ways.

2. *Subjectively assign a weight for each concept within each category* (§5.2.5).

   Concept weights are designed to mimic probabilistic correlation weights (§6.3.5.2). Humans are notoriously bad at estimating such correlations (17). Statistical systems do so precisely and reliably.

3. *Debug incorrect results* (§5.3.8).

   In a well-designed statistical system, any time a human expert determines that a categorization result is incorrect, this information is passed back to the system, which automatically infers what changes need to be made (§6.3.8). The Collexis methodology, in contrast, requires intense debugging effort on the part of IC experts, with no guarantee of generating an improvement.

One distinctive feature of the Collexis methodology is the intermediate mapping of terms to concepts (§5.3.2). There is no way to measure how much this affects correctness, but it is unlikely to be significant, and could easily be negative. The information gained by mapping to thesaurus entries tends to be offset by the information lost in discarding individual terms, particularly if many terms do not occur in the thesaurus.

Mapping to concepts could potentially help in resolving *synonymy* and *polysemy* (§5.2.3). For a rule-based system, resolving synonymy would not change the outcome, but could reduce the number of rules – if A, B, and C are identified as synonyms, a single inference rule can take the place of three. For a statistical system, resolving synonyms could marginally help or hurt.

Polysemy could be a significant problem for a rule-based system. If, say, a grant proposal described an `experimental vision`, but this was incorrectly interpreted as relating to the concept of `VISION`, the proposal could be miscategorized. Collexis provides a mechanism for writing rules to differentiate meanings by context, but this feature is currently not used (§5.3.2). Identifying and implementing all such rules would be a momentous task; there could easily be millions of such rules, assuming thousands of base terms and thousands of context terms.

Polysemy is also a problem in a statistical system. However, it should be far less significant than for a rule-based system, since a term's context is always taken into consideration (§6.3.5.1). Of course, polysemy is only a problem to the extent that it exists in the documents. There is currently no way to measure this.

The main advantage of mapping to concepts is the sense of transparency it provides for the human expert. An expert may readily agree that the concept of `DRUG EFFICACY` suggests the category of `CLINICAL TRIALS`, but may not so readily accept that the normalized phrase `efficac~ of the drug~` implies the same category. Likewise, if a statistical system infers that the normalized phrase `offer~ a choic~` also suggests `CLINICAL TRIALS`, the expert may balk at the scientific defensibility of this inference, despite the empirical evidence.

Another distinctive feature of Collexis is reordering of terms (§5.3.1). Again, there is currently no way to measure the impact of this feature on correctness. However, it appears at least as likely to be negative as positive, as the example of `vision research` vs. `research vision` illustrates. In general, reordering of nouns in a noun phrase does, in fact, change the meaning. Offering the Curator the ability to configure this behavior for every possible noun pair is again an intractable burden.

The incumbent RCDC implementation uses Collexis version 6.5, which is in the process of being upgraded to version 7.0. This will introduce various new features, such as allowing rules to disambiguate terms according to part of speech, and expanding complex noun phrases, e.g., `heart, lung and liver transplants` (§5.3.1).

8.1.1.4 Functionality

Crucially, Collexis offers no means for evaluating its *correctness* (§A.3.1.1), either overall or by category (§6.3.6). Aside from anecdotal observation, the only way to determine how well the system is categorizing documents is through an extensive study along the lines of the 2009 IDA report (9). Statistical systems automatically provide this capability (§6.3.6).

Furthermore, the current validity checking methodology (§5.3.7) only identifies *false positives*, i.e., documents that are incorrectly assigned to a given category. It does not identify *false negatives*, i.e., documents that should have been assigned to a given category, but were not.

The strongest factor in favor of retaining Collexis is its familiarity to RCDC domain experts. Adopting any competitor would require experts to learn a new system, and otherwise disturb the status quo (§A.1.3.3).

8.1.1.5 Integration

Collexis is designed as a standalone product. It can be integrated with other products using a Java API, but no integration is provided out-of-the-box.

8.1.1.6 Evaluation Summary

PROS:

- No disruption to status quo (§A.1.3.3).
- Thesaurus-based (§A.3.2.1).
- Nominally more comprehensible than statistical system (§A.4.1.1).

CONS:

- Extremely high demand on IC domain experts (§A.1.4.2).
- Extremely high annual licensing fee (§A.1.4.1).
- Inconsistent correctness (§A.3.1.1).
- No means for evaluating correctness (§A.3.1.2).

8.1.2 Collexis Best-Fit

In 2009, NIH and Collexis personnel produced an alternative RCDC implementation, termed the "best-fit model", which creates category fingerprints (§5.2.6) statistically rather than manually. So far, the best-fit system has only been tried in a few limited settings, primarily with users interested in querying for novel, ad hoc topics, less interested in creating or evaluating the category fingerprint. Users have generally been satisfied with the results, although there is some concern about "overfitting," i.e., skewing the analysis too strongly toward the specific training documents used, rather than generalizing appropriately.

8.1.2.1 Evaluation Summary

PROS:

- Thesaurus-based (§A.3.2.1).
- Lower demand on IC domain experts (§A.1.4.2).
- Anticipated better correctness (§A.3.1.1).

CONS:

- Extremely high annual licensing fee (§A.1.4.1).
- No means for evaluating correctness (§A.3.1.2).

8.1.3 Recommind Decisiv Categorization[5]

8.1.3.1 Overview

Recommind[6] was established in 2000 to spin off academic research in text analytics. It was a niche vendor focused on *e-discovery* and document archiving, but has recently broadened its market to more general information access applications such as enterprise search (18). Its current business is roughly half e-discovery, half information access (including search and categorization). Its business model emphasizes rapid but manageable organic growth, converting each customer into a positive reference (19).

Unlike most of its competitors, Recommind's core technology is content categorization, embodied in its flagship product, Decisiv Categorization (20). Other product offerings, such as e-discovery and *enterprise search*, are built on this core technology (21).

Gartner's *Magic Quadrant for Information Access Technology* for 2009 identifies Recommind as a Leader, i.e., excelling in both vision and ability to execute, up from Visionary the previous year. The reviewer observes (correctly) that Recommind "is unusually accomplished at using statistical analysis to extract meaning from documents in an automated fashion" (22).

8.1.3.2 Customer Relations

The Office of Energy Efficiency and Renewable Energy (EERE), a division of the U.S. Department of Energy (DOE), recently purchased a Recommind solution to replace an existing Autonomy (§8.1.6) installation (23). They first performed a head-to-head comparison of several tools, including IBM's InfoSphere Classification Module (§8.1.4). So far, they are pleased with

---

[5] Recommind appears to have recently renamed their flagship product line from MindServer to Decisiv. The latter name is used in this report, although the evaluation took place on the former product.

[6] Recommind is pronounced with a long 'i'.

their decision. An EERE representative indicated that it took about six months elapsed time from contract to production, including integration with Xythos Enterprise Document Management Suite (24). They invite any government agency representatives to come see their installed system and learn from their experience (25).

Other notable Recommind customers include Verizon, DuPont, Nationwide, WilmerHale, Nixon Peabody, and business.gov.au (26).

Subjectively, the Recommind customer service staff was the most responsive in providing information to the study team.

8.1.3.3 Technology

Recommind includes all the standard elements of a statistical categorization platform (§6). Its core technology is a patented machine-learning algorithm known as Probabilistic Latent Semantic Analysis (PLSA) (27). PLSA analyzes a collection of unstructured documents to infer concepts and clusters of related concepts (28). The underlying algorithm is a *support-vector machine* (SVM), the state-of-the-art for document categorization.

Recommind Decisiv Search maps terms to concepts within each document. Unfortunately, this functionality is not integrated into the categorization tool; categorization is based on terms, not concepts. The mapping to concepts could be integrated into the categorization interface, but this has not been implemented. The search tool can be used to generate a category structure from unlabeled documents, or suggest improvements to an existing structure (29).

8.1.3.4 Functionality

Recommind's focus on document categorization is reflected in the richness of Decisiv's functionality. By default, it continually monitors correctness of analysis within each category.

Uniquely among the products we reviewed, by default it also flags training documents within each category that appear to be mislabeled (19).

The training engine is highly configurable. In particular, it can weight based on the *freshness* and other document properties (§6.3.8), and it can differentially weight document sections (e.g., title). By default, it considers phrases up to four words in length, although this can be configured. Like most statistical tools, Decisiv Categorization can also process business rules, feeding the results of rules-based analysis into training for statistical analysis (30).

The default user interface (UI) is highly intuitive, reminiscent of an Internet browser. It is also easily extensible using Java and XML.

8.1.3.5 Integration

Decisiv Categorization interfaces directly with all major languages and databases, including Java and Oracle. Recommind documentation claims (and customer reports appear to confirm) that it is fully parallelizable, scalable to terabytes of data and millions of users (30).

8.1.3.6 Evaluation Summary

PROS:

- Automated: Minimal demand on domain experts (§A.1.4.2).
- Maximum expected correctness (§A.3.1.1).
- Default monitoring of correctness (§A.3.1.2).
- Intuitive user interface (§A.4.1.1).
- Comparatively inexpensive (§A.1.4.1).
- Reputation for excellent customer support.

CONS:

- Disruption to status quo (§A.1.3.3).

- Not thesaurus-based (§A.3.2.1).
- Nominally less comprehensible than statistical system (§A.4.1.1).

8.1.4 IBM InfoSphere Classification Module

8.1.4.1 Overview

IBM's InfoSphere Classification Module (ISCM) can be acquired on its own or as an integrated component of various IBM content management platforms (31). The individual product cost is low, $1630 per processor for 12 months, including support; or comparable figures based on number of users (32).

Gartner's *Magic Quadrant for Information Access Technology* 2009 identified IBM (not specifically IBM InfoSphere) as a Challenger – high ability to execute, but lacking in vision – down from Leader the previous year. Gartner cites IBM's shift in focus to a narrow list of business problems as one reason for the downgrade.

8.1.4.2 Customer Relations

IBM provides phenomenal ISCM product documentation and tutorials on their website (33). ISCM does not have its own user group, although several user groups exist for products with which it can be integrated (34). Information on customer satisfaction was not readily available.

8.1.4.3 Technology

ISCM, like Collexis, simulates the behavior of a statistical system using weighted rules. Unlike Collexis, ISCM uses statistical analysis to generate the weighted rules from categorized documents. The derived system therefore has much in common with a purely statistical system, with the weights on the results loosely corresponding to the likelihoods on statistical results (35).

8.1.4.4 Functionality

ISCM does not differentiate document sections. An integrated product, Quark XML Author, can provide this function, but only within the context of a content management platform (36). Alternatively, the feature can be programmed through the Java API (37).

ISCM does not map concepts to a thesaurus. However, it does differentiate homonyms based on context, and it does track and display which terms and phrases contribute to a given document categorization, both positively and negatively. NLP algorithms and heuristics are also configurable (37).

The ISCM Classification Workbench monitors correctness by default, and provides numerous analytic tools for improving system design, including precision-recall graphs by category (38). Corrections to incorrect categorizations are fed back into the system to improve future performance. The feedback can either be immediate or deferred (37), allowing a data steward to review feedback before integrating it into training. By default, the system weights training documents based on *freshness* (§6.3.8) (39).

8.1.4.5 Integration

ISCM appears to offer the best integration potential of all the candidates. Although it can be acquired separately, it is designed to integrate with any of IBM's full-featured content-management platforms. It is highly customizable, with complete APIs for several programming languages including Java, and excellent documentation. ISCM does not provide a database interface, but the Java API can be used to interface with Oracle indirectly. ISCM can also import documents from XML or standard file formats (40).

8.1.4.6 Evaluation Summary

This option considers acquiring ISCM as a standalone product, without the features built into a content management platform (§8.1.5). Notably, differentiation of document sections and integration with Oracle would not be available out-of-the-box, although they could be programmed through the Java API.

PROS:

- Automated: Minimal demand on domain experts (§A.1.4.2).
- High expected correctness (§A.3.1.1).
- Default monitoring of correctness (§A.3.1.2).
- Negligible licensing cost (§A.1.4.1).

CONS:

- Disruption to status quo (§A.1.3.3).
- Not thesaurus-based (§A.3.2.1).
- Nominally less comprehensible than statistical system (§A.4.1.1).
- Missing out-of-the-box differentiation of document sections (§A.3.3.3).
- Missing out-of-the-box integration with Oracle (§A.2.1.1).

8.1.5 Classification Module with IBM FileNet P8

ISCM can also be acquired in conjunction with an IBM content management solution, e.g., FileNet P8. This would provide out-of-the-box integration with Oracle, and differentiation of document sections with the addition of the integrated Quark product (§8.1.4.4).

8.1.5.1 Evaluation Summary

PROS:

- Automated: Minimal demand on domain experts (§A.1.4.2).

- High expected correctness (§A.3.1.1).
- Default monitoring of correctness (§A.3.1.2).
- Inexpensive compared to Collexis (§A.1.4.1).
- Best out-of-the-box functionality, little or no configuration.

CONS:

- Disruption to status quo (§A.1.3.3).
- Not thesaurus-based (§A.3.2.1).
- Nominally less comprehensible than statistical system (§A.4.1.1).

8.1.6 Autonomy IDOL

8.1.6.1 Overview

Autonomy's Intelligent Data Operating Layer (IDOL) content management platform tops Gartner and Forrester reviews in a host of areas, including enterprise search (18), information access (22), and e-discovery (41). Its market role is so pivotal that Gartner dedicated an entire report just to IDOL (42). Forrester terms IDOL "the most complete product evaluated with the best core technology architecture and security capabilities" (18). Gartner says it "is nearly unmatched in its breadth and extensibility of potential function" (42). Automatic Categorization, Taxonomy Generation, and the Autonomy Collaborative Classifier are among a huge array of IDOL add-on modules.

Functionality comes at a cost, however. Autonomy is the only leading vendor (excluding Collexis) not to offer a fixed pricing structure, and tends to have among the highest prices. Gartner warns that a company should plan on three full-time equivalents to manage the installed platform (42).

8.1.6.2 Customer Relations

According to Gartner, many customers reported negative experiences with sales and support, including a pattern of fading attention once the sale is consummated. Satisfied customers tended to be those with multimillion dollar contracts lasting several years (42). The sales representative tells us this reputation dates back to a time when Autonomy was leaner, focused on pre-sales, and dependent on partners for support services. Autonomy gained a new services arm with its Verity acquisition in 2005, and the sales representative, a vice-president, assures us he has made service to customers a top priority, including those at the lower end of the financial spectrum (43). Subjectively, Autonomy was less responsive than other companies in providing information regarding their system.

8.1.6.3 Technology

Autonomy IDOL's underlying algorithm is Bayesian inference (44). Bayesian techniques have been used extensively in machine learning because they are easy to implement and often produce reasonable results. For document categorization they have been largely overshadowed by SVMs for the past decade (45).

Automony claims that IDOL's core technology is Meaning Based Computing (MBC). What this means technically is unclear, and apparently unknown to the sales staff, but documentation seems to imply conceptual analysis involving both structured and unstructured information (46).

IDOL Automatic Categorization is intrinsically based on conceptual analysis. Documents are assigned to categories based on a weighted analysis of identified concepts and entities. The weights are initialized automatically based on analysis of example documents, but they can also be adjusted manually. While the system can incorporate negative examples, this tends to produce undesirable results, as is typical with Bayesian inference (43).

8.1.6.4 Functionality

The standard user portal comes with a variety of analytic displays. All analytic data is also exported as XML, allowing development of any desired presentation interface. Oddly, the system does not include an out-of-the-box mechanism for evaluating correctness, but analysis of correctness can be performed on the exported XML data (43). Indeed, there was a sense in the demo that the UI was deliberately left unfinished, under the philosophy that as long as the relevant data is exposed, each customer can design a UI according to its own needs.

The Autonomy Collaborative Classifier enables geographically disbursed domain experts to edit taxonomy definitions. Access controls define in which sections specific users can either suggest or finalize changes. There is no intrinsic model for handling disagreement among experts, however.

8.1.6.5 Integration

Categorization tools sit on top of the IDOL core. All of IDOL's extensive content management features therefore come with the package, including the industry leader enterprise search platform (18). The possibilities for leveraging the tool in other applications are virtually endless.

Autonomy actively seeks to integrate with other tools and to partner with other software vendors (22). Its standard user portal is written in Java/JSP, and it provides an API for all the most popular programming languages. Likewise, IDOL integrates easily with a variety of databases, including Oracle 11g.

8.1.6.6 Evaluation Summary

PROS:

- Automated: Minimal demand on domain experts (§A.1.4.2).

- High expected correctness (§A.3.1.1).
- Inexpensive compared to Collexis (§A.1.4.1).
- Highly flexible and extensible.
- Market leader.

CONS:

- Disruption to status quo (§A.1.3.3).
- Not thesaurus-based (§A.3.2.1).
- No intrinsic evaluation mechanism (§A.3.1.2).
- Nominally less comprehensible than statistical system (§A.4.1.1).
- Relatively expensive compared to other contenders (§A.1.4.1).
- Cumbersome installation and maintenance.
- Reputation for poor customer support.

8.2 Open-Source Tools

Open-source[7] software is licensed and distributed in such a manner that users can view and adapt the source code for their own use. Open-source software is often free, but not always.

A search for open-source document categorization tools did not turn up any candidates.

8.3 Custom Solution

A custom in-house solution would offer maximum flexibility for integrating cutting-edge academic research with the perfectly customized user interface, while maintaining complete internal control over the product. However, any improvement in correctness over Recommind

[7] This is "open-source" in the software sense. The term has an entirely different meaning in the intelligence community.

Decisiv Categorization (§8.1.3) would likely be minimal, as the state-of-the-art in categorization has not advanced significantly beyond Recommind's core PLSA technology (§8.1.3.3). Likewise, the customizable Recommind user interface is nearly as versatile as developing an interface from scratch. Internal control does not appear to be a high priority for NIH, and is outweighed by the potential cost of maintaining software development staff.

PROS:

- Maximum expected correctness (§A.3.1.1).
- Maximum flexibility.

CONS:

- High development and maintenance overhead.
- Inability to benefit from vendor support.

9 Platform Comparison

This section addresses the first purpose of the study: (1) identify and compare commercial alternatives to the Collexis implementation.

A top-level comparison of the main contenders for document categorization platform is provided in Table 10, with analysis broken down by criteria group (§7.1). A breakdown by key criteria is provided in Table 11. This includes only those criteria with a weight of 5 or more where not all contenders received the same rating[8].

Table 10: Platform Comparison by Criteria Group (§7.1)

	Weight	Collexis (8.1.1)	Collexis Best-fit (§8.1.2)	Recommind Decisiv (§8.1.3)	Autonomy IDOL (§8.1.6)	IBM ISCM (§8.1.4)	IBM P8 (§8.1.5)
Total	5	3.53	3.45	**4.08**	3.29	3.56	3.75
Business	1	1.79	2.55	**3.67**	3.01	**3.67**	3.27
Technical	1	**4.69**	**4.69**	**4.69**	**4.69**	3.91	**4.69**
Functionality	2	3.42	3.49	**4.18**	3.32	3.39	3.70
Usability	1	**4.34**	3.03	3.68	2.11	3.42	3.42

[8] Criteria for which all top contenders received the same rating include: Java (§A.2.1.2), Scalability (§A.2.2.1), Custom Categories (§A.3.3.1), and Transparency (§A.4.1.2).

Table 11: Platform Comparison by Key Criteria (§A)

	Wt	Collexis	Collexis Best-fit	Recommind Decisiv Categorization	Autonomy IDOL	IBM ISCM	IBM P8
Status Quo	5	**1**	0	*-1*	*-1*	*-1*	*-1*
License Maintenance	5	*-1*	*-1*	**1**	0	**1**	0
Domain Expert Time	10	*-1*	**1**	**1**	**1**	**1**	**1**
Oracle	10	**1**	**1**	**1**	**1**	0	**1**
Correctness	10	0	**1**	**1**	**1**	**1**	**1**
Evaluation Mechanism	5	*-1*	*-1*	**1**	*-1*	**1**	**1**
Thesaurus-Based	5	**1**	**1**	*-1*	*-1*	*-1*	*-1*
Distinguish Text Fields	5	**1**	**1**	**1**	**1**	0	**1**
Comprehensibility	10	**1**	0	0	*-1*	0	0

The overall ratings are resilient to moderate changes in weightings and assessments. It would be possible, for example, to make Collexis come out ahead of Recommind by raising the weight of Usability from 1 to 2 and raising the weight of Comprehensibility from 10 to 50. A detailed uncertainty analysis is shown in Appendix C.

9.1 Findings

The analysis implies two findings:

Finding 1: Under most weightings, alternatives to Collexis score higher than does the incumbent system.

Of the five alternatives to Collexis considered in detail (including the Collexis best-fit alternative) three (Recommind, IBM ISCM, and IBM P8) score higher than does the incumbent system using the weightings in Table 10. The alternatives all score substantially higher than does Collexis on the Business category, scored equivalently (except for IBM ISCM) on the Technical category, and scored equal to or higher than Collexis on the Functionality category.

Turning to the individual decision criteria in Table 11, Collexis, as the incumbent system, scored highest on the Status Quo criterion. Collexis also scored highest on the Thesaurus and Comprehensibility criteria. Alternatives scored at least as high as Collexis on the other six criteria, with all three other systems being superior to Collexis on the License Maintenance, Domain Expert Time, Correctness, and Evaluation Mechanism.

Part of the difference in scoring lies in the choice of an expert-driven as opposed to a statistically-based system. There are known concerns with the Collexis system that have been spelled out in this report. The degree to which an automated system will raise concerns in terms of its acceptability across NIH could not be determined.

Finding 2: Of the challengers, Recommind Decisiv Categorization scored highest.

Recommind Decisiv Categorization (§8.1.3) scores highest overall and in every category except Usability (Table 10). The latter assessment is based on the assumption that configuring concept weights manually is more comprehensible than doing so automatically (§A.4.1.1).

IBM may also provide a viable alternative to Collexis. The correctness of its results should be at least as good as those offered by Collexis. Its analytic tools are excellent, and it is highly configurable.

If IBM is selected, it would merit closer investigation to determine whether to adopt FileNet P8 or another content management solution (§8.1.5), or acquire ISCM as a standalone product (§8.1.4). One consideration would be how difficult it is to use the Java API to introduce key functionality, such as distinguishing text fields or interfacing with Oracle. Another consideration would be the level of demand for spin-off applications such as enterprise search..

10 Other Technical Factors

10.1 Category Reform

Much of the inaccuracy of the incumbent RCDC system would appear to have nothing to do with shortcomings in the implementation, but rather in the enumeration/definition of the categories. A key indicator of this problem is disagreement among domain experts over categorizations. If the experts do not agree with one another, the system cannot be expected to perform any better than they do. The degree to which human experts agree with one another is called *inter-annotator agreement*.

While inter-annotator agreement among RCDC experts has not been analyzed, there is anecdotal reason to believe it is lower than it should be. One potential explanation is that experts from different ICs may use terms in different ways.

RCDC is not unusual in this. Some studies suggest that, in general, experts disagree on document categorization half the time (28). For this reason, research on categorization typically includes an analysis of inter-annotator agreement. In general, we want to frame the problem in such a manner as to facilitate high agreement.

Typically in RCDC, only one expert evaluates the correctness of any given categorization (§5.3.7). The motivation for this practice in terms of efficiency is obvious. However, a formal evaluation of inter-annotator agreement requires a large number of categorizations that have been evaluated by two or more experts.

Addressing this shortcoming is largely independent of the choice of platform. Ultimately, the goal is to produce a set of well-defined categories to which human experts can agree with close to 100% consistency, but which also map unambiguously to the canonical set.

Of course, the challenge is not merely technical. Achieving 100% agreement on anything, in any group of people, is not easy.

10.1.1 Study 1: Perform a Study of Inter-Annotator Agreement

The first step in evaluating the current category structure would be to perform a study of inter-annotator agreement. Ideally, a full study would cover all 300+ categories; draw a random sample of 20 or more documents purportedly from each category; and ask three or more experts to categorize each document. This should provide enough evidence to determine the degree of agreement, and identify particularly troublesome categories.

A smaller initial study could be performed to determine whether the larger study is warranted; say, with 200 documents drawn at random, each evaluated by three experts.

10.1.2 Study 2: Revise Category Taxonomy

If significant inter-annotator confusion is discovered, the next task is to revise the category structure to produce better-defined categories. There are two broad approaches to doing this. One approach is to have a panel of experts sit down with the results of Study 1, consider the categories where there was the most confusion, and try to come up with better alternatives. Another approach is to let a statistical system perform automated *clustering* of the documents (§6.3.1), and see how the results map to the existing taxonomy.

Ideally, it will be discovered that each ambiguous category is simply the concatenation of two or more distinct subcategories, e.g., `GLAUCOMA` and `OPTIC NEUROPATHY`. In this case, the system can divide the larger category into its subcategories for analysis, and then recombine the results for reporting. The system could thus be optimized without affecting the canonical category list.

However, the results may not be so neat. For instance, it may be determined that there are, in fact, two or more overlapping taxonomies, categorizing the same documents using different sets of criteria. It may be discovered that large, overarching categories such as `CLINICAL RESEARCH` need to be treated separately from the main taxonomy(-ies). Such observations will require discernment.

Almost certainly, it will be beneficial to define which categories can be overlapping, and which are mutually exclusive. Likewise, it will most likely be beneficial to redefine the taxonomy such that a parent category is exhaustively defined by its subcategories (§5.2.4).

This study may prompt a petition to Congress for significant category revisions. If so, the statistical clustering analysis can demonstrate the empirical justification for the request.

10.1.3 Practice: Monitor Categories

Long-term, it would be useful to adapt the validity process to measure inter-annotator agreement. Any shortcomings can then be addressed as discussed under Study 2.

For example, suppose validity testing currently involves 50 domain experts each checking 150 categorizations, for a total of 7500 assessments. It would be more useful if, instead, 2500 categorizations were each reviewed by three different experts. The 2500 examples could be chosen at random or designed to be representative; presumably any recently added categories would merit extra testing. Thus no additional expert time would be needed, but the information would then be available for an analyst to assess the category model and identify opportunities for enhancements. The overall sample size would still be more than adequate to evaluate the correctness of the system as a whole.

10.2 Transition Plan

If NIH chooses to transition away from Collexis, it is recommended that discussions be held with other organizations that have made a similar transition.

In transitioning to any new system, there may be a need to maintain backward compatibility with previous approaches. It is possible that the new system may lead to shifts in the number of awards assigned to each Congressionally-mandated category. The NIH Office of Budget (OB) has directed that any annual change over 5% in reported funding for any category must go through a formal acceptance process.

11 Related Applications

This section briefly discusses technologies related to document categorization and some possible applications.

11.1 Related Technologies

Document categorization is considered part of *information access* technology, an umbrella discipline encompassing all techniques used to manage and process large amounts of (mostly) unstructured data (22). Information access technologies share much of the theoretical basis, and often underlying code, with document categorization.

11.1.1 Search

Search technologies provide ad hoc access to (typically) unstructured information. Many vendors offer forms of *semantic search*, where a user is guided to content related to relevant *concepts*, rather than just searching for key terms. The basic algorithms for identifying concepts are similar to those used to categorize documents, and the two functions often depend on the same core code.

Many statistical search tools, including all of the candidates in this study, can be configured to tune results based on observed characteristics of the user running the search.

The primary market for search technology is *enterprise search*. Most of the major players in information access offer enterprise search, either as a core product or as an optional part of an integrated suite.

11.1.2 Content Management

Content management (also *information management*, *document management*) solutions keep track of all of an enterprise's information, including structured documents, emails, databases and the like. Such systems typically offer the ability to arrange information by category, stored

hierarchically in a file system or database. Content management systems are often integrated tightly with document categorization and enterprise search.

11.1.3 E-discovery

A recent driver in commercial use of document categorization is *e-discovery*, the process of indexing unstructured corporate information for discovery in civil litigation. Changes to the Federal Rules of Civil Procedure in 2006 imposed significant penalties on companies that do not properly store and manage electronic information, creating a new industrial mandate overnight. The 2008-2009 recession fueled a desire to bring e-discovery capabilities in-house (41). This has sparked entry of new vendors and provided a boost to veterans such as Recommind, Autonomy and IBM. Gartner predicts e-discovery software may still be a few years from mainstream adoption, which could lead to short-term challenges for vendors with an overly narrow focus (47-48).

11.2 Potential Applications Inside NIH

- Enterprise search
- Enterprise content management
- Email routing
- E-discovery
- Identifying trends

11.3 Potential External Applications

- Searching for medical experts with specific expertise
- Enhancing maintenance of the National Library of Medicine (NLM)’s Medical Subject Headings (MeSH) thesaurus (49)
- Enhancing search within NLM's MEDLINE system (50)

- Enhancing search of NIH awards in RePORTER (3) and Grants.gov (51)
- Enhancing non-medical search on the NIH general website (52) and job search function (53)

A Criteria Definitions

For the sake of brevity, only criteria with a weight of 5 or higher are described here. Many other features were also considered (§7.1). Evaluations were based on industry reviews, product documentation, and sales presentations (§4).

A.1 Business Group

This group represents business factors, as opposed to product features.

A.1.1 Vendor Subgroup

This subgroup represents features of the vendor as an enterprise.

A.1.1.1 Market Presence

This feature indicates the vendor's current market standing, derived from relevant industry reports.

A.1.1.2 Corporate Strategy

This feature addresses the vendor's strategy for future growth and adaptation to changing markets and technologies, reflecting its potential ability to maintain or grow market presence and provide a competitive product. This assessment was derived from industry reports and the company's documentation.

A.1.1.3 Customer Support

This feature reflects the vendor's reputation for customer support.

A.1.2 Control Subgroup

This subgroup includes features related to NIH control of the final product, indicating NIH's ability to modify, disseminate, or otherwise make use of the product in manners of its own choosing.

A.1.2.1 Internal Maintenance

This feature indicates the degree to which NIH personnel control the deployment and maintenance of the final product. The highest value indicates that NIH has entire control, with no ongoing reliance upon the vendor.

A.1.3 Initial Cost Subgroup

This subgroup represents the initial cost of acquiring and developing the system.

A.1.3.1 Initial License

This feature represents the licensing cost for the software. The incumbent RCDC system is assigned an initial license cost of zero, since that cost has already been paid.

A.1.3.2 Development

This feature represents the cost in money and manpower for developing the system's technical framework, not including initial knowledge acquisition. This is largely a subjective assessment of information derived from product documentation and/or interviews with company representatives. The incumbent RCDC system is assigned an initial development cost of zero, since that task is already done.

A.1.3.3 Status Quo

This feature reflects general institutional cost of change. As the incumbent, Collexis receives the best value, while all other contenders receive the worst value.

A.1.4 Maintenance Cost Subgroup

This subgroup represents the ongoing cost for maintaining the system.

A.1.4.1 License Maintenance

This feature represents the (presumably annual) fee for continuing to license the software.

A.1.4.2 Domain Expert Time

This feature represents the cost in time of *domain experts* for configuring and maintaining the system.

A *domain expert* is an expert in a professional discipline, as opposed to a technical or administrative role. In the case of RCDC, the term refers to medical experts from the ICs (§10). The time of domain experts is generally considered to be more valuable than that of technical or administrative personnel.

A.1.4.3 Consulting

This feature represents the cost of securing additional consulting from the vendor not provided under the license agreement.

A.2 Technical Group

This group includes technical considerations of the product.

A.2.1 Interoperability Subgroup

This subgroup includes features related to interoperability with other tools.

A.2.1.1 Oracle

The tool must be able to store and access information in an Oracle Relational Database.

A.2.1.2 Java

It must be possible to call the tool from a Java program.

A.2.2 Capacity Subgroup

This subgroup refers to the physical capacity of the system to handle many simultaneous users and large amounts of information.

A.2.2.1 Scalability

Scalability refers to the ability of a system to run efficiently with many concurrent users, large amounts of data, and/or distributed over several computers running in parallel.

The RCDC system must scale adequately to hundreds of thousands of documents and dozens of simultaneous users. Information about scalability is derived from industry reviews, product documentation, and interviews with company representatives. Independent scalability tests are beyond the scope of this study.

A.3 Functionality Group

A.3.1 Correctness Subgroup

This group reflects the quality of the system's results.

A.3.1.1 Correctness

The system's categorization results should correspond to expert opinions, to the extent possible (§10).

A formal comparison of the incumbent Collexis system and its competitors is beyond the scope of this study. Indeed, it would be difficult to construct a fair experiment, given that RCDC has already been trained using countless hours of expert time. Rather, the assessment of the incumbent system is derived from the 2009 study, anecdotal reports, and validity data; while assessments of its competitors are derived from industry reports and customer testimonies.

A.3.1.2 Evaluation Mechanism

The system should provide a mechanism to measure correctness overall and by category. This reveals areas where the system needs improvement, and verifies whether intended improvements had the desired effect.

A.3.2 Concept Extraction Subgroup

This subgroup refers to the process of extracting concepts from mostly unstructured text, such as a grant application.

A.3.2.1 Thesaurus-Based

The system should perform categorization based on concepts in a thesaurus, as opposed to terms (i.e., text strings).

The primary purpose of this feature is instilling faith in IC experts that the system is categorizing documents based on scientifically defensible grounds. Note, however, that the need to manually assign concept weights (§5.2.5) undermines scientific defensibility. Contrast a statistical system, which assigns weights mathematically based on empirical evidence (§6.3.5.2).

A.3.2.2 Noun Phrases

The system should consider noun phrases as conceptual terms, e.g., `brain cancer`.

A.3.2.3 Synonyms

The system should identify synonymous terms as representing the same concept, e.g., `brain cancer` and `malignant brain neoplasms`.

A.3.2.4 Disambiguation

The system should distinguish homonym senses, e.g., **`organ`** `transplant` vs. **`organ`** `music`.

A.3.3 Categorization Subgroup

This subgroup refers to the specific process of assigning categories to documents.

A.3.3.1 Custom Categories

The system must assign categories from a customized canonical list.

A.3.3.2 Category Hierarchy

The system should support a hierarchical taxonomy of categories, as opposed to a flat list.

A.3.3.3 Distinguish Text Fields

The system must assign extra weight to concepts extracted from "the Title, Abstracts and Specific Aims of NIH grants and comparable descriptive text from other scientific documents (contracts, interagency agreements and intramural research projects)" (1), versus secondary text sections.

A.4 Usability Group

This group and its sole subgroup represent considerations related to ease of use of the final product, particularly by medical domain experts.

A.4.1 Usability Subgroup

A.4.1.1 Comprehensibility

The system must be transparent and comprehensible to stakeholders. Specifically, medical experts and other stakeholders must be able to understand the categorization process used by the system. Otherwise, they are unlikely to trust its results. This assessment is largely subjective.

A.4.1.2 Transparency

The system should provide a trace of the process used to generate a particular classification. Ideally, this should be understandable to a non-technical user.

B Rejected Candidates

This is a sampling of other candidates that were considered and rejected, with brief explanations.

B.1 Commercial Categorization Platforms

B.1.1 SAS Enterprise Content Categorization

Although SAS Enterprise Content Categorization (ECC) supports mainstream applications such as enterprise search and e-discovery, SAS does not emphasize these solutions, and has little market presence outside its analytic niche (54-55).

ECC is a rule-based system, like Collexis. Unlike Collexis, it has only one set of rules, generating concepts from terms. Each document is then categorized with the list of all concepts it includes. For instance, if there are 10,000 concepts in your ontology, then your set of documents will be categorized into 10,000 categories. If a given document contains 30 concepts, then it will be categorized into 30 categories.

B.1.2 Smartlogic Semaphore

Smartlogic's Semaphore product integrates with popular search solutions to introduce conceptual information. In essence, searchable terms are converted into concepts in an ontology.

Semaphore Content Classification Services is marketed as a document categorization add-on. However, each document is simply categorized with the list of all concepts it includes.

B.1.3 Oracle Text

Oracle is identified as a Challenger in Gartner's *Magic Quadrant for Information Access Technology* 2009 (22). Its Oracle Text product is already included under NIH's existing license agreement. However, Oracle Text addresses document classification only as an afterthought.

Classification rules are implemented as pattern-matching indexes into text documents stored in an Oracle database (56). Such rules would be untenable to maintain.

B.1.4 Microsoft

Microsoft is identified as a Leader in Gartner's *Magic Quadrant for Information Access Technology* 2009, in part because of its acquisition of Fast Search & Transfer in 2008 (22). However, their efforts are effectively limited to their SharePoint platform, and they do not offer document categorization except in partnership with Smartlogic (§B.1.1). Microsoft Research has produced numerous papers related to document categorization, but they have apparently never commercialized their results.

B.1.5 Rapid-I RapidDoc

Rapid-I RapidDoc is built on top of the open-source RapidMiner data mining platform. RapidDoc, however, is a proprietary solution, marketed as a web service, and maintained by Rapid-I staff; and does not offer conceptual analysis.

B.2 Open-Source Tools

A handful of widely used data mining and/or text analysis packages could provide a framework for a custom solution (§8.3).

B.2.1 Apache Lucene/Solr

Apache Lucene is the industry leader in open-source search engines, the lone subject of Gartner's *Open Source in Information Access, 2008* (57). Its adherents include LinkedIn, CNET, and the Smithsonian (58), and it was recently adopted by IBM as a foundation for their information access product line (22). However, Lucene does not currently support concept analysis or document categorization.

C What-If Analysis

The platform comparison in §8 (illustrated in Table 10 and Table 11) shows that three competitors score higher than the Collexis incumbent system, given the weights assigned to the various scoring criteria; and that Recommind scores higher than the other competitors both overall and on each individual criterion. This section uses a comparison between Collexis and Recommind (as the leading competitor) to explore the extent to which the weights on the categories have the potential to change the relative rankings of the two platforms.

Table 12 compares Collexis and Recommind at the level of the criteria groups shown in Table 10.

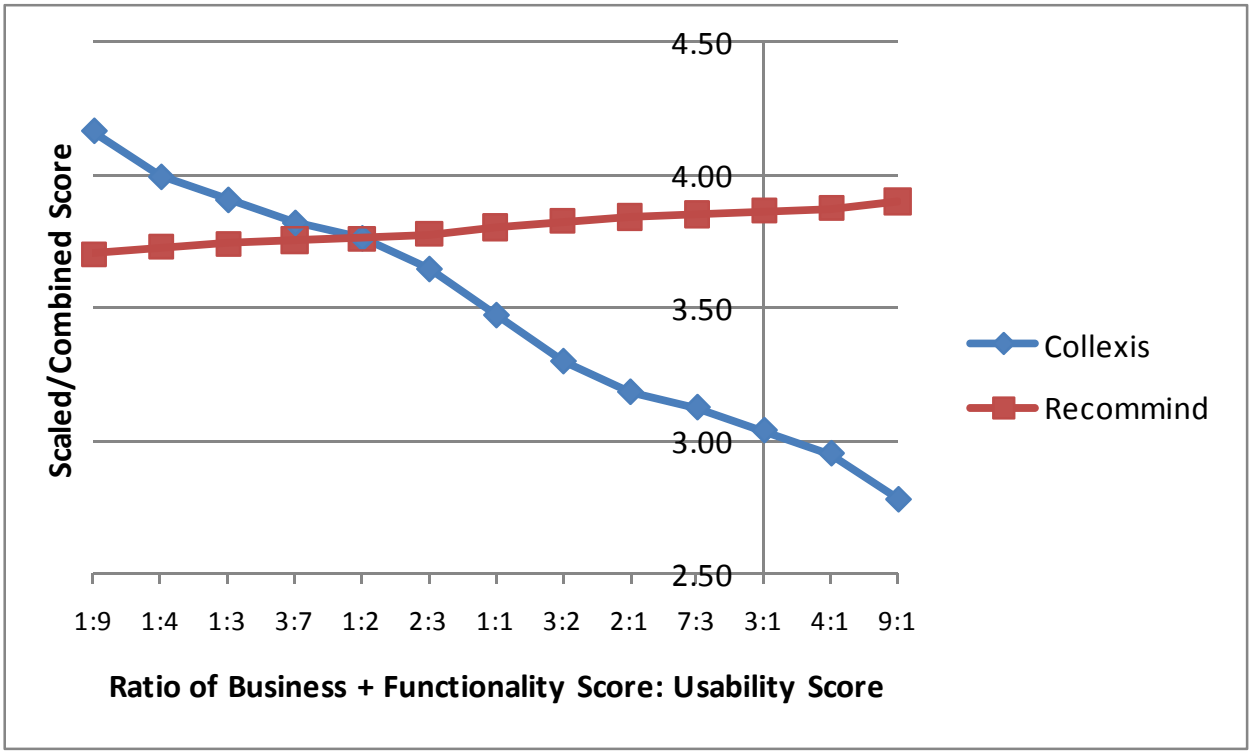
Table 12: Comparison of Collexis and Recommind Based on Table 10

Criteria Group	Collexis (8.1.1)	Recommind Decisiv (§8.1.3)	Difference
Business	1.79	**3.67**	*-1.88*
Technical	**4.69**	**4.69**	0
Functionality	3.42	**4.18**	*-0.76*
Usability	**4.34**	3.68	**0.66**

Recommind scores substantially higher than Collexis on the Business criteria, and higher on the Functionality criteria, while Collexis scores higher on the Usability criteria.

Figure 2 below explores the effect of different relative weightings of these three criteria groups, excluding the Technical criteria because the two systems score evenly. Figure 2 graphs the relative combined total score across these three criteria groups, using a variety of relative weightings of the Usability criteria (on which Collexis scores higher) and the Business and Functionality criteria (on which Recommind scores higher). The Y-axis crosses the X-axis at the 3:1 weighting of criteria groups used in the study.

Figure 2: What-If Analysis Based on Table 10



Note: Because Technical score was excluded, summed scores do not match the Total score reported in Table 10.

As shown in Figure 2, Recommind scores higher than Collexis at the 3:1 weighting used in the study, and continues to score higher unless the Usability score is weighted at more than twice the sum of the Business and Functionality scores. It is also worth noting in this exploratory analysis that the Recommind score is relatively invariant – because of the narrow variation between the Usability score (3.68) and the Business and Functionality scores (3.67 and 4.18, respectively) – while the 1.79 Business score Collexis received makes its score more sensitive to the relative weightings of the categories.

Table 13 compares Collexis and Recommind on the nine key criteria from Table 11. Of those nine criteria, Collexis scored higher on three (Status Quo, Thesaurus-Based, Comprehensibility), Recommind scored higher on four (License Maintenance, Domain Expert Time, Correctness, and

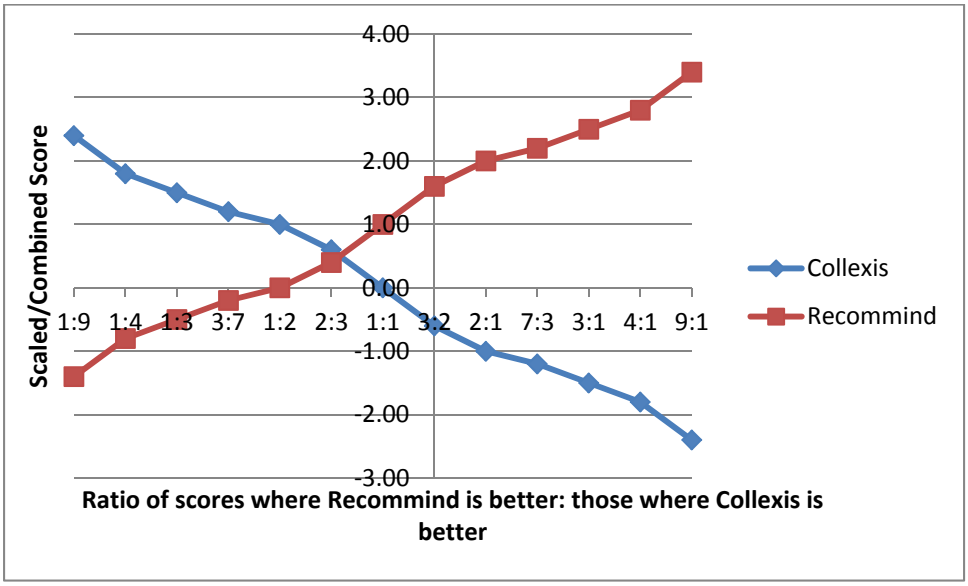
Evaluation Mechanism), and the platforms scored equally on two (Oracle, Distinguish Text Fields).

Table 13: Comparison of Collexis and Recommind Based on Table 11

Criterion	Collexis	Recommind	Difference
Status Quo	**1**	*-1*	Collexis higher
License Maintenance	*-1*	**1**	Recommind higher
Domain Expert Time	*-1*	**1**	Recommind higher
Oracle	**1**	**1**	None
Correctness	0	**1**	Recommind higher
Evaluation Mechanism	*-1*	**1**	Recommind higher
Thesaurus-Based	**1**	*-1*	Collexis higher
Distinguish Text Fields	**1**	**1**	None
Comprehensibility	**1**	0	Collexis higher

The exploratory analysis in Figure 3 below compares relative weightings of the three criteria on which Collexis scores higher versus those on which Recommind scores higher (rather than comparing each weighting individually). The Y-axis crosses the X-axis at the 3:2 weighting of criteria groups used in the study.

Figure 3: What-If Analysis Based on Table 11



The what-if analysis shows that a comparison of the two platforms based solely on these seven key criteria is sensitive to their relative weighting. The break-even point is where the ratio of the weighting of the Collexis-favoring criteria to the weighting of the Recommind-favoring criteria is 3:2, the opposite of the weighting used in the study. With any higher weighting, Collexis would score higher than Recommind.

D Index of Technical Terms

E References

1. Work Plan for Tasks 1 and 3 of RCDC Evaluation. National Institutes of Health Office of Extramural Research; 2010.
2. Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC). 2010 [updated 2010 Feb 2; cited 2010 Apr 16]; Available from: http://report.nih.gov/rcdc/categories/.
3. RePORT EXPENDITURES & RESULTS (RePORTER). Bethesda, MD: National Institutes of Health; 2010 [updated 2010 September 07; cited 2010 September 8]; Available from: http://projectreporter.nih.gov/reporter.cfm.
4. Committee on the Organizational Structure of the National Institutes of Health. Enhancing the Vitality of the National Institutes of Health: Organizational Change to Meet New Challenges. Washington, D.C.: National Academy of Sciences; 2003 [cited 2010 July 23]. Available from: http://www.nap.edu/catalog/10779.html.
5. National Institutes of Health Reform Act of 2006, P.L. 109–482. Sect. 402(b) (2007).
6. Summary of Meeting: September 14–15, 2005. Department of Health and Human Services-National Institutes of Health-National Cancer Institute-37th NCI Director's Consumer Liaison Group. Bethesda, MD: National Institutes of Health (NIH); 2005 [cited 2010 September 3]. Available from: http://deainfo.nci.nih.gov/advisory/dclg/14sep05mins.pdf.
7. eRA Working Group Explores Technology-Assisted Disease Coding. Electronic Research Administration (eRA) National Institutes of Health; 2004 [updated 2004 May; cited 2010 September 3]; Available from: http://era.nih.gov/eranews/news_article1.cfm?lobjectid=F12ABEF5-101C-4874-B9A87DE7882300B0.
8. NIH Reauthorization. 2009 [updated 2009 May 6; cited 2010 April 16]; Available from: http://www.nih.gov/about/reauthorization/.

9. Heagy J, Holzer J, Chappell I. Analysis of the NIH RCDC Grant Tracking and Categorization System. D-3755: Institute for Defense Analyses Science & Technology Policy Institute; 2009.

10. Research, Condition and Disease Categorization (RCDC) Fingerprint Process: IC Expert Training [PowerPoint briefing]: National Institutes of Health (NIH); March 17, 2009.

11. Recall-Precision Graph. National Institute of Standards and Technology (NIST); 2008 [updated 2008; cited 2010 August 24]; Available from: http://trec.nist.gov/presentations/TREC8/intro/sld033.htm.

12. Our Customers. Collexis; 2010 [updated 2010; cited 2010 July 14]; Available from: http://www.collexis.com/customers/.

13. Bell T, Gilbert MR, Knox RE. Cool Vendors in Content Management, 2008. G00155552. Gartner; April 8, 2008.

14. Elsevier acquires Collexis, a leading developer of semantic technology and knowledge discovery software for research and development institutions. Collexis; 2010 [updated 2010 June 10; cited 2010 July 7]; Available from: http://www.collexis.com/news/press100610.htm.

15. Collexis Grant Management Solution. Collexis; 2010 [updated 2010; cited 2010 September 11]; Available from: http://www.collexis.com/products/GrantMgtSolution.htm.

16. FW: Price list [Email]: Collexis; July 13, 2010.

17. Heuer RJ, Jr. Psychology of Intelligence Analysis: CIA; 1999.

18. Owens L. The Forrester Wave: Enterprise Search, Q2 2008. Cambridge, MA: Forrester Research, Inc.; May 28, 2008.

19. Bellopede D, Kaul S, Etheridge N. [On-site demo]: Recommind; July 15, 2010.

20. Content Categorization. Recommind; 2010 [updated 2010; cited 2010 September 10]; Available from: http://www.recommind.com/solutions/enterprises/content_categorization.

21. Solutions Overview. Recommind; 2010 [updated 2010; cited 2010 September 10]; Available from: http://www.recommind.com/solutions/overview.

22. Andrews W. Magic Quadrant for Information Access Technology. G00169927. Stamford, CT: Gartner; September 2, 2009.

23. EERE Division of the U.S. Department of Energy (DOE) Selects MindServer Categorization and MindServer Search From Recommind. Recommind; 2010 [updated 2010 July 20; cited 2010 August 4]; Available from: http://www.recommind.com/releases/20100720/DOE_selects_mindserver_search_and_mindserver_categorization.

24. Xythos Enterprise Document Management Suite 7.2. Blackboard Inc.; 2010 [updated 2010; cited 2010 July 22]; Available from: http://www.xythos.com/products/EDMS_72NEWFEATURES.html.

25. Palmer T. Telephone conversation. 2010, July 22.

26. Customer Success Stories. Recommind; 2010 [updated 2010; cited 2010 July 14]; Available from: http://www.recommind.com/customer_success.

27. Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning. 2001;42(1/2):177-96.

28. Puzicha J. Finding Information: Intelligent Retrieval & Categorization. White Paper. Recommind; n.d.

29. Decisiv Search. Recommind; 2010 [updated 2010; cited 2010 September 10]; Available from: http://www.recommind.com/products/decisiv_search.

30. MindServer Categorization. San Francisco, CA: Recommind; 2009.

31. IBM InfoSphere Classification Module V8.7: Reducing compliant information management costs through advanced content classification. 2009 [updated 2009 August 18; cited 2010 July 2]; Available from: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=ca&infotype=an&appname=iSource&supplier=897&letternum=ENUS209-234#h2-descx.

32. View Pricing and Buy. 2010 [updated 2010; cited 2010 July 2]; Available from: https://www-112.ibm.com/software/howtobuy/buyingtools/paexpress/Express?P0=E1&part_number=D61U6LL,D090ILL,D090KLL,D090MLL&catalogLocale=en_US&Locale=en_US&country=USA.

33. IBM InfoSphere Classification Module information. IBM; 2010 [updated 2010; cited 2010 September 11]; Available from: http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp.

34. User Groups: Belong, Share, Benefit. IBM; 2010 [updated 2010; cited 2010 September 11]; Available from: http://www-01.ibm.com/software/data/usergroup/.

35. Likelihood function. 2010 [updated 2010 June 3; cited 2010 July 7]; Available from: https://secure.wikimedia.org/wikipedia/en/wiki/Likelihood_function.

36. DITA and XML Authoring for Business Users. IBM; 2010 [updated 2010; cited 2010 September 13]; Available from: http://www-01.ibm.com/software/data/content-management/filenet-content-manager/dita.html.

37. Magdalen J. On-site demo. 2010, July 21.

38. Classification Workbench User's Guide. IBM InfoSphere Classification Module: Version 87: IBM; 2009.

39. Rueckert M. IBM Classification Module: Technical Overview [Powerpoint Presentation]: IBM Corporation; 2008.

40. Developer's Guide. IBM InfoSphere Classification Module: Version 87: IBM; 2009.

41. Logan D, Andrews W, Bace J. MarketScope for E-Discovery Software Product Vendors. G00171281. Stamford, CT: Gartner; December 21, 2009.

42. Andrews W. Autonomy IDOL for Information Access: Effective for Strategic Use; Difficult for Smaller Implementations. G00168556. Stamford, CT: Gartner; July 13, 2009.

43. Walton C, Smith C. [On-site demo]: Autonomy; July 9, 2010.

44. A Different Approach. Autonomy; 2010 [updated 2010; cited 2010 September 13]; Available from: http://www.autonomy.com/content/Technology/autonomys-technology-a-different-approach/index.en.html.

45. Joachims T. Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning. 1998:137-42.

46. Meaning Based Computing (MBC). Autonomy Power Solutions Overview 2010. San Francisco: Autonomy; 2010. p. 4.

47. Caldwell F, Bace J, Logan D, Andrews W, Chin K, MacComascaigh M, et al. Hype Cycle for Legal and Regulatory Information Governance, 2009. G00173013. Stamford, CT: Gartner; December 22, 2009.

48. MacComascaigh M, Knox RE, Gilbert MR, Shegda KM, Bell T, Andrews W, et al. Hype Cycle for Content Management, 2009. G00169486. Stamford, CT: Gartner; July 24, 2009.

49. 2010. Medical Subject Headings (MeSH). U.S. National Library of Medicine; [April 01; cited 2010 September 11]; Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

50. MedlinePlus. Bethesda, MD: U.S. National Library of Medicine; 2010 [updated 2010 September 3; cited 2010 September 3]; Available from: http://medlineplus.gov/.

51. Grants.gov. U.S. Department of Health and Human Services; 2010 [updated 2010; cited 2010 September 13]; Available from: http://www.grants.gov/.

52. Search the NIH Web Site. National Institutes of Health (NIH); 2010 [updated 2010; cited 2010 September 13]; Available from: http://www.nih.gov/google.search.nih.html.

53. Jobs @ NIH. National Institutes of Health; 2010 [updated 2010 August 3; cited 2010 September 13]; Available from: http://www.jobs.nih.gov/.

54. Sallam RL, Hostmann B, Richardson J, Bitterer A. Magic Quadrant for Business Intelligence Platforms. G00173700. Stamford, CT: Gartner; January 29, 2010.

55. Kobielus J. The Forrester Wave: Predictive Analytics And Data Mining Solutions, Q1 2010. 56077. Cambridge, MA: Forrester Research, Inc.; 2010.

56. Oracle Text: An Oracle Technical White Paper. Oracle; 2007 [updated 2007 June; cited 2010 June 22]; Available from: http://www.oracle.com/technology/products/text/pdf/11goracletexttwp.pdf.

57. Andrews W. Open Source in Information Access, 2008. G00155975. Stamford, CT: Gartner; April 3, 2008.

58. Lucene/Solr Application Showcase Wiki. 2010 [updated 2010; cited 2010 June 25]; Available from: http://www.lucidimagination.com/developer/Community/Application-Showcase-Wiki.