



# Unique Features of Content Analyst™

(and how those Translate into Increased Value for our Customers)

Whether they are reading electronic communications, monitoring publications for relevant information, digging into reference volumes for the right piece of relevant research, or even sorting through emails and correspondence, knowledge workers must spend a significant portion of the daily effort sifting and categorizing information. How can companies and organizations utilize the massive amounts of unstructured data they see, across multiple languages and platforms, while dramatically improving their ability to ensure that no piece of relevant, mission-critical information slips by?

Content Analyst provides the answer, giving users the technology they need to turn massive amounts of text and data into meaningful and relevant information. There are no languages, rules, or ontologies to learn. Content Analyst utilizes a major extension and refinement of technology that has evolved from the original concept of Latent Semantic Indexing. This technology advancement enables Content Analyst to do the time-consuming work that turns raw unstructured data into actionable information.

## Software Learning

A large part of our unique technology is the combination of our LSI core and our extended routines, resulting in a very powerful *Software Learning* solution. We found early on that LSI works best when you let the software *train itself* and *learn from what it finds*. Therefore, we developed the routines that allow Content Analyst to both self-learn and to learn-by-example. To the casual user, the software appears to learn all by itself; to the software architect, this is really about the power of LSI to find and retain relationships that are hidden to other software techniques.

Our Categorization feature uses another aspect of software learning called *Exemplars*. Exemplars are “example documents.” When a user is setting up Categorization, they simply define the categories they want Content Analyst to recognize and provide a representative set of documents for each category.

When a new collection of documents is passed through the Categorization engine, Content Analyst reads each new document and compares it against the categories as previously defined by the exemplars - and properly categorizes or associates each new document with the category or categories it has learned.



## There are Ten Technologies or Benefits provided *only* by Content Analyst:



1. Contextual Search across Document Sets
2. Deep Conceptual Generalization
3. Cross-Lingual Operation without Translation
4. Instant Context for all Items
5. Novelty-Based Ranking
6. Automatic Alias Identification
7. Social Network Structure Generation
8. Comprehensive Information Overview from a Single Tool
9. No Auxiliary Structures Required for Accuracy
10. High-Tolerance for Data Errors and Inconsistencies

### Contextual Search across Document Sets

**What it is:** Content Analyst operates on the basis of concepts – therefore, it does not depend on the words that users choose when they formulate queries. As a result, Content Analyst can retrieve relevant documents even if they contain no words in common with the queries.

**Why it is important:** Other search technologies parse documents into pieces and then go back and try to “umbrella” context over those pieces to gain a complete picture since they don’t operate across all documents at once. With Content Analyst, the entire relationship is preserved – *and in many cases, the relationships that Content Analyst finds are invisible to other technologies.*

**Commercial Value:** One Content Analyst customer has built a subscription news service that retrieves articles based on context, which their market sees as superior to other platforms that use keyword and proximity search.

### Deep Conceptual Generalization

**What it is:** as part of Content Analyst’s *transform techniques* (our software is considered to use transform technologies, in that it “transforms” documents into powerful mathematic formulae that our software can analyze) our software creates multi-level associations in multiple dimensions (like many business intelligence products).

**Why it is important:** These associations can be turned into an automatic taxonomy generation (i.e., no human intervention required), or used to power dynamic categorization (where the software automatically ranks and sorts data according to these concepts).

**Commercial Value:** A number of Content Analyst OEMs use these features to power solutions as diverse as an incoming email routing system for Customer Service to an automatic document identification system for an online medical records insurance processor.

## Cross-Lingual Operation without Translation

**What it is:** think *Rosetta Stone*: as part of Content Analyst's powerful *Software Training* routines, we "trained" the base engine in 17 native languages (including Asian and Oriental languages). Because all language is essentially mathematical, Content Analyst can look for associations across different languages without translation; it "knows" what an English concept would look like in French, for example.

**Why it is important:** All other commercially-available cross-lingual search techniques require some level of machine translation in order to be able to search and analyze non-English documents – not true with Content Analyst.

**Commercial Value:** One example is a Content Analyst customer who reviews daily articles across Europe and the Middle East looking for particular subjects of interest; by using Content Analyst, they no longer need to translate those documents first. Now they only translate a small subset of *resultant* documents. This has saved money *and* allowed them to offer "same-day" services to their customers.

## Instant Context™ for all Items

**What it is:** Content Analyst can display any word, term or concept relative to the context of other similar terms or concepts across the set of relevant documents it has found.

**Why it is important:** Any material retrieved by Content Analyst may contain new or unfamiliar terms – sometimes they are foreign, sometimes simply new. Instant Context means the user can click on the word or item and a "pop-up" will show all the associated contextual terms for that item (so you immediately know what it means).

**Commercial Value:** The customer who presents a daily "update" of relevant news articles had the ongoing problem of explaining why a given term was important; with Instant Context, his support issues were dropped in half across his user base.

## Novelty-Based Ranking

**What it is:** Content Analyst also tracks what it has seen or indexed before. Therefore, it can return results to specific users based on how *novel* or unique new information is, i.e. has it already been seen in some form. This works cross-lingually as well.

**Why it is important:** Any kind of repetitive service or program is greatly crippled if it continues to find the same information previously collected. Users want to see "what's changed" from what they have already seen.

**Commercial Value:** One Content Analyst-powered information retrieval service added this feature at the request of their customer base - the ability to suppress previously-shown information. This helped double their re-subscription rates.

## Automatic Alias Identification

**What it is:** When Content Analyst indexes document and catalogs relationships along with contextual relationships, it also learns "alias" information. This broad capability includes synonyms, nicknames, aliases, transliteration differences, speech conversion errors, OCR errors, and even trade names.

**Why it is important:** Other search solutions cannot recognize different spellings of terms, misspellings, or abbreviations. They either must be "fed" that information beforehand, or use routines based on wild cards (which naturally cause errors). Those which do provide error correction can only do so within 5-10% of the original term.

**Commercial Value:** One customer uses Content Analyst to review and categorize daily correspondence about individuals that includes field work where aliases and even code names are common; in fact, they found they could actually use Content Analyst to discover code names out of routine correspondence.



## Social Network Structure Generation

**What it is:** another “element” in Content Analyst’s multidimensional indexing and analysis is an automated linking between disparate yet connected objects (think customers). Simply put, Content Analyst will provide all the data required by visualization software to identify *social networks*, i.e. who is socially related, known-to, etc., whom.

**Why it is important:** Without Social Network Structure analysis, it is impossible to identify trends that will affect related individuals. With all other commercially-available search techniques, they only augment in-place visualization tools: they cannot automatically feed the necessary information without a pre-determined structure.

**Commercial Value:** One Content Analyst customer is powering an innovative system that looks at potential customer abandons (i.e., who is most likely to leave their service for a competitor) and correlates these to that customer’s social network of that customer. The idea is to pre-empt a disgruntled customer’s service issue from boiling over to affecting his friends’ and neighbors’ choice of provider.

## Comprehensive Information Overview from a Single Tool

**What it is:** In as much as Content Analyst’s unique features have individual merit, the fact that they are all invoked *from a single tool, against a single data index*, is very significant. Many competing text analytics providers offer a suite of products that mimic some of Content Analyst’s offerings, but they do so by linking a series of routines and techniques, often from different vendors.

**Why it is important:** a Boolean search technique will not yield what a Proximity Word Search technique does. If a product provides instant context by first indexing all the keywords (Boolean), then running a similar technique to find related words (Proximity), and finally running a third routine to find synonyms (Natural Language Processing) it will miss 20% or more of the data that Content Analyst can identify with a unified process, because each “change of gears” has some inherent data loss.

**Commercial Value:** One application into which Content Analyst was integrated uses our software to first create an automatic taxonomy and then apply it “on-the-fly” as their users scan in new documents. By merging these two concepts, the Content Analyst customer created a completely new market for their 700+ corporate clients.

## No Auxiliary Structures Required for Accuracy

**What it is:** Content Analyst performs its deep analysis of unstructured text without any pre-determined structures like Ontologies, Taxonomies, Thesauri, Categorizations, etc. There are no rigid data structures, nor are there lengthy lists of exceptions that must be processed separately, to achieve accurate results.

**Why it is important:** In order to use other systems, there is a lengthy, arduous, and often expensive set-up required before any work can take place. There are several problems with these solutions. One is the fact that unstructured data is *fluid*: structures like taxonomies quickly become obsolete and new ideas or terms are constantly being introduced into business. Another is that in many cases there is no way to create these auxiliary structures without knowing the resultant data outcome.

**Commercial Value:** Certain Content Analyst-based Intelligence applications use Content Analyst to identify aliases and deliberately-used false names. There was no way to develop an up-front taxonomy for these names *since the relationships weren’t uncovered until after* Content Analyst performed its work.

## High Tolerance for Data Errors and Inconsistencies

**What it is:** Content Analyst is extremely resistant to data errors and inconsistencies. Because it reads and understands entire documents, recognizing a misspelled or inconsistent term is no different than correlating two different terms that happen to apply to related concepts.

**Why it is important:** Data errors and inconsistencies are responsible for up to 20% of unstructured text being eliminated from analysis by other systems. Often, this 20% represents tremendous value that is overlooked.

**Commercial Value:** Content Analyst's resistance to data errors has been exploited by the Intelligence community when they scan and interpret "raw" machine input of sensitive material – permitting real-time analysis in situations where the same effort formerly took days to produce results.



Our customers represent a wide range of industries and solutions. For some customers, just one or two of our unique features were enough to make Content Analyst their engine of choice. For others, the *inclusion* of Content Analyst alongside other solutions has solved nagging problems, giving their products an "edge" over their competition by combining the power of Content Analyst with their other tools and software techniques.

From highly-sensitive government installations to innovative new commercial products, the power of Content Analyst is at work "behind the scenes" and being regularly employed by tens of thousands of users. How can we help put the power of Content Analyst and LSI to work for you?

**Learn More:** Call 888.349.9442/703.391.8700 or Email [info@contentanalyst.com](mailto:info@contentanalyst.com)



Stop searching. Start doing.™