

New Software transforms research with visualization techniques

A new tool is giving researchers an instant, visual way to swiftly analyze and assess data from even the most complex scientific searches. STN® AnaVist™, developed by Chemical Abstracts Service and FIZ Karlsruhe, is powerful analysis and visualization software that makes research faster and more productive by presenting results in interactive, dynamic charts.

"The information challenge of the 21st century is not information access, but information utilization," said Robert Massie, president of CAS. "This new software is an important first step from the established STN partnership of CAS and FIZ Karlsruhe to bring new technologies to bear on this challenge."

The new tool provides important benefits that can be felt throughout an organization. With its clear visualizations and powerful analysis capabilities, the software helps information professionals and management make sound decisions more quickly.

"STN AnaVist is an easy-to-use tool providing concise, graphically well-presented patent analyses," said Dr. Gisela Darges, competitive intelligence, Degussa AG. "Of particular interest are the interactive linkages between the individual charts allowing for multidimensional viewing of selected criteria. The two- or three-dimensional Research Landscape is a promising new way to analyze and visualize document content. Such tools support innovation and trend research and can possibly shorten the time expended in R&D."

STN AnaVist analyzes search results from several databases of the STN network that encompass an extensive range of scientific literature and patents. Results that can be analyzed with the software can be imported from STN Express® with Discover!™, the search tool most commonly used by information professionals to search STN. In addition, users may perform a concept search within STN AnaVist itself. Analysts can use STN AnaVist to find out what their competitors are doing; analyze which companies are obtaining patents in what areas; explore new applications for existing technology; learn whether a particular area of

research is on the rise, holding steady, or declining; and support strategic business planning.

"You can easily see who is doing research, where they work, what they have published," said Michael Dennis, vice president of planning and development for CAS, a division of the American Chemical Society. "You can track your competitors and find new companies to work with."

CAS managers are confident their new product is reaching an eager market because STN AnaVist was developed in conjunction with customers based on their needs and requests. Before introduction this summer, it was heavily beta tested by customers in Europe and the United States. Customers have already commented positively on the product's ease-of-use and the powerful, interactively-linked graphics.

NEW APPROACH

STN AnaVist is a new approach to analyzing data from STN, a collection of scientific, technical, patent and business databases from around the world.

STN AnaVist quickly assesses millions of pieces of information in key STN chemical and patent databases. It automatically creates a group of interactive charts that show the research performed on a topic, who has done the research, how various applications being researched relate to each other and where patents are held.

"Anyone who has ever performed a search with typical Web search engines and received a bewildering list of thousands of

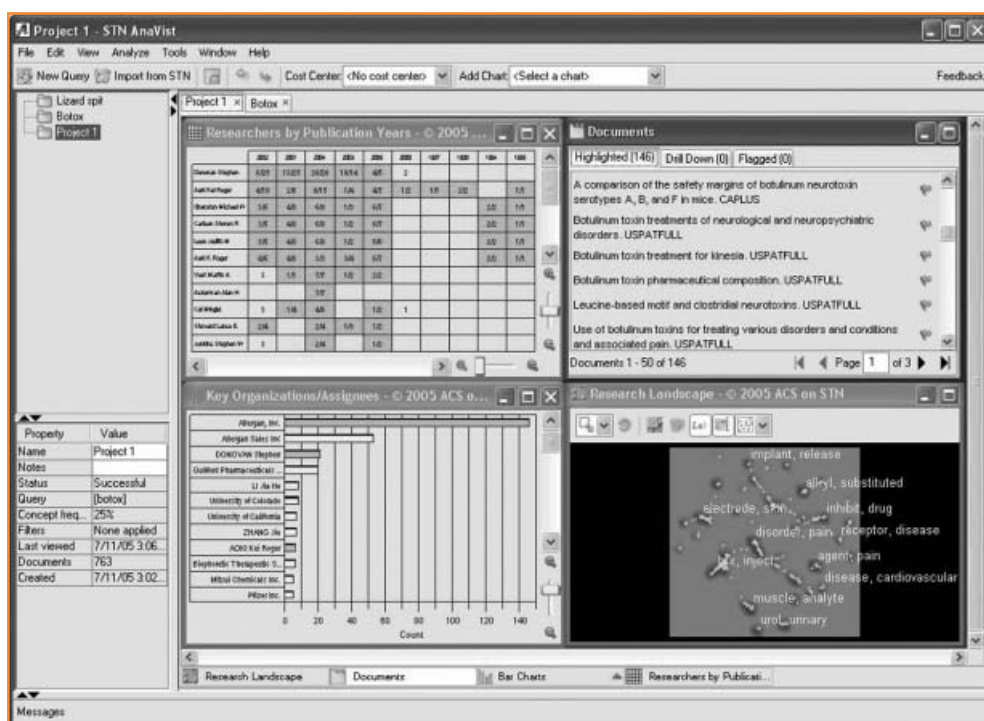


Figure 1. The software generates insights, not just answers to searches by creating interactive visualizations, including a matrix chart, bar chart, linked documents window and the Research Landscape overview.

documents in response can appreciate the time- and labor-saving value of STN AnaVist's ability to navigate through a multitude of literature," said Dennis.

Because STN AnaVist can organize large numbers of documents into charts and graphs so swiftly, searches and reports that used to take days, weeks, or even months have been reduced to hours or minutes.

While STN AnaVist's initial release focuses on chemistry-related searches, future releases will incorporate the numerous other fields of science covered by STN.

Currently, STN AnaVist analyzes data from three major STN databases: CAplusSM (CAS' most current and comprehensive chemistry bibliographic database), PCTFULL (a full text database containing Patent Cooperation Treaty applications) and USPATFULL (the complete U.S. patent database). STN AnaVist integrates the content from these sources, giving analysts a broad range of chemical and patent information from which to draw.

GENERATING A LANDSCAPE VIEW

Although many visualization tools provide only charts or landscapes, STN AnaVist provides an integrated view. Previously, to benefit from both data mining and text mining, analysts have needed to prepare different sets of data and work with multiple tools. STN AnaVist was designed to put powerful text and data mining tools together, so analysts can easily use both techniques and gain insights when they are interactively combined.

"The interactivity of the visualized data and the maps based on high-value databases make STN AnaVist a pioneering analysis and visualization tool," said Nicolas Lalyre, patent information specialist for Syngenta, "For example, you can select an area of interest on the map and simultaneously receive the key competitors, key inventors and database records related to this area."

As Figure 1 shows, the interactive visualizations automatically created by STN AnaVist include a matrix chart, bar chart, dynamically-linked documents window and Research Landscape. These graphics offer a new view of searching; a question doesn't just generate an answer, it generates insights.

One of STN AnaVist's most useful features, the Research Landscape, presents an overview of the research activity relating to a specified concept, represented in a visualization that resembles a topographic map. Analysts can easily see a complete picture of, for example, a specific company's research. To make it even simpler to interpret, the more similar the research areas, the closer they are grouped together. The density of documents is depicted on the Research Landscape as peaks. The more research that has been done in one specific area, the higher the peak on the Research Landscape. The Research Landscape can be focused or tilted to view peaks more easily. Analysts can easily obtain a deeper level of detail by clicking on an area of interest.

Along with the Research Landscape, searches yield bar charts on subjects like organizations conducting research, publication year trends and matrixes that list, for example, researchers at a particular company sorted by publication year. These can be viewed to immediately see overall trends or to find specific answers.

Once they have their initial results, analysts can click on individual bars or peaks in many of the charts for more detail. Doing so not only shows relationships among the data, but also presents the analysts with a list of related documents that can be viewed or flagged for future use.

In a patent classification chart sorted by dates, an analyst can learn more about new research directions by clicking on the patent classification code to access the definition. Cluster concepts created for the Research Landscapes can be edited to obtain a greater level of precision.

Analysts can further refine their answers by selecting concepts on the Research Landscape that they want to explore. Those selected areas will be highlighted in all other charts. Highlighted documents can be viewed in greater detail, adding a deeper dimension of analysis.

In many charts, analysts can customize the way they view information. For example, a bar chart of publication year trends can be sorted by document count or label. In a matrix listing researchers by publication years, specific documents can be highlighted and viewed against the total documents. To explore patent filing trends, the listed patents can be sorted alphabetically.

HARNESSING CAS' POWER

One of STN AnaVist's most powerful attributes is its ability to take advantage of CAS' CAplus database. CAplus draws from the indexing and vocabulary that CAS has applied to the literature for more than a century. Because CAS document analysts have tracked evolving scientific terminology over time and understand how terms relate, the database they maintain can be searched with a high degree of accuracy and relevancy. Analysts can be confident that the terms they search will draw the most comprehensive and reliable results. Key records that might be missed in other database searches can be found in CAplus. These strengths are evident to analysts as they search for relationships in the data from nine different charts available in STN AnaVist: clustering concepts, technology indicators, patent classifications, key researchers, key organizations/assignees, document distribution, patent-issuing organizations, publication years and priority applications dates. Thesauri that are intellectually created by CAS document

```

=> E E1+ALL
E1      0      CNUM CAS1003414/CO
E2     1376   --> AMGEN INC/CO
          NOTES 1980: AMGen established
          1981: AMGen changed company name to Applied
          Molecular Genetics, Inc.
          1983: Applied Molecular Genetics, Inc. changed name
          to Amgen
          1994: Amgen acquired Synergen Inc.
          2000: Amgen acquired Kinetix Pharmaceuticals
          2002: Amgen acquired Immunex Corp.
E3      215   RT1 AMGEN/CO
E4      20   RT1 AMGEN BOULDER INC/CO
E5      15   RT1 AMGEN CANADA INC/CO
E6       4   RT1 APPLIED MOLECULAR GENETICS INC/CO
E7     422   RT1 IMMUNEX CORP/CO
E8     337   RT2 IMMUNEX CORPORATION/CO
E9       7   RT1 KINETIX PHARMACEUTICALS INC/CO
E10     1    RT1 SYNERGEN A G/CO
E11     1    RT1 SYNERGEN AG/CO
E12     2    RT1 SYNERGEN ASSOCIATES INC/CO
E13     4    RT1 SYNERGEN BIOLOGICALS INC/CO
E14    104   RT1 SYNERGEN INC/CO
***** END *****

```

Figure 2. The company name thesaurus creates an answer set of all of Amgen's name variations and other companies acquired by AMGEN, to assure it is building a complete reference of research by the company.

analysts are applied in STN AnaVist to make clusters more relevant. Terms are extracted from a document's title and abstract, then standardized and consolidated using CAS vocabulary. This reduces data scatter, which saves time and produces more meaningful results.

STN AnaVist helps analysts retrieve all of the documents that relate to given organizations, which is important for competitive intelligence searches. CAS document analysts have developed an especially dependable and thorough thesaurus of company names. Over the years, they have recorded changes in company names and compiled the results in a company name thesaurus. This ensures that company name searches retrieve all relevant results.

Another of STN AnaVist's major strengths is its ability to help analysts find new relationships between structured and unstructured data. STN AnaVist is the only tool currently available that allows analysts to combine data mining of controlled terminology with text mining of unstructured data and offer seamless, dynamic interaction. This increases analysts' options and helps them take their searches in new directions.

COMPETITIVE INTELLIGENCE TOOL

With its ability to track research activity, STN AnaVist is an important competitive intelligence tool. STN AnaVist provides the data researchers need to do in-depth competitive analysis. Good competitive analysis requires reliable tools combined with authoritative content. STN AnaVist brings these together. As a result, analysts can spend their time analyzing, not processing data.

According to Dennis STN, AnaVist's value extends beyond competitive intelligence.

Organizations can use the search engine to gain information to help them in mergers and acquisitions, and employee recruitment and hiring.

An analyst can investigate not only what areas a company is doing research in, but also which employees are involved in that research, where the research has been conducted, and what patents a company holds. Other valuable uses for organizations are investigating trends in competitors' research, learning who has technology they would like to license and finding potential buyers for their own products.

For example, to learn about the recent research of the Amgen Company, an analyst first creates a search using the company name thesaurus in CAplus. After obtaining an answer set of all of Amgen's name variations and other companies acquired by Amgen, as shown in Figure 2, the analyst saves the answer set using the *Discover!* Wizard in STN Express with *Discover!*, Version 8.0. Next, the analyst imports the saved

results into STN AnaVist. The import summary shows document titles and a summary by publication year, document type and priority application year. By clicking "Start Visualization", the answers appear in the charts and other graphics that emerge on the screen. Analysts can then scan the results, click on individual areas for more information, further sort results, save, or print.

For competitive intelligence searches, the bar chart that identifies key organizations and their frequency publishing in a given research area is particularly useful. Key organization terms also can be edited for more detailed analysis. Analysts can group or ungroup names, groupings can be saved for future searches and names can be excluded. The power of searching technology by clustering concepts and indicators also comes into play in competitive intelligence searches. Analysts can easily compare controlled versus free-text documents in a matrix chart. Technology indicators are consistent because they are derived from CAplus controlled vocabulary. Analysts can see all of the documents in which they are most interested by clicking on a cell to highlight all the other cells in the chart that contain documents with the terms being searched, as shown in Figure 3. Analysts can then click on the highlighted answers to obtain details such as a document's name and authors, patent information and an abstract.

	rhl	prevent	grow	scdhd	nerve	glucose	pathway	polyol	thepobain	neopathy	accumable	human
Diabetes mellitus	10499	6565	5652	4295	6257	3255	3246	2234	10234	165	3200	195
Nerve	11223	7184	5150	8148	11253	171	7106	7111	12135	6231	5110	162
Eye	5163	4134	2108	2401	357	245	373	273	559	161	343	178
Cataract	2106	242	41	149	123	158	121	232	291	141	150	33
Kidney	260	178	297	46	126	64	65	162	240	38	145	44
Alcohols	164	154	54	128	104	33	70	194	100	123	147	18
Polyol pathway	351	341	268	105	123	44	291	391	325	122	124	33
Antidiabetic agents	2686	1583	1047	1741	2044	619	824	1022	2928	647	1329	158
Hypoglycemia	38	40	61	31	7	67	46	38	9	7	22	27
Blood vessel	246	251	154	32	215	42	37	28	219	17	9	23
Erythrocyte	127	121	26	147	122	28	118	119	117	9	100	36
Structure-activity relat.	225	117	10	216	214	7	5	7	24	2	210	15
Proteins	30	26	27	11	8	19	12	13	14	5	12	13
Antioesity agents	2	16	10	1	1	19				2		6
Lipids	14	24	26	7	4	22	6	3	7	8	3	11
Biological transport	30	31	33	24	18	31	21	21	15	10	22	12
Artery	175	176	153	9	12	21	22	18	16	5	5	5

Figure 3. The searching technology clusters concepts and indicators to aid competitive intelligence searches. Analysts can click on highlighted answers to see details such as a document's name and authors, patent information and an abstract.

Dennis projects that STN AnaVist will expand CAS' customer base because its many applications reach beyond their traditional audience to serve additional groups such as competitive intelligence professionals, a new focus for CAS.

CHRISTINE McCUE
 Director, Marketing
 CAS
 Chemical Abstracts Service
 2540 Olentangy River Road
 Columbus, Ohio 43202
 cmcuc@cas.org or help@cas.org