Pacific Northwest
NATIONAL LABORATORY

IN-SPIRE™ Frequently Asked Questions

What is IN-SPIRE™?

IN-SPIRE™ provides tools for exploring textual data, including Boolean and “topical” queries, term gisting, and time/trend analysis tools. This suite of tools allows the user to rapidly discover hidden information relationships by reading only pertinent documents. IN-SPIRE™ has been used to explore technical and patent literature, marketing and business documents, web data, accident and safety reports, newswire feeds and message traffic, and more. It has applications in many areas, including information analysis, strategic planning, and medical research.

The goal of IN-SPIRE™ is to:

- Quickly create meaningful visualizations of the text documents
- Provide effective ways for you to explore and understand large collections of text without reading every document

What does IN-SPIRE™ do?

IN-SPIRE's strength is its ability to quickly scan through thousands of documents, determine the topical content of those documents, and then present the documents in an interactive visual context, for further analysis. Since it requires almost no advance knowledge of the information being processed, IN-SPIRE™ is a great tool for getting a feel for information hidden in a large number of documents and understanding its "topical landscape." IN-SPIRE™ provides a number of query and display tools to support deeper analysis and interrogation of the information space.

Why was IN-SPIRE™ developed?

By the mid 1990s, the information age was burying information analysts in data. Analysts had access to more data than ever before, but lacked the tools to process and assimilate the overwhelming volume and diversity of the information. Most of the information was in textual form, but in different styles, for various purposes that could not be reliably processed for information content.

Researchers at Pacific Northwest National Laboratory began to explore whether a computer program could be developed to quickly and automatically convey the thematic content of large sets of unformatted text documents. The goal was to provide technology that enables analysts to spend quality time doing real information exploration by shifting workload from processing data to analysis.

The initial research project was the basis for Information Visualization program area. The SPIRE (Spatial Paradigm for Information Retrieval and Exploration) application was one of its first software products, for the Unix platform. IN-SPIRE is for personal computers running the Windows operating system.

What types of documents can it process?

IN-SPIRE™ organizes and visualizes the topical content of multiple types of text files. These files may come from web pages, databases, results from Optical Character Reading processes, message traffic, or other sources. Currently IN-SPIRE supports encodings for ASCII, UTF-8, UTF-16 and will also ingest most types of PDF, MS-Word, MS-Excel, and RTF files, as well as email and spreadsheet sources. IN-SPIRE™ is capable of ingesting XML formatted documents, and can read various types of web formats such as HTML, and RSS/XML formats. HTML documents IN-SPIRE™ retrieves directly from the web or a local file system are cleaned of markup. New document types and encodings are routinely added and are prioritized according to demand.

What do I have to tell it about the format of my documents?

The only thing that IN-SPIRE™ must know about your document collection is how to identify the beginning of each document. For example, if you had 1000 news articles and they were each stored in a file on disk, you would identify the files to IN-SPIRE™ and specify the string of characters that occur at the beginning of each document. If you have structured fields such as titles or dates in your documents, you may identify them also, so that during analysis they may be queried separately from other document content. For example, if a field were defined for “COUNTRY:” IN-SPIRE can automatically categorize documents into separate bins for each unique country represented in the dataset.

How do I get my data into IN-SPIRE™?

You create a dataset by specifying a data source such as local files or folders or a remote web site. You may also specify additional text processing and formatting parameters, if desired. IN-SPIRE's Dataset Editor provides a step-by-step walkthrough of the process, which allows you to create a visualization of almost any set of text data.

How does IN-SPIRE™ create a visualization with my documents?

In brief, IN-SPIRE™ creates mathematical representations of the documents, which are then organized into clusters and visualized into "maps" that can be interrogated for analysis.

More specifically, IN-SPIRE™ performs the following steps:

1. The text engine scans through the document collection and automatically determines the distinguishing words or "topics" within the collection, based upon statistical measurements of word distribution, frequency, and co-occurrence with other words. Distinguishing words are those that help describe how each document in the dataset is different from any other document. For example, the word "and" would not be considered a distinguishing word, because it is expected to occur frequently in every document. In a dataset where every document mentions Iraq, "Iraq" wouldn't be a distinguishing word either.
2. The text engine uses these distinguishing words to create a mathematical signature for each document in the collection. Then it does a rough similarity comparison of all the signatures to create cluster groupings.
3. IN-SPIRE™ compares the clusters against each other for similarity, and arranges them in high-dimensional space (about 200 axes) so that similar clusters are located close together. The clusters can be thought of as a mass of bubbles, but in 200-dimensional space instead of just 3.
4. That high-dimensional arrangement of clusters is then flattened down to a comprehensible 2-dimensions—trying to preserve a picture where similar clusters are located close to each other, and dissimilar clusters are located far apart. Finally, the documents are added to the picture by arranging each within the invisible “bubble” of their respective cluster. All of this information is then mapped onto the Galaxy and ThemeView™ visualizations that convey the document and topical relationships of your information.

How long does it take to process a set of documents?

Although this is largely dependent upon the speed and capacity of your computer, IN-SPIRE™ will process a typical dataset of 3,000 documents in about 2 minutes. The software is capable of processing upwards of 100,000 one-page documents in under 45 minutes on newer desktop computer configurations. Although there are no theoretical limits on the number of documents or size of an IN-SPIRE dataset, the practical upper bounds for maintaining responsive interactions with the visualizations ranges from 30,000 to 60,000 documents.

What are the hardware requirements?

IN-SPIRE™ was developed to operate on mid-level Windows workstations and Servers. IN-SPIRE is currently running at client sites with Windows 7 and Windows 10. A typical IN-SPIRE computer would include a 2.5 Ghz processor, 8 Gb of memory, and 200 Gb disk drive. About 20-50 Gb should be available to store the program and associated datasets. IN-SPIRE will operate on much less powerful configurations, but dataset size may be limited. There are no special graphics card requirements as with earlier versions of SPIRE.

How do I install the software?

The IN-SPIRE™ automated installer guides you through the process of choosing a location for IN-SPIRE. It installs executable files, configuration files, dlls, and searchable Help files, which can be viewed in an HTML browser. On workstation installations the datasets are placed in the same tree as the program. More recent versions of IN-SPIRE (since 2010) do not require administrative privileges for individual workstation-based installation. It is recommended that you accept the default settings suggested by the Installer program.

Please note: Almost all versions of IN-SPIRE™ are copy-protected and require an unlock code before the software will operate. Unlock codes are sent via e-mail and are based on information obtained from an Activation program installed with IN-SPIRE.

Is technical support available?

It is not unusual for an analyst to start using IN-SPIRE™ without any technical training or support. The video tutorials on this website provide a good start. Most users will also benefit from a short training session that covers the key aspects of using the tool. Training sessions usually consist of a 4-6 hour hands-on class that covers the general capabilities of the system along with tips and techniques for data import and analysis. Classes are usually held at the user's site. See the Training and Support page for details and pricing.
In some cases, an organization may have greater support needs, such as datasets that require some level of pre-processing. For example, some document or spreadsheets may contain meta data that is of value but may need special handling in order to extract the desired fielded information or convert from unsupported formats (e.g., converting from a special foreign language encoding/format to a supported Unicode format). In other cases, the data may exist within a database or enterprise data store, and there may be a need to interface directly to those data stores. PNNL can assist in these cases as well, on a time and materials basis. Contact us for more information.

Can IN-SPIRE™ be integrated with my database?

Some installations of IN-SPIRE™ process information exclusively from a database interface. IN-SPIRE™ can be configured to interface with most database systems that support http:// or https:// protocols. Installation of a database interface involves some level of software customization.

To make use of IN-SPIRE™, I really need a new feature. Can it be incorporated in a future release?

IN-SPIRE™ enhancements are driven by specific project requirements that are sponsored by s by different federal agencies. These projects provide funding and implementation priorties for different aspects of the system. PNNL does it’s best to generalize the solutions to these project-specific requirements so that they can be used by all or most IN-SPIRE users. PNNLis interested in advancing this technology and welcomes the opportunity to partner with organizations that would like to sponsor new functions and features. If you are interested in becoming an IN-SPIRE™ development partner, please contact us.

What is Galaxy visualization?

In the Galaxy visualization, individual documents are represented as gray dots. With this visualization, the goal is to give you a view of your dataset where closely related documents are generally located close to each other, and dissimilar documents are far apart. It is not a perfect representation of the document relationships due to the squeezing that occurs in reducing from high-dimensional space down to two-dimensional space, but:

- It's pretty good
- It's pretty fast
- It gives you a good starting point and general overview to work with.

What are the blue shaded areas in the Galaxy?

The shaded areas on the Galaxy are "ThemeClouds" which are analogous to ThemeView™ Peaks. ThemeClouds provide a two-dimensional representation of theme strength. Areas with higher thematic content and/or document density are more intensely colored in blue. Areas with less document density and thematic content are more lightly colored.

What is the ThemeView™ visualization?

The ThemeView™ visualization is the fastest way to get an overview of your document collection. It translates the Galaxy into a three dimensional “landscape” of your information space.

Think of the Galaxy as the “flat” sea-level foundation for a ThemeView. Each document which has content related to a major theme in the overall document collection will add a little to the height of the peak in that location (how much it adds will depend on the strength of that theme's relevance to that document). If a document is not at all related to that theme, it won't add any height to the layer there. Repeating this layer-building process for all 200 or so major themes (i.e., topics) in the dataset, stacking the layers on top of each other and smoothing the results, creates the thematic summary view, the ThemeView.

What does the ThemeView™ peak height and color tell me?

The labels flagging the peaks reveal what the strongest themes are under those peaks. Areas of documents having very similar thematic content contain tall peaks, while areas of documents having weaker thematic relationships to each other never rise above sea level. The coloring of a ThemeView™ lets you know how far above sea level a region is—yellow being the highest. If the documents in a region are practically void of any thematic content, they are represented at sea level height on the ThemeView. If there are only one or two documents in a region, but they are unusually packed full of topical content, they are represented as tall peaks on the ThemeView.

How are the ThemeView™ peak labels related to the cluster labels?

The ThemeView™ landscape is created by piling up the topicality of individual documents, so you will generally see higher peaks in areas of high document density. The number, placement, and height of peaks are really only an indirect correlation to the clusters, however, since they are based strictly on the Galaxy documents underneath, not the cluster groupings. An area under the peak may, and often does, include documents from multiple clusters.

In addition, the words used to label the cluster centroids are terms with the highest frequency count only, whereas the ThemeView™ labels are the words with the highest topical content in the region. These factors help explain why the ThemeView™ peak labels often differ from cluster centroid labels.

What if I have some text that isn't in English?

IN-SPIRE™ visualizations are language-independent, although the use of system or custom stopwords is recommended for optimal visualizations. For some languages such as Chinese, pre-processing with a segmentation tool may be necessary. . If your data contains multiple languages, the documents from one language use very different terms than documents from another, and you will find that the visualizations naturally show the divisions.

If the documents are in a language you cannot read, IN-SPIRE™ does support some third-party language detection and machine-translation software. Queries from your native language can be translated into the language used in the documents, and the Galaxy and ThemeView™ labels, and documents titles and text are translated on demand.