
**Pacific Northwest
National Laboratory**

Operated by Battelle for the
U.S. Department of Energy

LER Data Mining Pilot Study Final Report

J. Young
M. Zentner
D. McQuerry

October 2004

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830



DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY

operated by

BATTELLE

for the

UNITED STATES DEPARTMENT OF ENERGY

under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: <http://www.ntis.gov/ordering.htm>



This document was printed on recycled paper.

(9/2003)

LER Data Mining Pilot Study Final Report

Letter Report

October 2004

Prepared for the

U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research

By

Pacific Northwest National Laboratory
M. Zentner
D. McQuerry
J. Young

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

Acronym List

Acronym	Definition
ASCII	American Standard Code for Information Interchange
BWR	Boiling Water Reactor
CRDM	Control Rod Drive Mechanism
EDG	Emergency Diesel Generator
GIS	Geographic Information System
HPCI	High Pressure Core Injection
HPCS	High Pressure Core Spray
HTML	Hypertext Markup Language
LER	Licensee Event Reports
LOOP	Loss Of Offsite Power
NRC	Nuclear Regulatory Commission
OCR	Optical Character Reader
PC	Personal Computer
PDF	Portable Document Format
PERL	Practical Extraction and Report Language
PNNL	Pacific Northwest National Laboratory
PWR	Pressurized Water Reactor
SPIRE	Spatial Paradigm for Information Retrieval and Exploration
Starlight	Starlight Information Visualization System
XEE	XML Engineering Environment
XML	Extensible Markup Language

LER Data Mining Pilot Study Final Report

Letter Report

1.0 Introduction

Licensee Event Reports (LERs) consist of a one page standard form with a standard header and free text data, followed by additional continuation pages of free text data. Currently this LER data is analyzed by first inputting the heading and text data manually into a categorical relational database. The data is then evaluated by enumeration of data in various categories and supplemented by review of individual LERs. This is labor intensive and makes it difficult to relate specific descriptive text to enumerated results. State of the art data mining and visualization technology exists that can eliminate the need for manual categorization, maintain the text relationships within each report, produce the same enumerated results currently available, and provide a tool to support potentially useful additional analysis of the informational content of LERs in a more timely and cost effective manner.

1.1 Purpose of Study

The purposes of this pilot study are to demonstrate 1) the usefulness of data visualization techniques to evaluate reports of incidents and events occurring at nuclear facilities without the need to input the data into a categorical, relational database, and 2) to show that the use of this technology can both lower costs of analysis and provide additional insights into accident precursors, potential common cause failures, system performance, plant operation, and other safety related issues.

Data visualization techniques have been developed to:

- Directly analyze free text and/or categorical data;
- Improve analysts ability to comprehend complex interrelationships present in the data;
- Help identify patterns and trends in large data sets;
- Analyze data without the need to “predetermine” all the categories, ontologies, and taxonomies the data may fall into; and
- Eliminate enforced reliance upon "stovepipe" information systems that address only individual aspects of typically multifaceted problems.

The visual information analysis tools used in this study are designed to integrate and concurrently analyze the contents of large, complex, multimedia information collections. They incorporate advanced information models, a suite of pattern recognition algorithms and a variety of effective visualization tools into their design.

Two types of tools are used in the study:

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

- a. A tool to effectively analyze the relationships in unstructured text data. The tool selected for this analysis is a PC application named IN-SPIRE.
- b. A tool intended primarily for use on structured data to explore the correlations, trends, geographic distribution and relationships of data objects. The tool selected for the correlation analysis is named Starlight.

1.2 Scope of Study

For this study, plant LERs from the period 1998 to 2003, consisting of over three thousand six hundred documents, were analyzed. The LER data consisted of two files:

- a. The free text portion of the LERs
- b. An Access data base version of the first page LER header data

The two files were merged to create a database was constructed to capture the native and derived features of the information in an explicit information LER dataset.

The LER dataset was then evaluated using two separate data visualization tools, IN-SPIRE (see Appendix A for detailed description) and Starlight (see Appendix B for detailed description), to demonstrate examples of the interrelationships among the database elements and their properties that can be identified by this technology.

A sample problem analysis of a specific type of event, “*emergency diesel generators failure to start*” was performed using both tools. Additionally, two further event types were analyzed using Starlight. 1) Events related the electric power grid failure of August 14, 2003 and 2) reports related to control rod drive mechanism (CRDM) leakage.

Additional analyses of the LER data were performed to demonstrate the potential capabilities of both types of data visualization tools. This included trending, grouping, and mapping analyses.

2.0 **Summary of Results**

Two data visualization software tools, IN-SPIRE and Starlight were used in the pilot study. The results of the study are as follows:

- Using the appropriate query or search strategy, an analysis can be performed using the LER header and free text data identify events of a specific type without requiring preparation of a relational database. This means that textual data can be directly analyzed without a pre-interpretation process. For example, eight (8) Emergency Diesel Generator Failure to Start events were found.
- The language used in the LER can effect the results of the analysis. Therefore, alternate query strategies and/or language normalization will need to be applied together to produce robust, comprehensive results. For example, a search for Loss of Offsite Power events yielded only four such events on 14 August 2003.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

However, also searching on the date (14 August) and combining the two sets of queries yielded all six LERs generated for LOOP events associated with the northeast blackout. The additional search was necessary because the LOOP event was referred to as loss of grid or disturbance on the grid.

- This technology eliminates the need to preprocess the data and use tools such as relational databases to perform analysis of events, event types, etc.
- Temporal and geographic relationships can easily be shown from information contained within the reports using data visualization tools.
- Trending can be performed in a variety of ways depending on the needs of the particular analysis. For instance, analysis of temporal trends, equipment failure trends, or trends in failure types can be performed using either IN-SPIRE or Starlight.
- Analysis using the various types of visualization tools is complementary. For the IN-SPIRE analysis, LERs were reviewed to identify those related to emergency diesel generators failure to start. Using insights into the data structure and terminology gained from this part of the analysis an enhanced study was performed using the Starlight tool,

3.0 Data Conditioning

The first attempt to process new data to use in solving new problems is a time consuming activity that requires developing and refining data engineering approaches. Once a well designed data processing approach has been developed, it can be automated and reused on additional data. During this exercise a number of routines and techniques were developed along with the requirements for future automation of these designs.

3.1 Data Description

The original LER data was provided to PNNL in two separate formats, PDF (Portable Document Format) format and an Access database containing extracted and derived data fields. The Access database files contained structured data files summarizing each LER. The PDF files contained the original LER information. The need for the Access data in addition to the PDF-derived text is that the information in the text file is entirely unstructured. Accordingly, the Access database provided an effective and inexpensive way to generate reliable, structured fields such as "Date of Event", "Facility name", etc.

3.2 Data Conditioning Activities

3.2.1 Data Engineering

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

To import this data into Starlight and IN-SPIRE, each PDF file was converted into a text file and the each row in the Access database was exported as an XML¹ file. A PERL² script was used to insert the text files into the XML format as an additional XML element. Once the files were converted and matched, the documents were processed by a tool developed at PNNL, XEE (XML Engineering Environment), for manipulating data. This processing included tagging the primary items such as the LER Number, Plant Name, Plant type, relevant dates, and LER Title with XML element tags. Additionally, the geo-coordinate (longitude and latitude) for each plant was inserted into the data to allow mapping.

3.2.2 Data Modeling

Data relationships were defined and parameters for the visualizations were determined. Data field properties were assigned and “Stop Word Lists³” were created to improve the text processing accuracy by eliminating extraneous, non-descriptive words and phrases. The data was then processed by both IN-SPIRE and Starlight to produce their respective interactive visualizations.

3.2.3 Data Normalization

It was initially assumed that considerable effort would be needed to invest in “normalizing” the language used in the LERs. For example, one LER might refer to “EDG failed to start,” while another might say “emergency diesel generator failed to automatically start.” Both LERs are addressing the same issue, but using different language. It was thought that if queries were to be formulated that would successfully identify all LERs related to a particular issue, without inadvertently including a number of irrelevant documents, it would be required to have a method for altering the language using an ontological approach, such that all LERs related to a particular subject would be using the same terms to discuss that subject. After a relatively short time it was discovered that suitable queries could be devised that would locate all the documents relevant to a particular inquiry without incorporating a lot of superfluous unrelated documents. For more detail on this approach, see section 4.2.2 describing the EDG analysis.

4.0 LER Analysis

4.1 Analysis Objectives

Demonstrate data mining and data visualization technology that can:

¹ XML: "Extensible Markup Language" - A flexible format for structured documents.

² PERL: "Practical Extraction and Report Language" -- An interpreted language optimized for scanning arbitrary text files, extracting information from those text files, and generating output to a printer or file. It is commonly used for system management tasks and text manipulation, among other things.

³ Stop Words: A list of words that are provided to a text analysis algorithm which are to be ignored by the algorithm when they are encountered in an associated dataset.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

- produce results currently published by NRC directly from the LERs
- Identify additional insights into events directly from the LER data at reduced cost

4.2 IN-SPIRE Analysis

IN-SPIRE presents the data in two different visual metaphors. The first of these, the Galaxies visualization (Figures 1-2), uses a point on the screen to designate a document. Documents which are thematically similar tend to be located near each other on the screen. That is, points on the screen which are close together represent documents that are related to each other. The resulting visualization resembles stars in the night sky – hence the name, Galaxies. The second IN-SPIRE visual metaphor consists of a thematic “terrain map,” called the Themeview (Figure 3), which represents areas of high thematic density with mountain peaks. These peaks tend to parallel the placement of the clusters in the Galaxies display, since they are calculated from them. That is, the Themeview and Galaxies view both show the documents in the same location in the X and Y dimension. But while the Galaxies view is showing the individual documents and how they cluster together, the Themeview is showing the distribution of themes amongst those clusters.

The visualizations created with IN-SPIRE are based upon signatures which the software calculates for each document, based on a series of statistical algorithms. Consequently, the first phase of analysis within IN-SPIRE consists of tuning and refining the visualization. This is done by reviewing the preliminary visual representation and identifying terms which introduce “noise” into the visualization. These “noise” terms are words which tend to produce misleading results if not added to the Stop Word List. Putting a word in the Stop Word List will remove it from the calculations used to determine cluster membership and the individual signatures of the documents. A typical “noise” term would be something like “heretofore,” or “page.” These words in no way serve to differentiate the documents or tell anything meaningful about their content. A key place to identify potential candidates for the Stop Word List is the list of Major Terms. These are the words which are specifically being used to calculate the vectors for each document. The list is available by clicking on the Major Terms tool in the Tool Bar. A quick review of this list will often yield several words which need to be removed from the analysis.

Once the list of Stop Words was complete, the focus shifted to the document cluster display. There were three sets of “outliers” in the Galaxies visualization (Figure 1). The first one investigated contained two documents. Both were related to security violations; specifically to contractors being denied unescorted access to the facility because they’d failed to reveal their criminal history in their applications. The second and third outlier clusters consisted of two group of Safeguards LERs which contained no abstract or text section. Consequently they all had an identical signature, based upon the text from the titles, all of which were the same.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

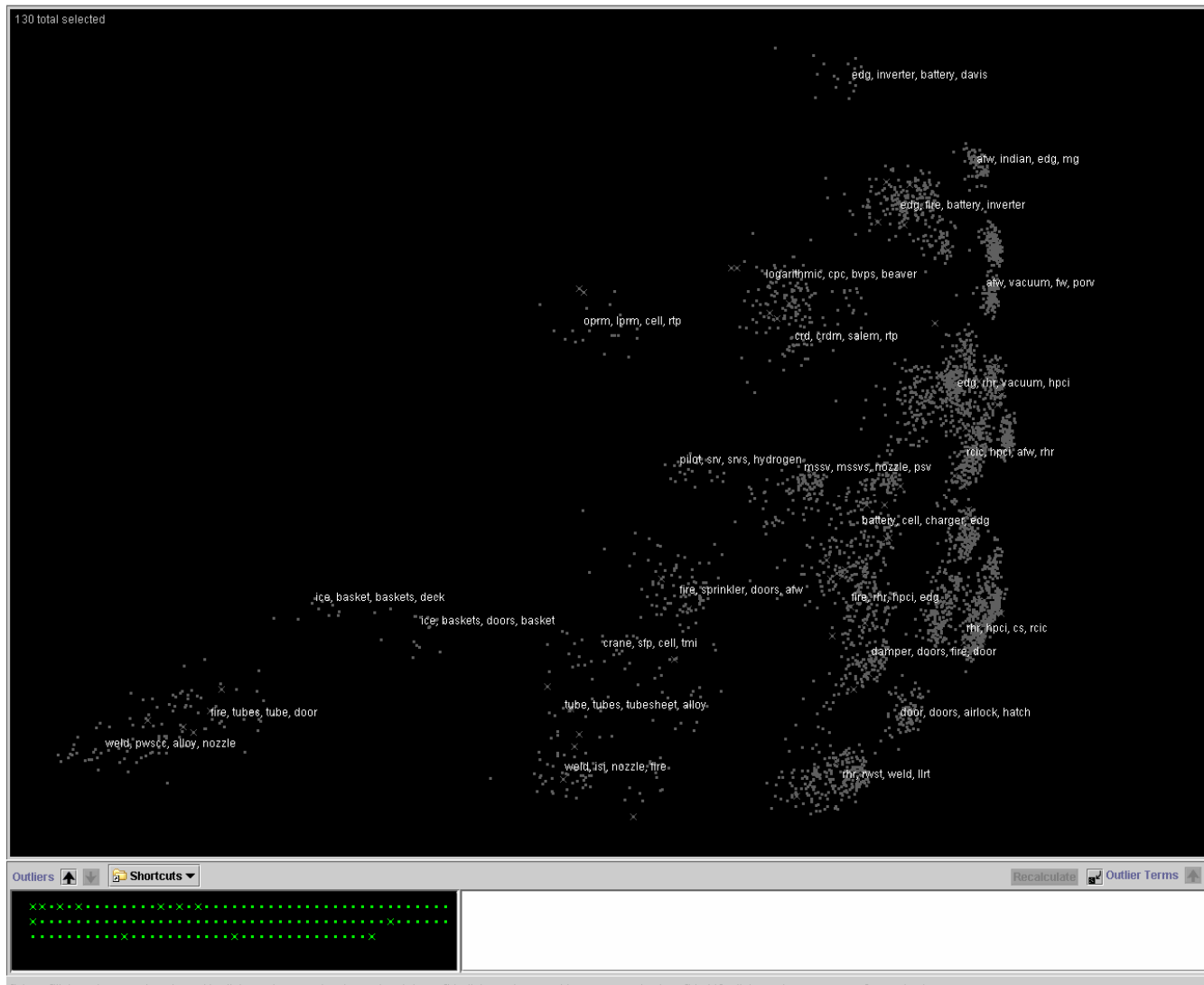


Figure 2 Outlier clusters removed and Galaxies display recalculated

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

All three of these clusters (130 documents in all) were added to the “Outlier Box” in IN-SPIRE and the Galaxies view was recalculated to show the updated relationships. After the Galaxies view has been recalculated, the relationships between the different issues discussed in the LERs becomes more evident (Figure 2). Note that the outliers, highlighted in green in the box below the Galaxies view, remain visible in the display. This enables the analyst to continue to work with them: they will respond to queries and can be re-introduced into the larger Galaxies display at any time. Removing them allows a better understanding of the relationships amongst the other documents.

4.2.1 General Analysis Case

The first step in an IN-SPIRE analysis generally consists of a high-level exploration of the dataset. This is accomplished first by exploring the regional “peak labels” from the Themeview visualization. These labels can be overlaid on the Galaxies view, as has been done in the previous figures, or used directly on the Themeview display (Figure 3).

It is also possible to explore individual clusters – either through the use of the Gisting⁴ Tool, that is, select a cluster and generate a list of gisted terms – or by turning on the individual cluster labels to see the top three terms in each cluster.

The next step in the analysis involves executing queries to see how different terms are distributed. A variety of terms likely to represent specific sources of risk, such as vulnerable subsystems, plant names, BWR vs. PWR reactors, etc. were selected. Figure 4 shows the groups created during this step.

After the creation of groups from the series of queries used in the initial explorations, the groups were explored in greater detail, combining them in various ways to identify correlations and looking at them over time to identify potential trends. In addition, the “browser” tool⁵ was used to examine individual documents that had been identified through a particular query of a series of queries.

⁴ Gisting: A technique for extracting the central theme of a document or set of documents based upon simple word-occurrence statistics.

⁵ The *browser tool* allows analysts to review document contents in detail.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

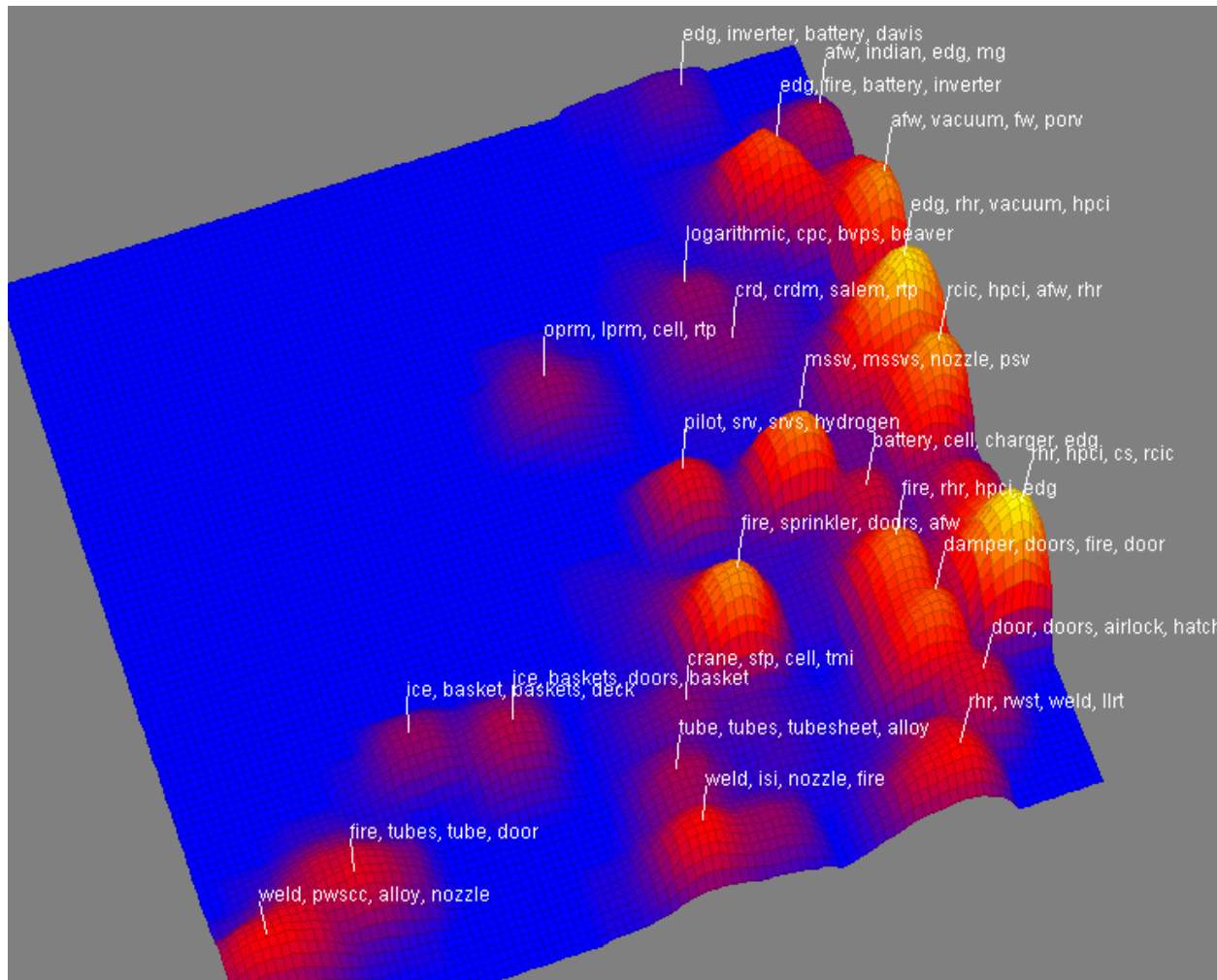


Figure 3 Peak labels help identify general content distribution

PNNL-14910: LER Data Visualization Analysis Pilot Study

10/15/2004



Figure 4 Groups created by queries are shown on the right side of the figure. In the Galaxy view, PWR and BWR LERs are respectively highlighted in red and yellow.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

Query List

Power load
“fast” “3 start”
What: “EDG” “4 fail” “3 Sta”
Systems
EDG or “Emergency Diesel Generator”
hpci
tube
mssv
condensor
crdm OR 'control ro
containment
actuat'
pump'
feedwater
damper
door'
batter”
rcic
rhr
hpcs
Risk Issues
ice
fire
hydrogen
steam
radiation
security
shutdown or 'shut do
Plant type
BWR
PWR

The query list recreated to the left shows the initial queries that were performed. The software assigns colors to the document groups identified by each query as shown. The documents can then be seen in the Galaxy display. After a query has been performed, the results are saved and can be used in other operations. An examination of the results of each of these queries provides insights to the analyst about the structure of the data, the way that the documents cluster together, and the information contained therein. A simple example is demonstrated in Figure 5. An examination of the query list at the left shows that two of the queries were run to identify documents that contain the word “ice”, and another query that contains the plant type “PWR”. At the bottom of the list it shows that 2446 documents were identified in these two queries, of which 68 contained both terms.

An examination of the Galaxy view related to this query, Figure 5, shows that the documents related to PWRs are colored red, the documents related to “ice” are purple, and those that overlap are blue. It can be seen on Figure 5 that there is a “cluster” of blue documents, those related to both “ice” and PWR. An examination of the documents contained in this blue cluster reveals that these are related to problems of PWRs with ice condenser containments.

Total Documents - 2446

Overlap--68

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

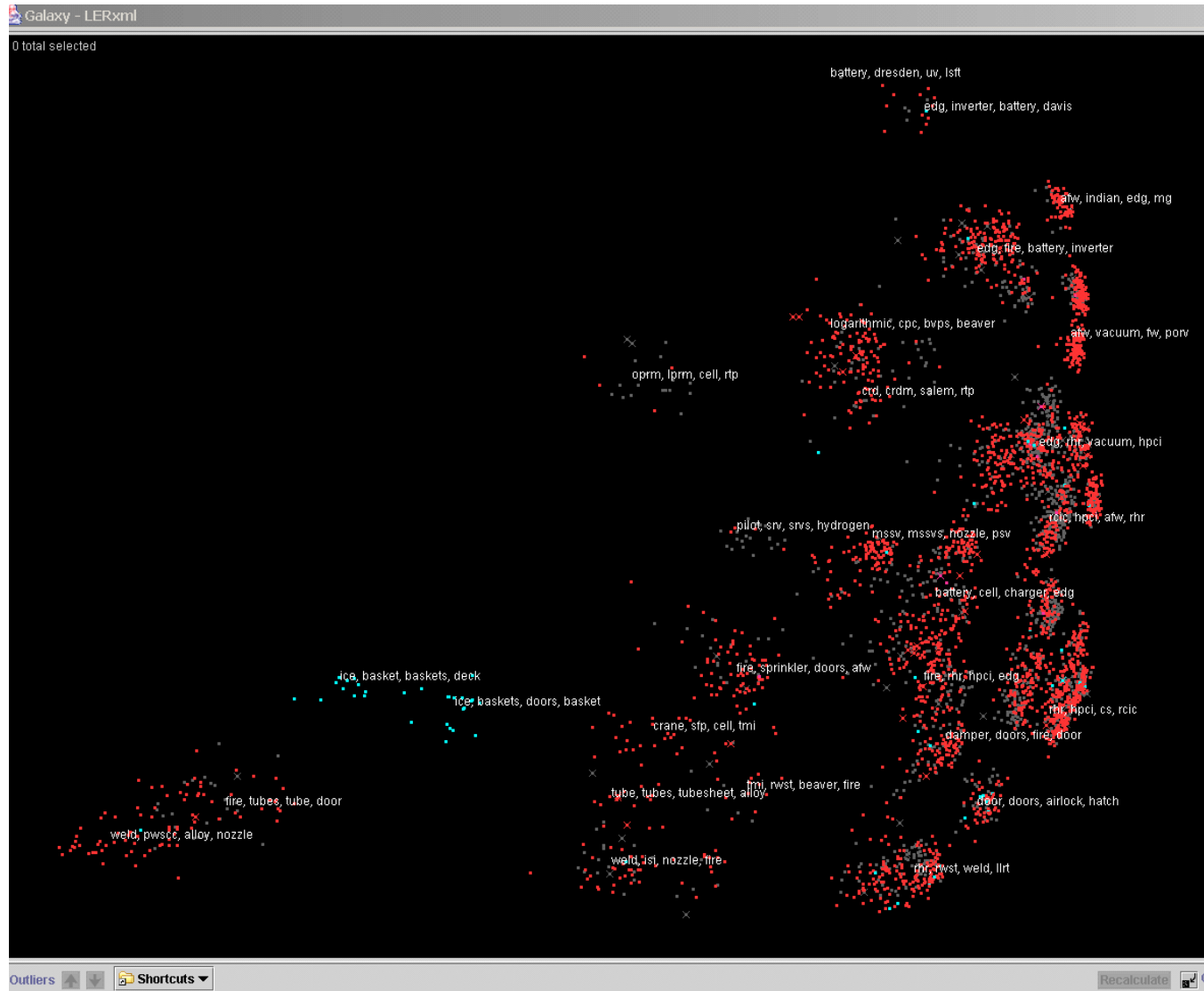


Figure 5 Union of search terms "ice" and PWR

PNNL-14910: LER Data Visualization Analysis Pilot Study 10/15/2004

4.2.2 EDG Analysis Case

One example of in-depth analysis in IN-SPIRE was the investigation performed on “EDG failed to start.” As has already been mentioned in the section on language normalization, this was initially expected to be a time-consuming operation, with considerable up-front work to standardize the terminology. However, the investigation required relatively little effort, since queries were easily developed that were quite successful at identifying the relevant documents almost immediately. In addition, non-responsive documents such as ones that contained phrases such as, “if we had run the test under these conditions it is possible that the EDG would have failed to start” were able to be quickly eliminated. Since the intent was to locate LERs describing situations in which the EDG had actually failed to start, that document was not one that fit the intended criteria. The query strategy employed enabled was able to reduce the number of responsive documents sufficiently that it was possible to manually review them for relevance.

Several query strategies were employed before discovering the correct approach. The initial search was on “*EDG OR 'emergency diesel generator'.*” This returned 459 documents which contained references to EDGs. Then a query on “*'fail* *3 start'.*” was executed. This is a form of proximity search combined with a specific phrase search. The search looked for any documents that contained the four-letter sequence, “fail*” (this would include, fail, fails, failed, failure), within 3 words of the word, “start”. The results included 80 documents. When these two queries were combined, it was found that there were 25 documents that satisfied the criteria for both queries. However, in many of these the two phrases being looked for were unrelated. For example, the “failed to start” phrase may have been referring to a system other than the EDG. So while both concepts; “failed to start” and “EDG” were contained within all of these documents, the majority of them had nothing to do with EDGs failing to start.

In order to focus on only the documents that contained references specifically to EDGs failing to start, both searches were combined into a single composite proximity search, thus assuring that the “failure to start” reference would be linked to the “EDG” reference. The resulting query, “*'EDG *4 fail* *3 start' OR 'emergency diesel generator* *4 fail* *3 start'.*” returned only 8 documents in which the reference to “EDG” occurred within 4 words of the reference to “fail*,” which in turn occurred within 3 words of the reference to “start.” Having such a small number of responsive documents made it simple to manually review them for relevance. It was discovered that six of the eight documents contained explicit references to actual failures of EDGs to start. The other two were references to hypothetical failures of EDGs.

4.3 Starlight Analysis

Performance of a Starlight analysis involves the use of visual metaphors to allow an analyst to perceive patterns, anomalies and form hypotheses based on the observed data relationships. The ability to manipulate the visualizations allows analysts to test hypotheses and drill down to individual documents and records to retrieve specific information and develop an understanding of key data relationships.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

4.3.1 General Analysis Case

The process of performing an analysis using Starlight begins with what is called a “General Analysis”. In the General Analysis process, the data is presented and evaluated in several different high level views, described as follows.

In the *Context View*, the concepts contained in the reports are captured and reports dealing with similar concepts are clustered together in a simulated 3-D display (Figure 6). Figure 6 displays the clusters of the top three key concept words in each document and their proximity locations to other clusters. The yellow and red encodings distinguish between pressurized water reactor (PWR, yellow) and boiling water reactor (BWR, red) plant types.

By using the query tools subsets of records are created. A subset might deal with a specific problem, plants in the same region, or incidents that occurred in a given time period.

In the iSpace view window (Figures 7 and 8), a map of the continental United States shows the locations of United States nuclear power plants. Using this tool, records from the concept clusters can be connected to the plants and their locations, The Link Array is a query visualization that summarizes the values in different fields of a record and then shows the relationships between fields. One link array showed the relationships among Document ID, Plant Name, Plant location, and LER Year. By viewing a temporal filter on LER Year, a time sequence showed which plants were reporting LERs, where the plants were located, the frequency of LER reports, and which documents were tied to which plants. Figure 7 shows the frequency of occurrence for LERs for the Indian Point power plants.

The following discussion describes how these approaches were used to demonstrate how Starlight is used.

4.3.2 Evaluation of Emergency Diesel Generator Failures to Start

The EDG Analysis began with a text query for all of the LERs to determine which contained the character strings “EDG”, or “Emergency Diesel Generator”, or “EDGs”. Next, another Text Query was performed against the entire record set to identify all LERs that contained the strings “Failure”, or “Fails”, or “Failed to start” or “Failed to initiate”.

The datasets resulting from the two were then “intersected” which resulted in a new record set that held all of the LERs containing both results. Documents contained in this new data set were then examined. It was determined that several different types of emergency diesels were contained in this data set. In particular, some plants had emergency diesel power pumps on their high pressure core spray (HPCS) or high pressure core injection (HPCI) systems.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004



Figure 6 Starlight Context View

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

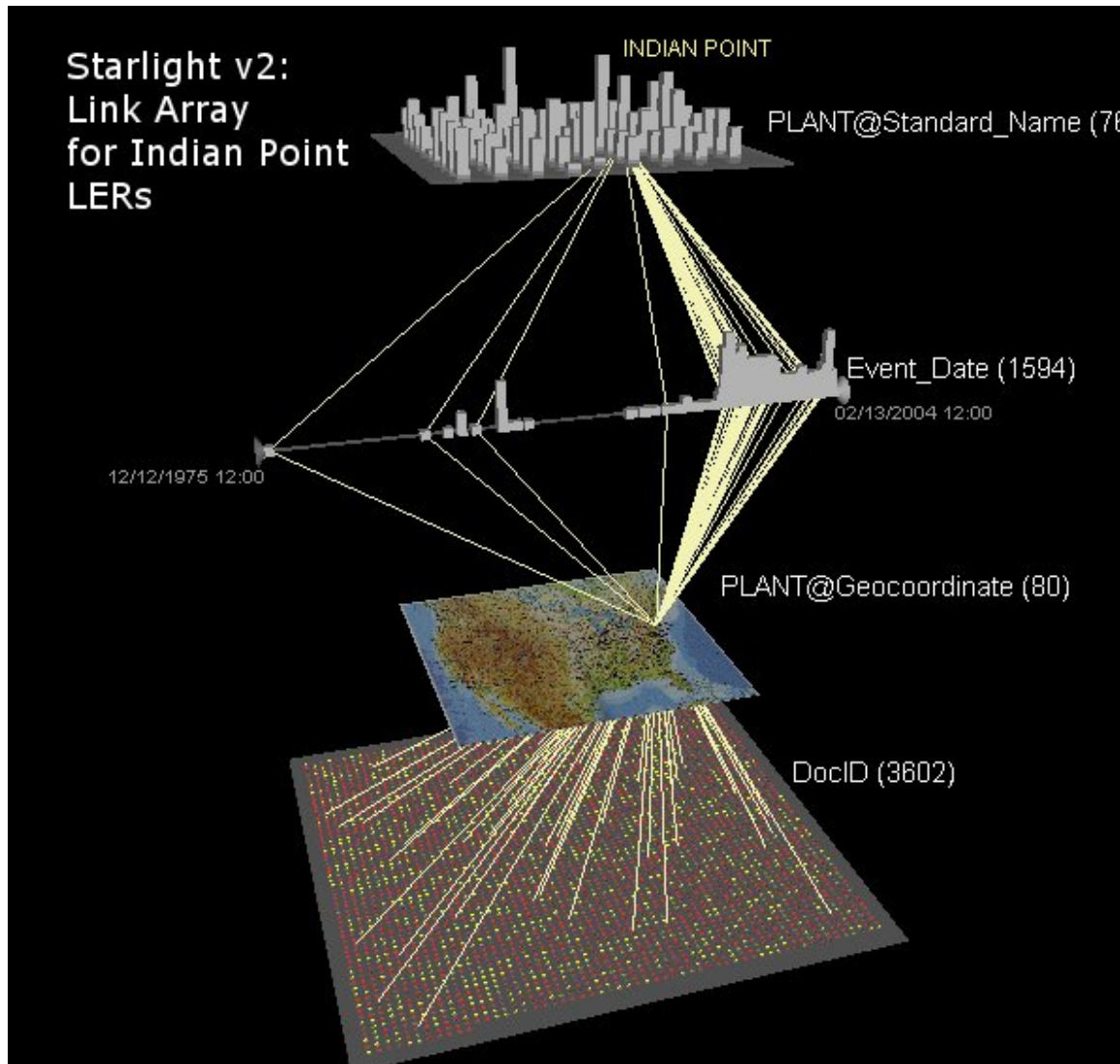


Figure 7 Temporal trends in LERs at Indian Point Nuclear Power Plants

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

Since the purpose of this task was to identify failure to start of emergency diesels that provide back-up electric power to the entire plant, a new query was developed to eliminate these LERs.

After further evaluation, a final subset was identified that contained six records that describe a “failure to start due to an EDG”. This confirmed the results of the IN-SPIRE analysis.

4.3.3 LOOP Analysis Case

To demonstrate how Starlight could be used to examine a common cause failure mechanism, it was decided to evaluate events related to the multi-state electric grid power failure to evaluate loss of offsite power events (LOOP). The first step in the LOOP Analysis began with a Temporal Query on all LER Event Dates to identify events that occurred on “8/14/2003”. Next, a Text query was initiated against all the records to find the records which contained “Loss of Power” or “LOOP”. The results of these two queries created a new record set which contained all the LERs which reported an event date of “8/14/2003” and also reported a “Loss of Power”. As a check, a further query was performed to identify any LER that contained the phrase “grid upset”. An additional power plant was identified that used that phrase to describe the events of August 14, 2003. When viewing the results, the relationship was displayed as the geographical map was connected to each of the Plant names which reported an LOOP. (Figure 8) It can be seen that as expected all the plants were located in the northeastern power grid area, where the power loss occurred.

Although this example seems simple and obvious, the approach demonstrates the utility of using geographical data to evaluate locations of a particular failure type. This could be used to provide a relatively simple way to evaluate potential environmental or geographical effects on nuclear power plants.

4.3.4 Leakage in Control Rod Drive Mechanisms

A final analysis was performed to identify LERs associated with leaks related to control rod drive mechanisms (CRDM). The first query proposed to identify LERs related to CRDMs. The next step was to identify LERs related to leaks. When these two record sets were combined, it was discovered that LERs had been written for many types of leaks, and leaks related to seal failures dominated the resulting record set. A new query was performed to identify LERs identifying problems with welds. When this record set was intersected with the two previously developed record sets, five LERs; three from PWRs and two from BWRs, were identified as being related to CRDM leakage from welds. Figure 9 displays the results.

5.0 Comparison of IN-SPIRE and Starlight Results

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

As expected, IN-SPIRE and Starlight both identified the same LER dataset describing emergency diesel generator failures. The thinking behind the LER analysis process was simplified for the Starlight work using insights in the query procedures gained during the initial use of IN-SPIRE.

Each tool has its own strength and weakness as an analytic tool. The query structure used in IN-SIRE allows for proximity searches for important words and terms. For instance, in the EDG assessment, the query structure “*‘EDG *4 fail* *3 start’ OR ‘emergency diesel generator* *4 fail* *3 start’*” searched for EDG used within four words of “failure” or “failed” or “fail” and with in three words of “start”. Additionally, IN-SPIRE can use the *gisting* process (a technique for extracting the central theme of a document or set of documents) to select documents.

Starlight uses a conceptually similar but a slightly differing approach to finding information about document relationships. Two basic types are used, word association queries and content queries; i.e. queries specifying a set of constraints (e.g., a concept, pattern or field and values) that documents must meet to be considered.

Both IN-SPIRE and Starlight revealed sets of data outliers, in particular those related to security issues and Safeguard LERs. When those were detached from the analytic process, document thematic grouping became clearer.

The IN-SPIRE tool allows incorporation of sophisticated statistical analysis methodologies to provide a more detailed and quantitative understanding of the information contained in the data under consideration. Starlight visualization tools employ information visualization models capable of effectively capturing multiple types of relationships that may exist among information of disparate kinds.

This pilot study demonstrates the complementary capability of IN-SPIRE and Starlight to process large sets of documents, establish the document’s relevant content, and then provide an interactive visual display that can enhance and provide guidance for further investigation.

6.0 Conclusions

The results of this pilot study shows that using the appropriate query or search strategy an analysis can be performed to identify documents addressing specific issues with out requiring preparation of a relational database. This means that textual data can be directly analyzed without a pre-interpretation process. Temporal and geographic relationships can easily be inferred from information contained within the reports. Trending can be performed in a variety of ways depending on the needs of the particular analysis.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

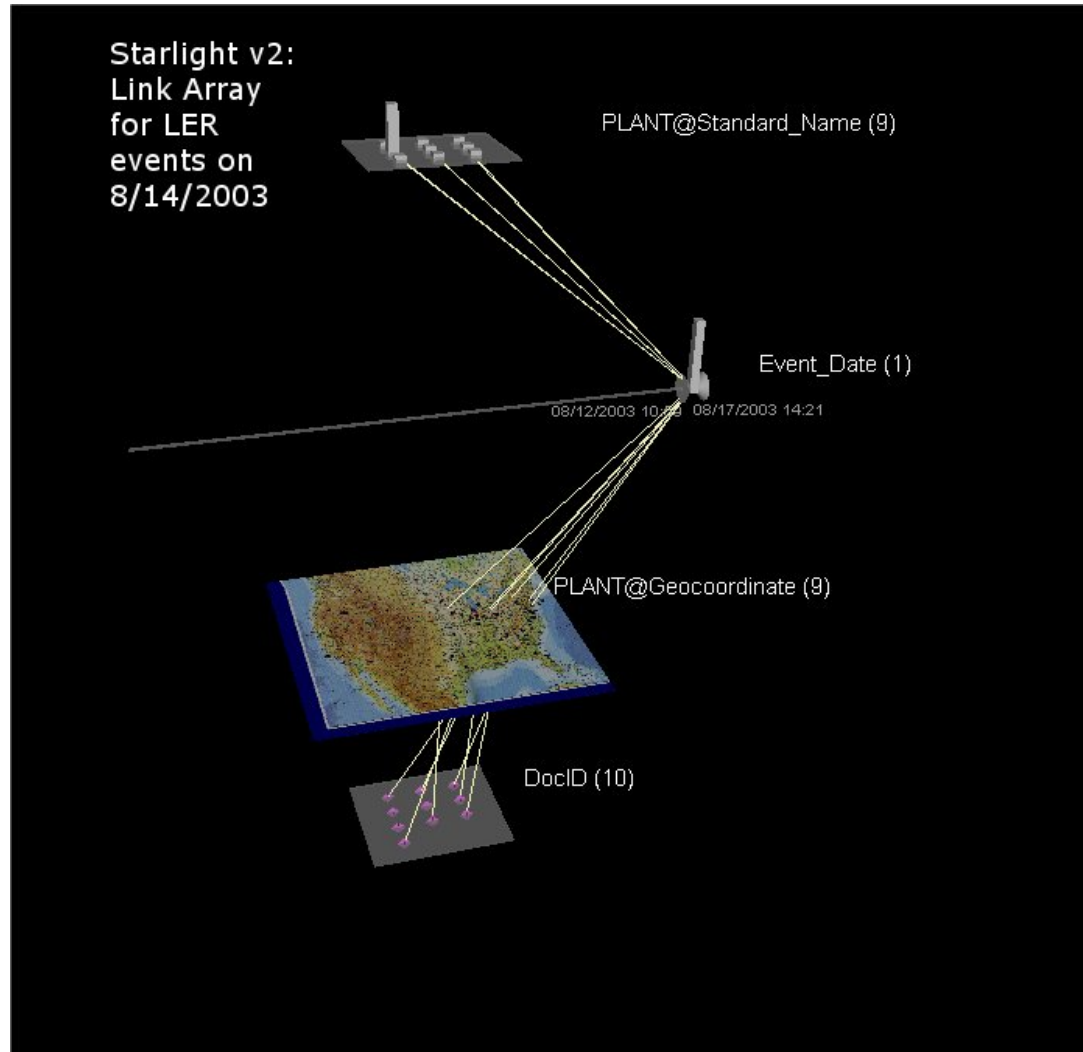


Figure 8 Loss of offsite power events for 8/14/2003

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

For instance, analysis of temporal trends, equipment failure trends, or trends in failure types can be performed using either IN-SPIRE or Starlight.

The capability to perform these types of analyses can be impacted by the loose language structure used in the LERs. For example, in the Loss of Offsite Power analysis of August 14, 2003 there was a discrepancy in the terminology used. In all but one of the LERs describing plant shut down as a result of the multi-state electric grid power failure, the terminology “*Loss of Offsite Power*” was used to describe the cause. In one LER, the cause was termed a “*Grid Upset*”. However, alternate analytical strategies can be used to properly identify similar events using differing terminology.

This pilot study primarily demonstrated approaches for identifying and classifying documents and groups of documents. The study demonstrated the use of technologies that can enable analysts to spend quality time doing real information exploration, emphasizing analysis rather than processing data. In this pilot study, the analysis was performed to understand and investigate the information provided by specific examples of thematic grouping of LERs. When necessary, other analytical approaches can be used, including the application of sophisticated statistical analysis methodologies with IN-SPIRE to further understand and describe the collective information about nuclear power plant operation contained in the LER database.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

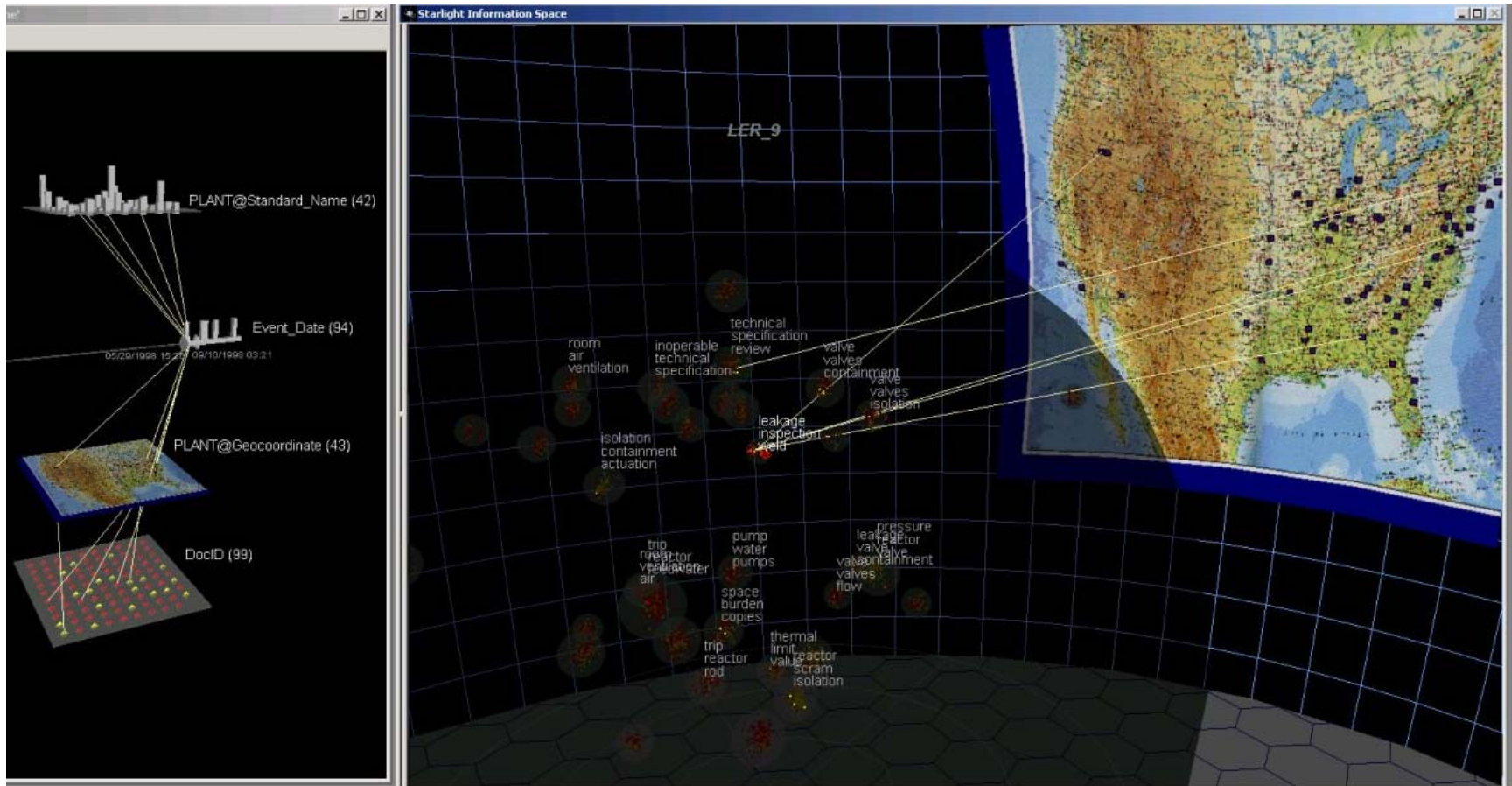


Figure 9 iSpace and Link Array View of LERs related to CRDM Weld related leakage

Appendix A: InSpire Descriptions

1. 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 interactive visualizations of the text documents
- Provide effective ways for users to explore and understand large collections of text without reading every document.

2. What does it 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.

3. Why was IN-SPIRE™ developed?

By the mid 1990's, 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.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

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.

4. What types of documents can it process?

IN-SPIRE™ organizes and visualizes the topical content of ASCII or XML text files. These files may come from web pages, databases, OCR documents, message traffic, or other sources. They must be available to IN-SPIRE™ as ASCII text (i.e., plain text) or XML. IN-SPIRE™ currently cannot read documents in special formats such as MS Word or PDF.

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

The only thing that IN-SPIRE™ must know about a document collection is how to identify the beginning of each document. For example, if a dataset consisted of 1000 news articles stored in a single file on disk, the user would identify the files to IN-SPIRE™ and specify the string of characters that occur at the beginning of each document. Structured fields such as titles or dates in the documents may be identified also, so that during analysis they may be queried separately from other document content.

6. How do I get my data into IN-SPIRE™?

You create a dataset by specifying a set of source data (files or a folder containing files) and the text that identifies the beginning of each document. 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 the user to create a visualization of almost any set of text data.

7. How does IN-SPIRE™ create 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.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

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 the information.

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

Although this is largely dependent upon the speed and capacity of the computer used for the analysis, 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 30 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.

9. Is technical support available?

It is not unusual for an analyst to start using IN-SPIRE™ without any technical training or support. However, most users will 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 in order to extract the desired fielded information or convert from unsupported formats (e.g., HTML to ASCII). PNNL can assist in these cases as well, on a time and materials basis. Contact us for more information.

10. 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

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

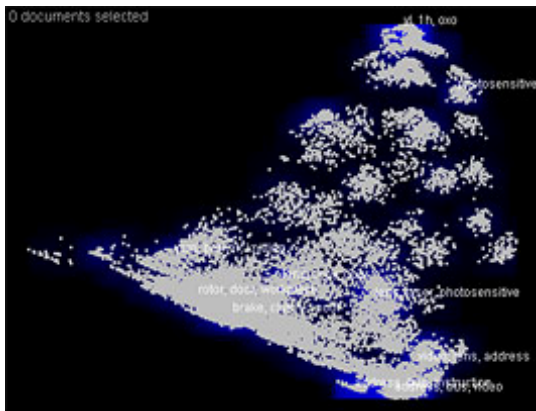
support http:// or https:// protocols. Installation of a database interface involves some level of software customization.

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

Development of the current version of IN-SPIRE™ has been supported by multiple federal agencies, each providing funding for different aspects of the system. PNNL is 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, contact Dennis McQuerry.

12. What is the Galaxy visualization?



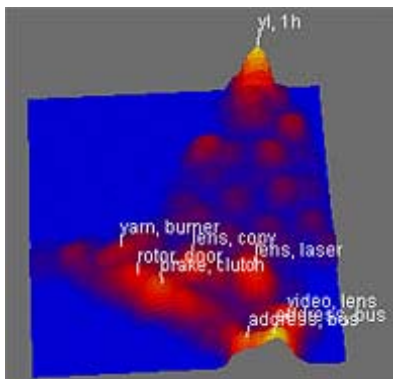
In the Galaxy visualization, individual documents are represented as gray dots. With this visualization, the goal is to give you a view of your dataset where closely related documents are generally located close to each other, and dissimilar documents are far apart. It is not a perfect representation 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
- * And it gives you a lot to work with

13. 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, where areas of higher thematic content are more intensely colored.

14. 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 thematic summary landscape map.

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

Think of the Galaxy as the sea-level foundation of a ThemeView. Each document having content related to that theme will add a little height to that layer (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 topics in the dataset, stacking the layers on top of each other and smoothing the results, creates the thematic summary view, the ThemeView.

15. 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.

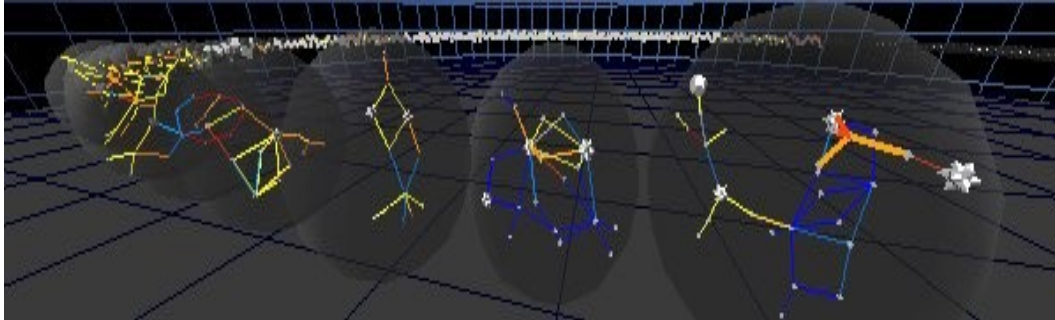
16. How are the ThemeView peak labels related to the cluster labels?

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.

Appendix B: Starlight Description

1. Introduction



To understand is to perceive patterns.
- Isaiah Berlin

The Pacific Northwest National Laboratory's *Starlight Information Visualization System* (Starlight) is a forerunner of an emerging new class of information system, one that couples advanced information modeling and management functionality with a visualization-oriented user interface. This approach makes relationships that exist among the items in the system *visible*, enabling exciting and powerful new forms of information access, exploitation, and control. The product of over six years of information visualization research, Starlight is simultaneously a powerful information analysis tool and a platform for conducting advanced visualization research.

2. Starlight - First Glance

- Novel "visible information" system
- Advanced information model
- Visualization-oriented user interface
- State-of-the-art information graphics
- Sophisticated query tools
- Data/text mining functionality
- Integrated Geographic Information System (GIS)
- Extensible Markup Language (XML)-based
- Client-server software architecture
- Windows NT/2000 platform

Starlight represents the first attempt to marry a variety of different types of "conventional" (and novel) information visualization capabilities into a single, integrated, information system capable of supporting a wide range of analytical functions. Further, Starlight visualization tools employ a common XML-based information model capable of effectively capturing multiple types of relationships that may exist among information of disparate kinds. Together, these features enable the concurrent visual analysis of a wide variety of information types. The result is a system capable of both accelerating and improving comprehension of the contents of large, complex information collections.

3. Principal Benefits

- Information Integration
- Complexity Management
- Workflow Continuity
- Accelerated Interpretation
- Improved Understanding

4. Motivation: Why visualize information?

Consider an arbitrary set of "information objects," for example, a collection of Web pages or database records, or perhaps a group of related email messages. What makes such a collection useful and valuable? We argue that it is potentially valuable because it can be used to help solve problems and, further, that its value for problem solving lies in one or both of two places:

- Within individual items (i.e., taken in isolation)
- In the relationships among the items

Deriving value of the first sort is an *information retrieval* problem: strictly a matter of finding and examining the item or items that have a certain property. Deriving value of the second type is an *information analysis* problem. Human cognitive analysis is largely a matter of comparison: comparing various properties of items with one-another, and comparing such properties with prior knowledge. As the volume and complexity of information increases, however, human ability to make these kinds of comparisons mentally degrades rapidly. Visualization technologies can effectively reverse this trend by capturing relationships in a kind of external, graphical, "memory" where they can be more easily compared and evaluated.

Visualization is a potentially powerful tool for information analysis because it enables humans to make rapid, efficient, and effective comparisons.

Note: A good rule of thumb to use when evaluating visualization designs is to ask yourself two questions: 1) What information does this design let me compare?, and 2) How easy is it to make the comparison?









5. Making it Work: How is information visualized?

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

In practice, enabling "visual" analysis of information is a two-step process. First, relationships among information objects (as well as the information itself) must be explicitly captured in a computer-manipulable form. Once this is achieved, interactive graphical representations of the relationships can be generated for analytical purposes.

Relationships are captured in a digital construct generically referred to as an *information model*. A model intended to support information visualization should be comprehensive (i.e., support many different relationship types), flexible (in order to support many different information types), and, above all, human-oriented. By this, we mean that, ideally, the model will capture relationships in a form that mimics the way humans naturally relate information.

The Starlight Information Model is our attempt to effectively meet these criteria. The Starlight Model is comprehensive, capable of accommodating a wide variety of relationship types, including discrete property (i.e., field/value pair) co-occurrences, free-text similarity, temporal relationships, parent-child associations, network relationships, and spatial (e.g., geospatial) relationships. Because the model is designed to capture relationships among XML objects, it can flexibly accommodate the full range of information types expressible in XML (i.e., almost any type of digital information). Finally, the model is human-oriented, explicitly designed for capturing and manipulating the types of relationships humans need to understand in order to solve complex, multifaceted, real-world problems.

Relationship Type:	General Similarity	Explicit Reference	Field/Value Co-occurrence	Parent/Child	Spatial	Temporal
Model Type:	<i>Vector-space</i> 	<i>Network</i> 	<i>Multidimensional Index</i> 	<i>Hierarchical</i> 	<i>Spatial</i> 	<i>Ordinal Index</i> 
Examples:	<i>Reports, articles, DB records</i>	<i>References & citations, hyperlinks</i>	<i>DB records, document metadata</i>	<i>File paths, taxonomies, IP addresses</i>	<i>Geolocations, CAD models</i>	<i>Event descriptions</i>

Components of the Starlight Information Model

Once relationships have been explicitly captured, Starlight can generate graphical representations of various aspects of the model that enable the underlying relationships to be visually interpreted. Importantly, Starlight visualizations are interoperable, enabling viewers to interactively move among multiple representations of the same information in order to uncover correlations that may span multiple relationship types. For example, email messages can be related to one another in a number of different ways. There may be topological relationships among the senders and recipients. There may be conceptual similarities among the message contents, or temporal correlations among the messages. Different email messages may even mention different places that are, in fact, physically near one-another: a spatial correlation. We are working to develop an information model

PNNL-14910: LER Data Visualization Analysis Pilot Study
10/15/2004

capable of seamlessly accommodating all of these relationship types, and visualization tools to enable users to quickly understand the potentially complex interdependencies among them.

To illustrate the potential power of this approach, consider again an arbitrary collection of email messages. A Starlight user may choose to graphically depict such "email spaces" in any of a number of different ways, depending on the problem he or she is trying to solve at any given moment. An analyst may initially wish to view the collection as a network diagram in which the emails are portrayed as edges connecting nodes that represent senders and recipients. This method enables the viewer to identify important topological relationships among individuals based on "who sent what to whom." Once a particular subset of email had been identified based on its network topology, an analyst might switch to a "conceptual" representation of the same information that summarizes the concepts described in the items of interest. Following that, the user could switch the display to another alternate representation that spatially groups the items according to author or recipient. In this way, even extremely complex and multifaceted relationships that exist in the collection can be quickly and easily characterized and assimilated.