Explorer at Los Alamos: A Library for the Future

Martha Waters
Los Alamos National Laboratory
mwaters@lanl.gov

Jim McDonald
Los Alamos National Laboratory
jmcdonald@lanl.gov

Abstract

Since 1993, Los Alamos National Laboratory has been developing World Wide Web (WWW) applications to facilitate access to vast quantities of information critical to the successful operation of a nuclear weapons facility. Explorer is a web-based tool that integrates full-text search and retrieval technology, custom user interfaces, user-friendly navigation tools, extremely large document collections, and data collection and workflow applications.

Explorer's first major thrust was to enable quick access to regulatory and policy information used by Department of Energy facilities throughout the country. Today, Explorer users can easily search document collections containing millions of pages of information, scattered across Web sites around the country. Over fifteen large applications containing multiple collections are searchable through Explorer, and the subject areas range from DOE regulations to quality management-related resources to technology transfer opportunities. Explorer has succeeded because it provides quick and easy access to stored data across the Web; it saves time and reduces costs in comparison with traditional information distribution, access, and retrieval methods.

1.   Introduction

Los Alamos National Laboratory (LANL) is a major Department of Energy (DOE) research facility covering almost 50 square miles of northern New Mexico. Home of the United States' largest supercomputing facility, LANL is also one of the largest multi-disciplinary research institutions in the world. Its 10,000 employees, contractors, and visiting scientists perform research in a wide range of scientific disciplines including materials science, supercomputing, biotechnology, nuclear materials, and national security. The Laboratory was founded in 1943, and in its 50-year history scientists have generated millions of pages of research data. This data represents an incalculable wealth of information, but because most of the documents are stored on paper and microfilm (some boxed in underground vaults), location and retrieval of specific information is time-consuming and costly. Recent and currently produced research data is more accessible, but the amount of information that is available or required to perform research and support functions in accordance with regulations is voluminous.

The Explorer Project Team was created in 1993 to build applications that would provide Laboratory employees with easy, fast, electronic access to very large data collections. Until this time, computers had been used to store, access, and manipulate business information systems data such as accounting information, but using computers to access large document repositories was a new concept. With the explosive growth and use of the Internet and the World Wide Web (WWW) in the last three years, this concept of managing information assets is now broadly accepted.

All Laboratory workers and customers need to see at least some of the data in the Laboratory's repositories; many require access to all of it. Scientists require access to classified and unclassified technical documents. Government regulators need to evaluate practices and policies of the laboratory and ensure compliance with federal law. Internal environmental cleanup staff must be able to locate and retrieve old research project data. And scientists responsible for stewardship of the nuclear stockpile rely on the Laboratory's millions of design and configuration records.

The Project Team determined that the *core requirements* for satisfying the needs of Laboratory staff were common, universal access to all information sources; access to external sources of relevant information; timeliness and data integrity; and, authentication and security controls for regulating access to sensitive information. To support these requirements, the Project Team determined the an effective search and retrieval application would have to underlie the primary interface to the repositories and that collection and

conversion tools as well as document management and work flow applications would be critical to the implementation of the Laboratory's information management plan.

Explorer was developed as a World Wide Web application to provide an accessible mechanism for searching various information repositories at the Laboratory. Explorer integrates off-the-shelf full-text search and retrieval technology, custom user interfaces, user-friendly navigation tools, and extremely large document collections. The search and retrieval component of the Explorer application works by presenting a graphical user interface (GUI) containing navigation buttons and a menu of searchable collections. Over time, the Explorer interface was extended to cover access to relational databases and other information sources, as well as the systems with which data is supplied to these information sources.

Through Explorer, users can easily search document collections containing millions of pages of information. For example, a researcher can search the SciSearch® collection of 13 million citations by author name to find out what other papers cite that author in their references (papers citing the author would contain related research). In a few short seconds, a list by title, author, and publication appears. The researcher can browse the list, view abstracts of the cited articles, or view the bibliographies of the cited articles. Quick access to the latest research in a particular field saves hours of searching through journals—even electronic journal indexes— individually.

To date, over 20 large document collections or databases are searchable with Explorer, and the subject areas range from DOE regulations to quality management-related resources to technology transfer opportunities to weapons design records.

As an electronic tool for search and retrieving information, Explorer has succeeded because it provides quick and easy access to stored data from any computer networked to the WWW; it saves time and reduces costs in comparison with traditional information distribution, access, and retrieval methods. Because Explorer was so successful as a search and retrieval application, the Project Team proceeded to integrate applications that would satisfy the remaining requirements for collection and conversion tools, access to external information sources, and document management applications. The original design of Explorer was open enough to accommodate this extension of functionality with only minor changes to the user interface.

2. State of technology

When the project began in late 1993, the goal was to make information previously available only on paper available through electronic means, recognizing that information is an asset and its value increases with accessibility.

Explorer was made possible by the major technological advances that occurred in computer technology in the early 1990s—the emergence of electronic document management software, client-server technology, scanning technology, cross-platform software, readily accessible mass-storage capabilities, and WWW networking and browser software. This new technology meant that information could be easily exchanged between computers with different operating systems. It also provided the capability for employees at their desks to obtain masses of information material that could be searched quickly and effectively.

3. Issues and Solutions

Explorer development began when the Internet and the WWW were still in their infancy. Modems were slow. Mass storage tools were new. Networking was young. Computers could not communicate cross-platform. Displaying and transferring images was slow, and the results poor. Each image file contained only one page of a document. The most significant breakthroughs that allowed full implementation of Explorer were the development of applications programmer interfaces (APIs) to various commercial search engines, the proliferation of Web browser software, and the introduction of Adobe Acrobat portable document format (PDF). With these tools, Internet access was no longer limited by technical competency, query interfaces to information repositories could be customized and made configurable, and WYSIWYG ("what you see is what you get") renditions of documents could be made available across platforms (e.g., PC, Macintosh, Unix) from a single source file.

There were also problems of perception and attitude on the part of Laboratory users. Many users were convinced of the value of computers to access and manipulate business operations data, such as payroll systems, but did not see the value in using computers to make vast amounts of information available.

On a larger scale, Explorer was created during a time when DOE was embracing the "openness" of historical information to the public. For nearly half a century, DOE had closely guarded all its information with a tenacity deserving of information that protects US citizens. However, not all information generated within DOE deals with national security, but rather with research

areas that advance understanding on a number of commercial and health-related topics. Once these records were opened, the challenge became making this information accessible to as many people as possible.

4. Requirements Mapped to Components

4.1. Common, universal access to all information sources

For some time, Laboratory staff had been using electronic information sources (e.g., Dialog, Lexus, Nexus, ISI Science Citation Index) to aid research and administrative activities. The major drawback to these applications was their relative independence of each other and the wide variations in user interfaces associated with each application. Not only did users have to learn to navigate each of the various interfaces, but they were unable to conduct searches across multiple data sources with a single query; therefore, a usable search metaphor for one application would likely be incongruous and unusable in another.

Figure 1. Explorer Multi-collection Search Screen

Figure 1 is an example of an Explorer multi-collection search screen. Users simply select one or more collections to search, without regard to source location or type. As a result of extensive usability testing, we discovered that the majority of users – regardless of skill level – were most comfortable when allowed to remain completely ignorant of the location, design, layout, and structure of the various information sources to which they required query access. On the other hand, the test results indicated that users performed requisite searches most effectively when allowed some control over which collections would be searched.

4.2. Access to external sources of relevant information.

The value of any information source is directly dependent on the facility with which users can locate and retrieve the materials they require. Though sequential searches of various electronic repositories represented a vast improvement over traditional reviews of printed materials, and whole-Web searches (e.g., AltaVista, Hotbot, Yahoo) are useful for locating entire collections and repositories, neither provide an effective mechanism with which a user can select specific target sources for a single query.

Though more and more Web search applications now allow more focused searching of selected sites, few support so-called "push" technolgies (e.g., alerts, notifications, agents) and none that we have discovered support more than minimal customization at the individual user level. Unsolicited feedback and usability test results from Los Alamos and throughout the DOE indicate that users are particularly interested in relating data available from external web sites to specific internal information sources. Moreover, we have learned that the vehicle used to notify users of new or changed data should be the same regardless of data source (e.g., internal web, external web, intranet systems).

4.3. Timeliness and data integrity

Feeder applications to automate the flow of information from the producers or producing system to the publishing application or web site are critical to maintaining the timeliness of information. In most cases, a combination of batch processes and real-time data stream will suffice for read-only and query access publishing; however, direct query access through application gateways (usually at the API level) is often the best course, particularly for data stored in transaction systems.

Even sophisticated feeder applications and real-time query access are often insufficient to provide the timely flow of information required for mission-critical operations. The Explorer application contains advanced features – called "Alerts" – which allow users to save named queries associated with specific collections or groups of collections. At regular intervals specified by the
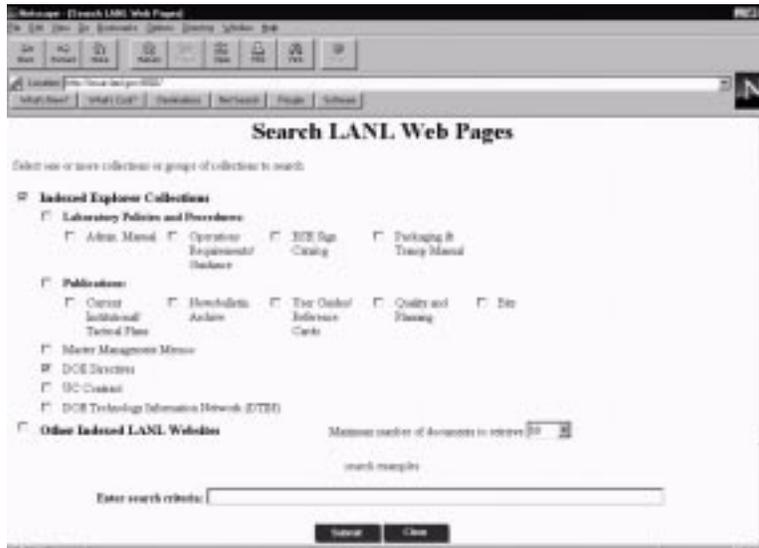
user (after any change, daily, weekly, monthly, and on demand), Explorer executes these queries and send the results via electronic mail to the user. Users may also request that new or changed documents matching the query be transmitted directly to their desktop. In addition to providing the necessary means to distribute information to the people who need it – and asked for it – in a timely manner, Explorer Alerts can also be configured to support controlled distribution protocols which require acknowledgment of receipt of specific documents. Figure 2 illustrates the web form used to create and save queries for an Explorer Alert for the SciSearch® at LANL application

For most purposes, data integrity can be divided into two major problems: the perception that what one sees online is actually the "official" version, and real server-side security to prevent unwanted posting of new or modified data. The former problem was largely solved by the introduction of the various image formats and freely available viewers; however, the Acrobat Portable Document Format (PDF) is largely responsible for ability we now have to distribute WYSIWYG *documents* from various media and various electronic source applications. The problem of server-side security to maintain the integrity of the information is mostly an extension of the security requirements for any networked computer, though Web server software has complicated the problem – as has the explosive growth of Web and Internet usage as a whole.

Figure 2 Sample Explorer Alert Screen

4.4. Authentication and security controls for regulating access to sensitive information

Authentication is rarely as simple as it presents itself, though making use of existing authentication systems (e.g., on the Los Alamos intranet) eliminated the usual angst associated with the yet-another-password problem. Validating users for access to applications and individual collections of data proved rather simple, with the help of such technological advances as the Enterprise Server from Netscape and secure socket layer encryption. On the other hand, validating users for access to specific documents or data elements has proved more difficult as there are intricate business rules associated with such accesses, rules that are difficult to implement and enforce without extensive intervention at the document level.

Providing approved access to validated individuals can be accomplished by coding documents for access categories and granting users access by category. In general, this will solve most need-to-know requirements, though it does little to reduce the overhead required to actually code the individual documents. Applying thesauri and simple data mining techniques can go a long way to identifying documents and associating them with potential categories; however, the technology is not so sound yet that we can apply it confidently without substantial human intervention.

5. The Explorer application

Explorer was developed to accommodate all of the requirements identified by users and stakeholders at the Laboratory and the Department of Energy. Rather than select several commercially available, off-the-shelf applications and attempt to seamlessly integrate them into a larger application with a single, configurable user interface (a Herculean task, at best), the project team set out to uncouple the user interface from the underlying applications. The resulting product is a highly-configurable application which employs gateways and APIs to integrate commercial and home-grown technology with a single consistent user interface.

5.1. Interface: the Explorer tool bar

The Explorer tool bar contains all of the navigation functions that we consider central to most search-and-retrieval and information access applications. To enable
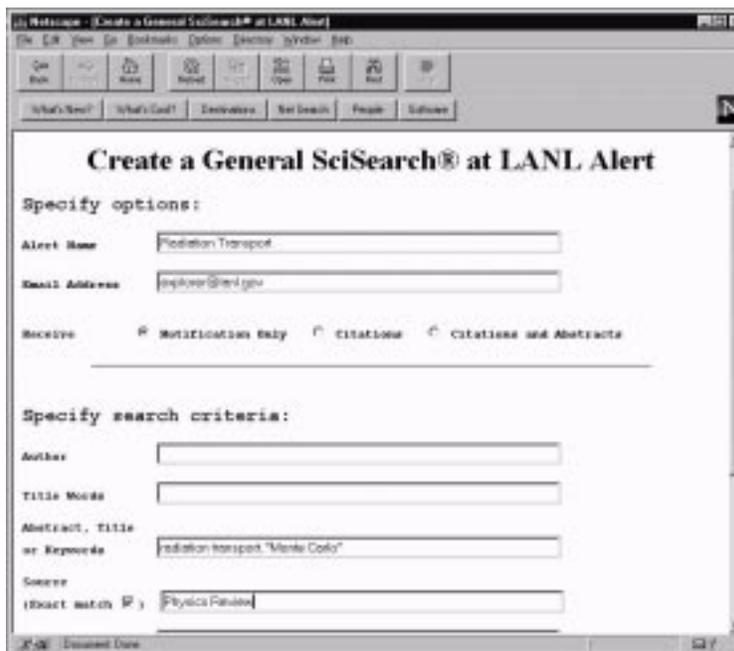
navigation capabilities independent of the Web browser software used throughout the Laboratory and DOE, we designed a set of "buttons" to be presented on every Explorer page, down to the document level. Through Explorer version 2.4, this toolbar was a single image map found at the top and bottom of the page. Following the user interface design principles for window-based applications, the buttons are either bright or dim, based on their availability in a given context.

Each of the items on the Explorer tool bar (Figure 3) corresponds either to navigating the site or collection, information about the application or the materials, or actions to be performed on individual documents.

From anywhere in an Explorer application, **Search** and **Home** always return the user to specific pages.

The **Up** button supports logical navigation through the hierarchy of the site. For example, if while viewing a document retrieved as the result of a search the user clicks on the Up button, Explorer will re-load the search results page from which the user selected the document.

The **Next** and **Previous** buttons are only active at the document level and allow the user to navigate through some or all of the items on a search results list without having to depend on the browser's Back button to return to the results list in order to select another document.

User feedback is critical both to the success of the Explorer application and to the information providers who supply source documents for Explorer collections. The **Comments** button points to a configurable form with which we collect identifying information and comments; when submitted, this form is then e-mailed to the Explorer technical support desk and to the owner of the information in the collection where the button was selected.

The **Help** button directs users to the Explorer online help, which contains information about the respective collections, search instructions and examples, and basic instructions for using Explorer.

Mark and **Download** allow users to mark records and documents as they browse the collections and, on demand, download the list of marked documents for specific actions (e.g., save to disk, mail to another user, request copies or information from the source).

The **About** button is reserved for use by the information provider. In general, the organization that provides the information for a specific collection may choose to re-label this button and route it to any web page or web site they choose.

5.2. Web crawler and indexer

The Explorer web crawler enables the project team to index information sources anywhere on the Web. Using the crawler, Explorer provides access to groups of Web-based information collections organized and identified by subject, interest, or source, and integrates these collections with existing Explorer collections to provide a single point of entry for searching. For example, *Success-through-Sharing* (StS) is an Explorer application that provides full-text search and retrieval across over twenty quality-related Web sites from around the world. Users need not know where all of these sites are located or even how to access them – they need only know about StS.

5.3. Alerts and notifications

Explorer alerts (Figure 2) provide users with the means to learn about changes and additions to collections relevant to their interests and responsibilities. Users create one or more profiles that Explorer uses to dynamically identify documents. When documents or data elements matching a user's profile are either changed or added to a collection, Explorer automatically notifies the user of their existence and, optionally, sends a copy of the document(s) directly to the user's desktop. This feature has been particularly well received by users who have neither the time nor the inclination to access the Web every day just to repetitively search for new information.

5.4. Online review and comment

The ability to review and comment on documents is an essential component of any "living" policy, procedure, or regulatory collection, and this particular feature has become an integral part of the feeder applications that supply documents and data to Explorer collections. With this feature, DOE and Laboratory users have the ability to review and comment on draft documents electronically, immediately after these documents are posted. Explorer collects, collates, and disseminates comments based o the

Figure 3. Explorer tool bar

business processes established by the information providers.

5.5. Knowledge sets

Knowledge sets are perhaps the most significant of all Explorer features. This is essentially a thesaurus that is tied directly to searches conducted in Explorer applications. Information providers and subject-matter experts create these knowledge sets to help users find the information they need in the various Explorer collections. When knowledge sets are invoked, users searching for a word of phrase will also retrieve records deemed by the subject-matter experts to be relevant to the original query.

6. The Collections

The unclassified Explorer collections are organized into specific categories: policy, procedures, and regulations; archives and research resources; administrative support; and, technology transfer and industrial partnership information. Organizing information in this way facilitates both casual browsing and focused searches. The following are brief descriptions of representative samples from each of the major categories.

6.1. DOE Directives

Explorer's first major thrust was to enable quick access to over 10,000 DOE Directives. It is critical that Laboratory personnel have access to the latest versions of the Directives, such as safety regulations. In the past, once a directive was approved at DOE headquarters, it took 10 weeks to publish 1700 copies and distribute them to 85 DOE sites. At each site, another 80 to 100 copies would be made and distributed to those employees who actually needed them to perform their jobs. Now, with Explorer, an approved directive is added to the DOE Directives collection and accessible to users at their desks in less than 10 hours. DOE saves thousands of dollars in printing and distribution costs, and the information is available almost immediately to those who need it. Because of the success of Explorer and the DOE Directives collection, DOE fully supports the extension of Explorer to other federal agencies and to most other DOE information repositories. The DOE Directives collection receives approximately 1,000 queries per day.

6.2. LANL Policies and Procedures

Administrative, health and safety, computer user manuals, and supply catalogs also have been put online at Los Alamos, searchable by Explorer. Employees can find and read or print the exact pages they need from their desks. Previously, it cost $100,000 per year to print and distribute just one these manuals to every employee. Hundreds of thousands of taxpayer dollars are saved every year with Explorer's application to document distribution.

6.3. Library Without Walls

The goal of the Library without Walls project is to provide an Internet interface for existing library systems and information repositories nationwide. At Los Alamos, we are converting to PDF files tens of thousands of unclassified reports (more than 500,000 pages), policies, procedures, and other official documents from microfiche, microfilm, paper, and various electronic formats.

With Explorer technology, the Library Without Walls project produced the Science Citation Index, a SciSearch® electronic database of science and technology. This database includes over 15 million citations, 100 million cited references entries, and over 41 million hyperlinks. Through Explorer, users can easily search document collections containing millions of pages of information. For example, a researcher can search the SciSearch® collection of citations by author name to find out what other papers cite the researcher in their references (papers citing the researcher would likely contain related research). In less than 5 seconds, a list by title, author, and publication appears. The researcher can browse the list, view abstracts of the cited articles, or view the bibliographies of the cited articles. Quick access to the latest research in a field saves hours of searching through journals – even electronic journal indexes – on an individual basis.

In the first 15 months that SciSearch® was operational at Los Alamos, it received 300,000 queries representing 1,000,000 accesses from 3,500 different users. A recent study found that over 5,000 queries per month come from the major non-government user of SciSearch® at this time, the University of New Mexico.

SciSearch® institutional users have subscribed to the service through the Laboratory because of Explorer. The SciSearch® data collection was available through Explorer for a free 6-month trial to members of the Library Services Alliance of New Mexico. When the free trial ended, all of the institutions (including Sandia National Laboratories and many universities) asked for subscription access through the Laboratory rather than acquire the data themselves – they preferred to use Explorer. The Project team receives a number of calls per week from other institutions that desire access to SciSearch® through the Laboratory because of Explorer's capabilities. Stanford University began

subscribing through the Laboratory in Spring 1997 and Lawrence Livermore National Laboratory signed on in the Fall of the same year.

6.4. Success-through-Sharing (StS)

Not only can Explorer search every page of every document in multiple collections, it can also search all the electronic files at a specific list of WWW sites for a word or phrase and return a ranked list and description of specific WWW addresses that contain the item. These addresses are presented as hotlinks, so the user can browse the list and simply click on an entry to access a file. The Success-through-Sharing (StS) collection is an example of such an Explorer application: the StS home page contains a menu of 27 other sites whose focus is quality management. Explorer can search all the files in each of these sites in a matter of seconds. The ranked list of findings helps the user to quickly identify useful sites for further exploration.

7. Measures of Success

The Project team, Laboratory users, and DOE consider Explorer a success and are actively encouraging its expansion to other information collections. Some indications of success include the following:

- While the number of accesses to Explorer collections has rapidly increased, the number of trouble calls has decreased. These trends indicate that users find Explorer useful and easy to navigate.
- The DOE Directives collection receives approximately 1,000 queries per day.
- DOE is encouraging the Project team to approach other federal government agencies, such as the Department of Defense, about employing Explorer with its Directives or other information repositories.
- DOE has asked the Project team to use Explorer on all Government Information Locator Service (GILS) sites. Every federal agency has such a site online, and Explorer will be implemented on the DOE GILS site in spring 1997.

A significant factor of Explorer's success is that its development occurred with constant feedback from users. No application is successful unless its users actually use it and are happy with it. All Explorer screens offer users an opportunity to provide comments or suggestions to the Project team.

The primary beneficiaries of the Explorer application are Laboratory and DOE employees and contractors, other government agencies that fund work at the Laboratory, American industries, and researchers at universities and other institutions. Though some collections are restricted to the Laboratory and DOE use, many are available to the general public. All of the information available at the DOE Explorer site (www.explorer.doe.gov) is open to the public. Some collections are accessible through the Laboratory home page (www.lanl.gov).

Since its inception, the project team has receive several awards recognizing the success of the Explorer application.

- The DOE Technology Information Network and the DOE Directives both received Energy Quality Awards.
- The LANL Environment, Safety, and Health Manuals on Explorer received a Scientech/NUS Pioneer Award and a Los Alamos Achievement Award.
- The Explorer project team was recognized by the Smithsonian Technology Innovation Awards program and added to the permanent collections of the Smithsonian Institution. Information about the project is currently on display in the Smithsonian American History Museum in Washington, D.C.
- The DOE Directives recently received the Hammer Award from the National Performance Review.

8. Conclusion

As with any software tool, the Project team is constantly improving Explorer. Future enhancements include further refinement of the "alert" service capabilities and improved methods for source providers to add information to a collection and collections to the system. Additions to the DOE Directives collection include cradle-to-grave processing—the capabilities to publish, comment upon, collect and collate comments, and address and resolve comments on draft directives, all online, culminating with posting the final official directive. Another enhancement will be refinement of the search technique so that explicit information can be found more easily. Another goal is to tie SciSearch® to electronic journals, so that users can access not only the abstract, but also the entire journal article that appears in the search results list.

Today, Explorer extends the wealth of information invested in the national laboratories to business, academia, and industry, so that users within and outside of the Laboratory and DOE can share in these time-saving, cost-saving benefits as well.