INSPIRE: A VIEW FROM THE OTHER SIDE

by Mary-Elise Haug, INCOLSA

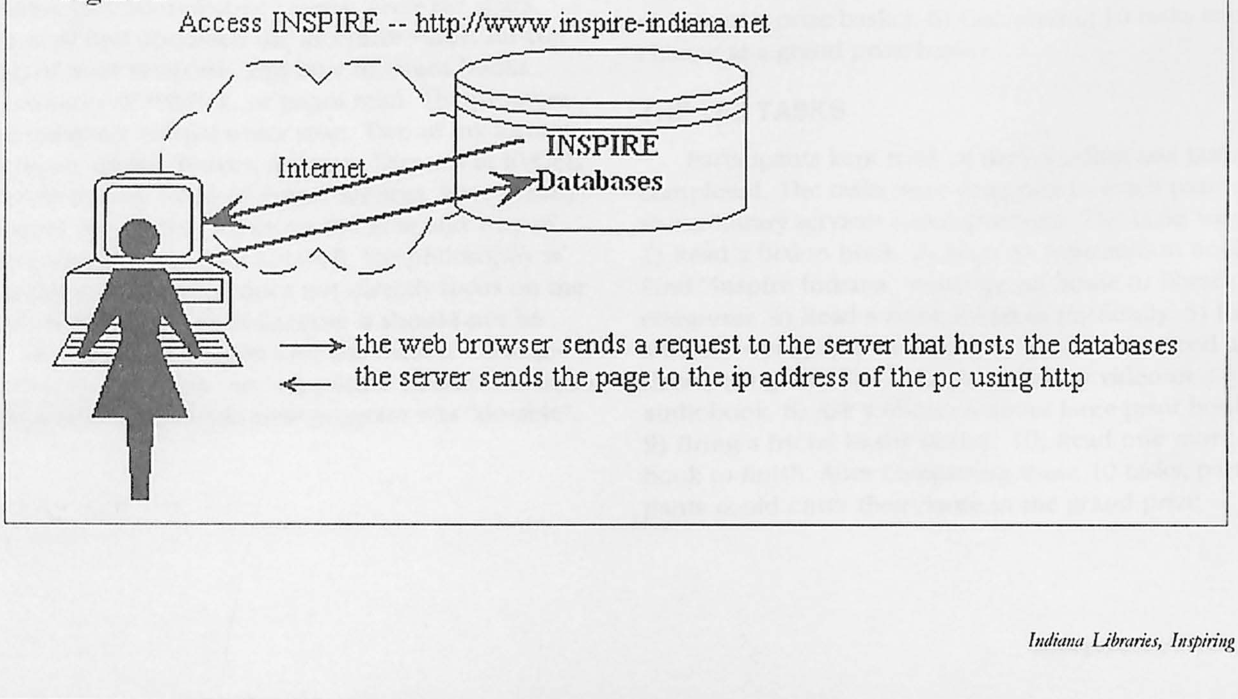
Inspire, Indiana's Virtual Library on the Internet, is a collection of commercial databases and other information resources that may be made accessible to any Indiana resident. Access requires a computer and an Internet connection. A virtual library is possible because Indiana has an existing network (IHETS/IndNET TCP/IP backbone) and there are standard protocols (http, z39.50) for communications as well as mechanisms for authentication, which in turn provide a client server infrastructure — a web browser connecting to a web server — that is available when developing a system. The greatest technological challenge in implementing Inspire has been to limit access to computers in Indiana. As Inspire staff have worked with individuals and institutions around the state a number of questions about our technical configuration have emerged. This article is an attempt to explain "how things work" inside the Inspire network. By looking at the components of the system: the client, the servers, the network, access control, and database interfaces, users may be better able to understand both the power and limitations of the system. For security reasons certain details will be omitted. Also keep in mind that technology evolves rapidly. Components of the Inspire system are regularly upgraded to add features, improve performance, and streamline processes.

THE CLIENT

I would imagine that most people reading this article have at least tried to access Inspire, Indiana's virtual library on the Internet. With an Internet connected PC and web browser, anyone in Indiana is a few clicks … and a little typing … away from retrieving full text journal articles.

Figure 1 depicts a seemingly simple web transaction, where a user sends a request with appropriate Internet protocols (http[1]) and receives a reply in (http) from an Inspire server. Inspire functionality is based on the premise that the PC is configured with a client, the web browser. Inspire development work is not done on the client side; however, the capabilities of later browsers are utilized by server applications. We recommend Netscape 4.x or Internet Explorer 4.x with a screen resolution of 600x800 and 256 colors and do not guarantee that database interfaces will work with earlier browsers. (A no frames interface designed to work with lynx, a text browser, and meet ADA guidelines is in development.) Users have been known to experience difficulties if Java/JavaScript is not enabled.

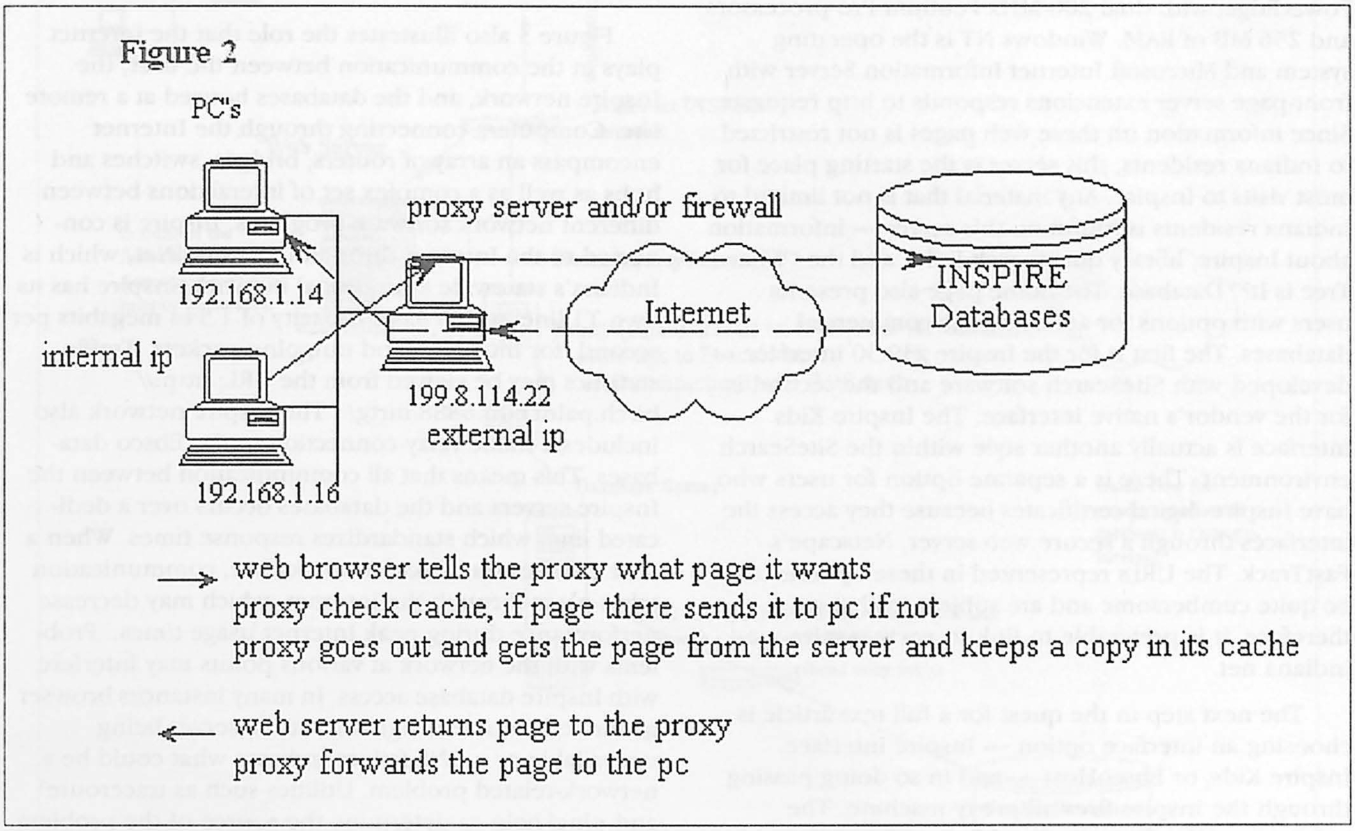
Figure 1


Access INSPIRE -- http://www.inspire-indiana.net
Internet
INSPIRE
Databases
→ the web browser sends a request to the server that hosts the databases
← the server sends the page to the ip address of the pc using http

THE CLIENT GOES THROUGH A PROXY OR FIREWALL

So you opened the URL http://www.inspire-indiana.net, selected Search Inspire NOW!, things did not go as expected — and you did not get the Inspire access denied page either. Such a scenario suggests that the PC in question may not access the Internet directly. It is becoming fairly common for schools, businesses, and other organizations to deploy firewalls[2] and/or proxy servers[3] between their internal networks and the Internet. Inserting a firewall or proxy server into the http transaction yields figure 2.

Figure 2



web browser tells the proxy what page it wants
proxy check cache, if page there sends it to pc if not
proxy goes out and gets the page from the server and keeps a copy in its cache

web server returns page to the proxy
proxy forwards the page to the pc

The configuration of a firewall or proxy server can affect a client's ability to access Inspire. This often manifests into one of three problems.

1. Traffic from a particular port is reject by the proxy/firewall. Due to the complexity of the Inspire servers and software, we have had to use non-standard ports[1] for the SiteSearch. Getting to a full text document requires the ability to receive http traffic from port 8008. *A change to port 80 is in the works.*

If the configuration is real strict, you may have to ensure that access to the inspire-indiana.net domain is allowed. Access to the database vendor's domain may also be required.

2. A school corporation (or other entity) uses a proxy service from an out of state vendor. The IP address seen by Inspire servers is no longer only used in Indiana. *An alternate means of authentication, digital certificates, is required.*

3. A user receives incorrect search results or cannot start a new session. In the case the proxy server is caching dynamic pages and serving them back to PCs instead of sending the request to the Inspire server. If at all possible, the proxy should be configured so that it does not cache URLs with *sessionid*. *The newest release of SiteSearch uses URLs that conform to a new standard, RFC-2396, which should help with caching problems. Staff is also investigating server settings and adding a page count to each URL to further reduce caching problems*

Adjusting browser settings so the document in cache is compared to the document on the network every time will also help with caching problems.

With the variety of proxy servers and firewalls available with a multitude of configurations, it is impossible to address every situation. In configuring Inspire servers and software, consideration is given to common firewall/proxy arrangements and every effort is made to conform to existing and emerging standards.

THE SERVERS

Figure 3 shows an expanded view of the http request depicted in figures 1 and 2. As illustrated, Inspire uses three servers in its current deployment. When a user opens the URL http://www.inspire-indiana.net, the primary World Wide Web server sends back the Inspire home page. This server is a Dell PowerEdge, with dual 200 MHz Pentium Pro processors and 256 MB of RAM. Windows NT is the operating system and Microsoft Internet Information Server with front-page server extensions responds to http requests. Since information on these web pages is not restricted to Indiana residents, this server is the starting place for most visits to Inspire. Any material that is not limited to Indiana residents is found on this server — information about Inspire, library quality web links, and the "What Tree is It?" Database. The home page also presents users with options for accessing the commercial databases. The first is for the Inspire z39.50 interface — developed with SiteSearch software and the second is for the vendor's native interface. The Inspire Kids interface is actually another style within the SiteSearch environment. There is a separate option for users who have Inspire digital certificates because they access the interfaces through a secure web server, Netscape's FastTrack. The URLs represented in these options may be quite cumbersome and are subject to change; therefore, it is preferable to link to www.inspire-indiana.net.

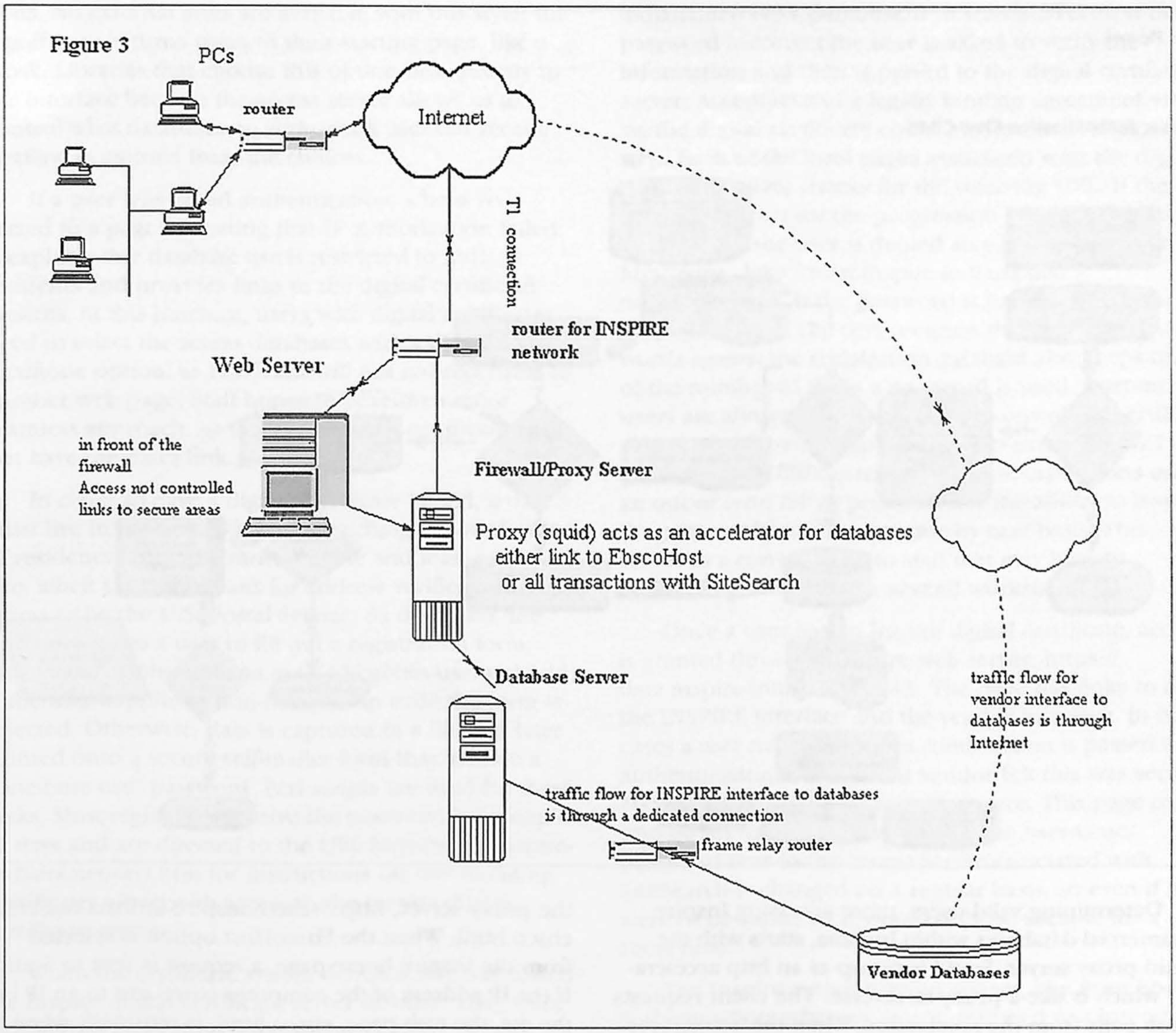
The next step in the quest for a full text article is choosing an interface option — Inspire interface, Inspire Kids, or EbscoHost — and in so doing passing through the Inspire firewall/proxy machine. The existence of the firewall, CheckPoint's firewall 1 product, is invisible to the user. All incoming and outgoing traffic to data is routed through the firewall, which listens on the external interface of a Sun Enterprise 250, Search, that has one 250 MHz processor and 512 Mb of RAM. All the Inspire Sun servers run under the Solaris operating system, which is a flavor of UNIX. The firewall is configured to optimize performance, so its impact on response time is negligible. Users are limited to selected ports and protocols on this machine and the machine behind it, increasing the security of the Inspire network and limiting down time that results from malfeasance. The downside is that if the firewall crashes both servers are inaccessible to network traffic as is access to the commercial databases. This machine also hosts a proxy server, Squid, which plays a substantial role in user authentication, to be discussed below.

The machine behind the firewall is a SUN Enterprise 450, Data, with 4 250 MHz processors and 768 MB of RAM. Data has significant storage capacity and is outfitted to accommodate local databases, if any are acquired. Thus far databases available through Inspire have been housed remotely (on a computer at the vendor's site). The Inspire interface, which has been developed using OCLC's SiteSearch product, operates on this machine. One of the SiteSearch components communicates with the databases to retrieve the full-text article and return it to the user.

NETWORK

Figure 3 also illustrates the role that the Internet plays in the communication between the user, the Inspire network, and the databases housed at a remote site. Computers connecting through the Internet encompass an array of routers, bridges, switches and hubs as well as a complex set of interactions between different network software programs. Inspire is connected to the Internet through IHETS/IndNet, which is Indiana's statewide educational network. Inspire has its own T1 line, which has a capacity of 1.544 megabits per second, for incoming and outgoing packets. Traffic statistics may be viewed from the URL: http://birch.palni.edu:8888/mrtg/ . The Inspire network also includes a frame relay connection to the Ebsco databases. This means that all communication between the Inspire servers and the databases occurs over a dedicated line, which standardizes response times. When a user chooses the EbscoHost interface, communication takes place through the Internet, which may decrease performance during peak Internet usage times. Problems with the network at various points may interfere with Inspire database access. In many instances browser generated error messages about the server being unavailable or a DNS failure indicate what could be a network-related problem. Utilities such as traceroute[5] and ping[6] help to determine the source of the problem. Many firewalls, including Inspire's, reject the IMCP packets sent by these tools. For troubleshooting purposes, trying another URL such as http://www.incolsa.net, which uses a different router and T-1 line, or http://www.ihets.org, which is nearby, provides useful information. If the browser generates error messages for these sites as well there is bound to be a network problem. When a network interruption occurs, it is beyond the control of Inspire staff. We do keep in contact with IndNET staff until the situation is resolved.

Figure 3


PCs
router
Internet
T1 connection
router for INSPIRE network
Web Server
in front of the firewall
Access not controlled
links to secure areas
Firewall/Proxy Server
Proxy (squid) acts as an accelerator for databases
either link to EbscoHost
or all transactions with SiteSearch
Database Server
traffic flow for vendor interface to databases is through Internet
traffic flow for INSPIRE interface to databases is through a dedicated connection
frame relay router
Vendor Databases

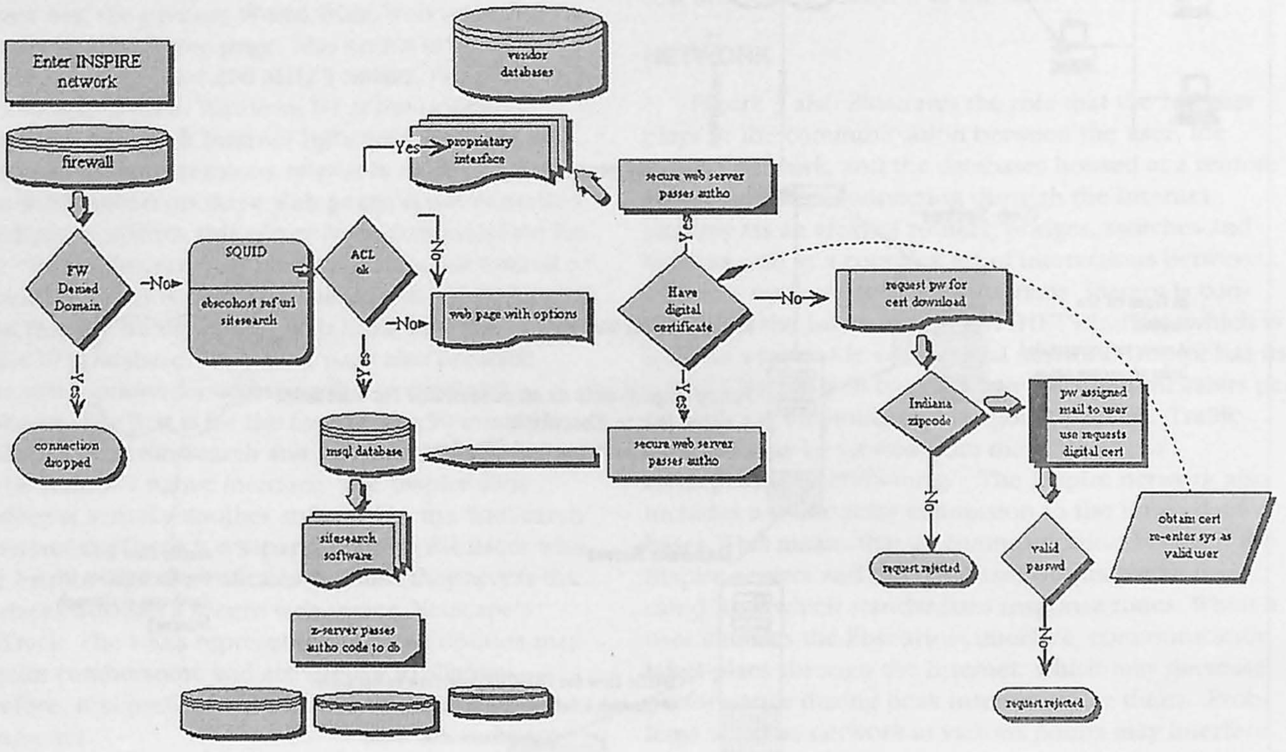
ACCESS CONTROL

The next step in retrieval of a full-text article is getting past Inspire's access control. What distinguishes Inspire from other statewide information access projects is the commitment to provide access to a diverse constituency, including homes and businesses, in Indiana. Traditionally, statewide library projects have authenticated users via IP addresses/domain name or library barcode, often leaving the implementation to participating institutions. There were no models available for extending access to any state resident without requiring a person to have a library card. When evaluating existing authentication schemes, it became apparent that the available technologies were premised on an identifiable user group, usually employees or students. Business enterprises, concerned about secure transactions for commerce, would register folks, regardless of geography. The more realistic solutions, notably username/password scheme (htaccess or database), cookies, and proxy servers, failed to address how to identify that someone was from Indiana in the first place. It became clear that there are two issues at work for establishing an authentication system for statewide access. The first one is how to verify that a user is a resident of a particular state and the second one is how to maintain information about valid system users once residency is verified.

The full-blown details of the user authentication system are outlined below. The short version is that Inspire staff maintain a list of IP addresses that are used only in Indiana. If the IP address of a computer matches one on this list, the user is passed on to the selected database interface. If an IP address is not on this list, but an Inspire digital certificate has been installed in the browser, access to the interfaces may be gained through the secure web server (https://data.inspire-indiana.net:443). If neither scenario applies, an Indiana resident may fill out a registration form and receive a password in the mail. That password is used to download a digital certificate. Figure 4 illustrates this process.

Figure 4

User Authentication Flow Chart



Determining valid users, those accessing Inspire commercial databases within Indiana, starts with the Squid proxy server. Squid is set up as an http accelerator, which is like a proxy in reverse. The client requests a web page from the accelerator. When the accelerator receives the request it gets the page from either its cache or the web server and then returns the page to the browser. Squid will only get the database interface pages for IP addresses and domains listed in specified access control lists (acl). The majority of libraries, schools, universities, and other Indiana institutions are authenticated in this manner. Local ISPs, who use specific IP addresses or domains in Indiana, are also included within the list.

When the proxy server communicates with the web server, requests appear to come from Inspire's internal network. Access to the database interface web server is limited to the internal network. For additional protection, the firewall software only lets machines behind it use the internal network address. The result is that a user's web browser cannot gain direct access to the web server associated with SiteSearch or the protected link to EbscoHost.

EbscoHost checks for a referring URL for its authentication. That referring URL is a link on page behind the proxy server, http://search.inspire-indiana.net/links/ebsco.html. When the EbscoHost option is selected from the Inspire home page, a request is sent to Squid. If the IP address of the computer correlates to an IP in the acl, the web page, ebsco.html, is returned; when the user clicks on the link to EbscoHost, the correct referring URL is in the browser properties. If a user types the URL for EbscoHost, http://www.epnet.com/cgi-bin/refurl30?incolsa.main.web, directly, an error message will be returned. If libraries want a direct link to EbscoHost on their web pages they need to link to the Inspire page or make arrangements with Ebsco for an alternative referring URL and logon identity.

The SiteSearch interface has a built in access server that uses a msql database. The authorization table accepts connections from Squid, as there is an entry for the internal network's IP address. If a user attempts to access port 8000 on the Enterprise 450 directly, the connection is rejected because the appropriate IP or domain is not listed in the authorization table. The system is set up to prompt for a user name and password. Incorrect values are embedded in the form. In some instances the Access Server authorizes valid users. For instance, Inspire staff support a version of the interface designed for libraries that lock down workstations. No external links are available with this style; the logoff page returns users to their starting page, like a kiosk. Libraries that choose this option link directly to the interface because the access server allows us to control what databases an authorized user can access; NetFirst is omitted from the choices.

If a user fails Squid authentication, s/he is redirected to a page indicating that IP authorization failed. It explains that database use is restricted to Indiana residents and provides links to the digital certificate services. At this juncture, users with digital certificates need to select the access databases with a digital certificate option, as FastTrack will not redirect users to another web page. Staff hopes to develop a more seamless approach, so those users with certificates will not have the extra link.

In order to have a digital certificate issued, a user must live in Indiana. When approaching the verification of residency issue, the most reliable and least expensive, albeit slowest, means for address verification seems to be the U.S. Postal Service. As deployed, the system requires a user to fill out a registration form, http://worf.inspire-indiana.net:443/cgi-bin/usera.pl/Add. If the user supplies a non-Indiana zip code the form is rejected. Otherwise, data is captured in a file and later printed onto a secure self-mailer form that include a "one-time use" password. Perl scripts are used for these tasks. Most registrants receive the password in a couple of days and are directed to the URL http://www.inspire-indiana.net/cert.htm for instructions on downloading certificates along with access to the Inspire digital certificate server.

When the certificate system debuted, users accessed the Inspire digital certificate server directly. They were required to enter their password with other data for the digital certificate. An auto-verification program associated with the server checked the password against the registration database. Passwords expired after 21 days. Most people abided by the one-time use concept; staff noticed that many users required several attempts before successfully downloading a certificate. A few seemed to abuse the lack of enforcement of the one-time use and send passwords to friends out of state (or were out of state college students with relatives in Indiana). The firewall became part of the authentication process, dropping such out of state users. It is possible to revoke certificates as well.

To compensate for suspect use and at the same time accommodate users who had difficulty downloading both the server and client certificate in one visit, a new mechanism for password checking was devised, forms were streamlined, and instructions were updated. Now, the process for accessing the digital certificate server begins with checking the mailed password at the URL: http://worf.inspire-indiana.net:443/cgi-bin/usera.pl/SearchIDForm. If the password is correct the user is asked to verify the information and then is passed to the digital certificate server. Acceptance of a legally binding agreement vis-à-vis the digital certificate could incorporated in this step. Each of the html pages associated with the digital certificate server checks for the referring URL. If the URL is incorrect for the progression of pages that issue a certificate, the user is denied access and sent to an html page, http://worf.inspire-indiana.net/nopasswd.html. If the password is incorrect an error message is sent. The perl program that checks passwords against the registration database also keeps track of the number of times a password is used. Currently users are allowed three attempts to download certificates and the passwords continue to expire after 21 days. We have the capacity to provide institutions using an out of state ISP or proxy service the ability to by-pass the password process on a case by case basis. This serves as a convenience to staff that may have to download certificates for several workstations.

Once a user has an Inspire digital certificate, access is granted through a secure web server, https://data.inspire-indiana.net:443. The page has links to both the INSPIRE interface and the vendor interface. In both cases a user name/password combination is passed for authentication. The current vendor felt this was secure enough because the server was secure. This page could also be set up as a referring URL. The username/password sent to the Access Server associated with SiteSearch is changed on a regular basis, so even if it were to be captured it would only be useable for a short time.

The Inspire authentication system, like most library authentication schemes, is not designed nor intended to be 100% secure. We do believe it is on a par with methods, such as referring URL, used by the vendors we work with. It does keep most people honest. Reasonable efforts are made to improve the system as any weaknesses are exposed as well as take advantage of advances in technology.

THE INTERFACE SOFTWARE

After being properly authenticated, a user is closer to the full-text article and now only has to perform a search/receive results using the database interface software. The interface software acts as a gateway between http requests and the databases stored in their raw format. EbscoHost is the native interface to EBSCO databases. Ebsco maintains EbscoHost almost entirely with modest adjustments, such as the home library link, possible through an administrative module. Both the Inspire Interface and the Inspire Kids Interface were developed using OCLC's SiteSearch Suite.

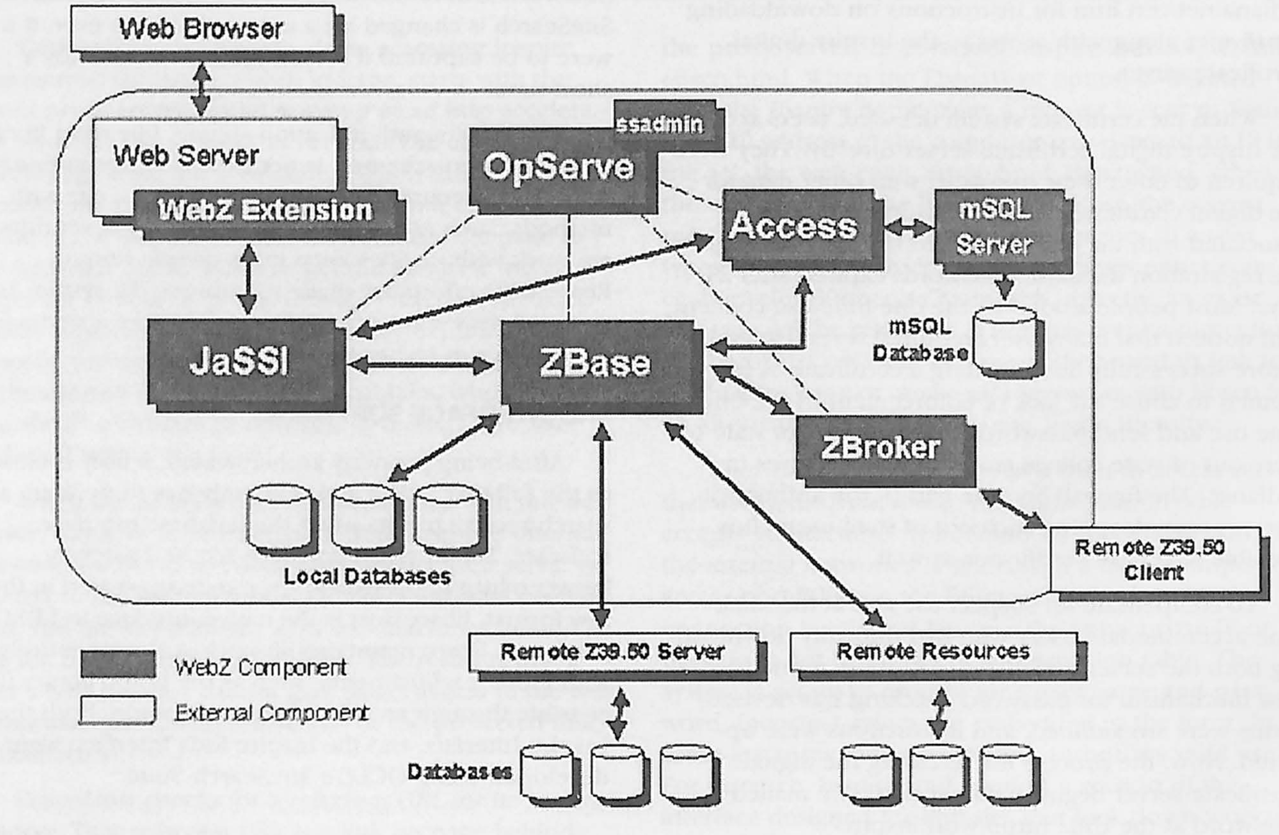
SiteSearch is written almost entirely in the Java programming language and is highly customizable. By employing the z39.50[7] standard, it allows users to search multiple databases from different vendors at the same time.

The SiteSearch software and interfaces are enhanced and administered by Inspire staff. Because searching is built on the z39.50, SiteSearch interfaces have some limitation when compared with vendor native interfaces; for example, browsing and many of the special search limits are not available through Ebsco's z39.50 server — the remote server in Figure 5[8]. Figure 5 illustrates the components of SiteSearch. The web browser talks to the web server through the Squid proxy server as explained earlier. In order to get to a full text article, a series of http requests are sent to an Apache web server, with SiteSearch's WebZ Extension. The requests are forwarded to JaSSI, which translates between http and Z39.50. Before any translation, JaSSI checks the access database. If all is clear a session ID number is assigned. This is how WebZ maintains a persistent connection with a web browser such that the server knows where to return search results. The JaSSI goes through Zbase (z39.50 server) to process queries and result sets.

The web pages that go with the SiteSearch interfaces are a combination of hard coded html and html generated from Java classes. Database results, returned through the Zbase, are formatted by an additional Java code based on database attributes. The differences between the Inspire interface and the kids interface demonstrate the flexibility in design and layout. Anything presented in html, such as search screens and results screens, can be significantly modified. Java classes that interact with Zbase and the remote server, such as parsing queries, are more cumbersome to alter.

From time to time z39.50 error messages appear in the Inspire interfaces. In many instances, after logging out of the current session and starting a new session the problem simply goes away. Such glitches may be a result of heavy traffic or packet loss or other communication problems that can occur anywhere in the complex array of interactions between the web browser and the remote databases. Errors indicating problems with results often fall in this category and are isolated rather than symptomatic of wider system problems. Error code 25 — "Z39.50 Search Error - 109/Database Unavailable" — happens when Ebsco takes a database off-line, making it unavailable through their z39.50

Figure 5



server. Sometimes the database is available in EbscoHost and other times it is out of commission for both interfaces. This error is outside the realm of the INSPIRE network. Additional errors are seen at times when either SiteSearch or the databases have reached their maximum user limit.

CONCLUSION

After looking at how a client and the Inspire servers interact through layers of networking, access control, and database interface software, it should be clear that the seemingly simple request for a full-text articles has considerable complexity to it. By understanding the role different components — clients, servers, networks, authentication, and database interfaces — play in Inspire database access, librarians should be better able to communicate with technical staff about Inspire access as well as to assist users with Inspire related questions or problems.

ENDNOTES

[1] HTTP stands for hypertext transfer protocol, a set of standards for transferring files from computer to computer across the Internet.

[2] A firewall is a combination of hardware and software buffer that organizations use between internal networks and the Internet. A firewall allows only specific kinds of messages (protocols) to flow to and from the internal network and the Internet.

[3] A proxy server acts as a gateway between an individual computer and the Internet. It speeds up loading web pages, while reducing bandwidth requirement of the Internet Service Provider. Some proxies are configured to prevent traffic from certain web sites based on a variety of criteria.

[4] The port refers to the number part of a URL. It is to the right of the colon, i.e., http://search.inspire-indiana.net:8008. Every service on the Internet listens to a particular port. Web servers normally listen to port 80.

[5] Traceroute (tracert) utilities show the route taken from a PC to a particular Internet connected machine. Problems getting to the destination address suggest network difficulties.

[6] Ping is used to tell if there is some rudimentary connection between a pc and an Internet connected device.

[7] Z39.50 is the Information Retrieval Service Definition and Protocol Specification for Library Applications. This standard, used by WAIS, specifies an OSI application layer service to allow an application on one computer to query a database on another. Z39.50 is used in libraries and for searching some databases on the Internet.

[8] Hagler, Mike. "WebZ System Diagram". Available http://cypress.dev.oclc.org:7301/help/sa/sa_04-10-01r.html (5 Oct 1999).