APPENDIX H

DISCUSSION OF LSN DESIGN ALTERNATIVES

1   Background

The LSN is the combined system that makes documentary material available to parties, potential parties, and interested governmental participants to the proceeding for a license to receive and possess high-level radioactive waste at a geologic repository operations area pursuant to part 60 [of 10 CFR], as part of the electronic docket or electronic access to documentary material, beginning in the pre-license application phase. The specific method of providing electronic access to documentary material is not specified in the rule in order to allow flexibility to accommodate current and future technology advances.[1]

It is anticipated that the functionality of NRC's ADAMS system can be readily adapted to ensure that NRC high-level waste documents are LSN-available, to establish the electronic docket, and to facilitate electronic motions practice via Electronic Information Exchange (EIE).

2   Researched Technologies

2.1   Search and Retrieval

ASLBP staff attended a Delphi Consulting Group seminar on portals and the business is dominated by perhaps only 4-6 companies with current, deployed, and competitive software. The consensus market leader in this technology is Plumtree™.

> ". . . most corporate portal vendors we've talked to mention that one of their primary competitors is Plumtree. This means that Plumtree right now is the company to beat in this space. Vendors who offer similar types of content management capabilities are startups Glyphica, KnowledgeTrack, and 2Bridge. . .Viador. . . Sequoia and DataChannel. . . SAP, PeopleSoft, Lawson. . .Netscape and Yahoo!. . ."[2]

> ". . . Pointcast. . .Dataware. . .OpenText Livelink. . .Viador. . .Verity's Search97 / Agent Server / Knowledge Organizer. . ."[3]

The Plumtree Corporate Portal has applications in sales and field support, government, competitive intelligence, customer service, and product research & development. Plumtree's customers include Texas Instruments, U.S. Department of Energy, Compaq Computer, Automated Data Processing, Monsanto Consumer & Nutrition, W.W. Grainger, and Caterpillar.

---

[1] [63 FR 71735, SUPPLEMENTARY INFORMATION]

[2] Patricia Seybold Group. "Plumtree Blossoms: New Version Fulfills Enterprise Portal Requirements", in Information Assets: Transforming Information into Profits, June 23, 1999. P.9.

[3] Molly Lyman of Project Performance Corporation at the Delphi Seminar on Portal Software.

Other products such as Excalibur™ could be utilized to implement portal sites, and will be closely examined in the design phase of the project. At the October 1999 TWG meetings, Dr. Tom Nartker of UNLV's Information Science Research Institute made note of the capabilities of some of the software products they have been evaluating for DOE/YMP, with special emphasis on the Excalibur software package. Excalibur provides a capability to establish a uniform software base across multiple sites and it then handles the process of running a distributed query against each site in the enterprise network. It was noted, however, that this would require that all participants license the same software - which would require the LSNA to issue mandates for use well beyond what the LSNA is currently authorized to propose or request. Additionally, this would require all participants to purchase, install, and populate within about a 20 month window and this was deemed non-viable.

We then met with DOE's ES&H organization about their experiences with developing a portal site and found that little customization was needed to develop the portal presence and that the project took approximately seven months from start of design until operational deployment. It should be noted that the DOE document collections are all found in a single "corporate organization", are individually smaller than the projected DOE and NRC contributions to the LSN, but conversely, connect to more locations and on as wide a range of platforms and operating systems as is anticipated for the LSN.

Certain attributes of Plumtree's portal software would be especially useful. It has its own underlying SQL database and full text indexes built from data extracted from target sites. The software has an additional feature of building in a "data dictionary" that keeps track of the different field naming conventions encountered in each target collection. This approach allows the portal site to present a single user interface to do search and retrievals. It also allows the participants' sites to act as a backup should the portal site become inoperative (by a user going directly to the participant's homepage), and, allows the portal to continue to identify (but not retrieve) the existence of a document even if the participant's site is temporarily inaccessible. It also has some utility as a central caching and replication mechanism for documents routinely retrieved from the target collections.

The use of portal software was vetted in great detail by the LSNARP TWG. Each participant's web-accessible (outside the firewall) collection may use any number of software management systems for structured data (bibliographic) and unstructured data (text and images) under the operational control of the participant. The LSN portal software scans through these collections and builds its own index to structured data or its own index to text terms found at one of the target sites. Options within this Alternative include making decisions about the level to which the system is developed. With more memory, the system can cache the most frequently used files (text or image) right on the portal machine in order to speed response time, but this increases the amount of memory that needs to be acquired and integrated into the LSN. The portal could be used to store other media types, such as full motion video or audio files.

Software for the LSN solution may include software development tools, web authoring tools, a universal interface module, search engine and indexing software, utilities and additional enabling features such as help software. Software for the audit data capture and analysis resource may include various Internet and database auditing software and development tools, a database package, a report generator package, and various other utilities and analytical tools.

CD-ROM authoring hardware and software may be required for mass dissemination of training tools for the technical staff of the parties and potential parties.

2.2   Audit

Per the LSN Rule, the LSNA is required to verify the integrity of the data.   To do this on the geographically dispersed sites, requires that we crawl the documents on a regular basis and bring back and store at least the following information: file name, file location, date of file, and finally file fingerprint. With this information alone, the LSNA can identify that data has changed but is not able to determine what data has changed. Without knowing what data has changed, the LSNA has no way of identifying the identical version of the document, previously placed on the participant server, which is to be replaced by the participant.

In the case of image objects, the fingerprint would be very difficult to match if changed data were present. However, in the case of structured data records and text files, embedded but non-viewable-character bytes could be more easily used to compensate for deleted data so that the files appeared to be comparable. This means that true fidelity to content, especially of textual documents, must be accomplished over and above merely collecting file metadata. To accomplish this, content must be characterized at the word level, and this requires the construct of an accounting mechanism for term occurrence within each document, e.g., building an index against which to compare.

Below is a summary of our research into the architecture needed to support the LSN audit requirement under LSN design alternative one and the conclusions we reached based on that research.

A close examination was performed on the class of resources required in order to determine whether savings could be achieved in the hardware and software configuration. It was found that the cost of computation has fallen dramatically, in fact, to the point where even the least capable CPU available in the market is now capable of what were once considered daunting computational tasks. Second, modern software techniques for indexing require far less computational power than earlier technology, many advances have been made in this area. Third, it is expected that participant sites will be crawled relatively infrequently. In summary, indexing was not a significant factor in determining the CPU requirements.

A close examination was performed on the implications of design solutions that anticipate significant data storage. It was found that, especially in Alternative 1, the design presumes retaining the indexes, pointers to the documents themselves, and related information but discarding the source material after indexing. This requires far less storage than retaining the entire document collection. In any case, it would be necessary to retrieve the documents if for no other reason to "fingerprint" them for later identification. There exists no reliable way of having the participants perform this fingerprinting and it is doubtful that a requirement to do this could be imposed upon them if it did exist.

A close examination was performed on whether audit solutions were influenced by COTS vs custom code approaches. It was found that lessened development requirements are reflected in lowered staffing costs. In all design alternatives it was assumed that development would be

expedited through the use of highly-effective off-the-shelf software for site development bundled with the portal software. If the portal is not used to expedite development of the audit capability in Alternative 1, it was felt that the staffing costs will be far greater. Security and hardware and software maintenance will be required no matter what design is adopted.

In focusing just on the audit capability, emphasis was placed on finding a substitute for the portal software that served as the focal point for the other design alternatives, specifically to find an alternative for "crawling" participant sites to enable implementation of the audit facility. This software is an expensive component and places significant requirements on supporting software and hardware that add greatly to the overall cost so finding an alternative for it would have a cascading affect on all parts of the system.

We spent several days researching crawler (AKA "web spider") technologies and found that two categories of products existed. We did not find a single product that fit between these categories, and recognize that further research during the design phase might prove fruitful.

First examined were bare-bones crawlers that focused on information retrieval with very limited information processing capabilities. Typically, information processing was limited to performing site replication and backup, or site-wide HTML code validation. We looked at about 30 crawlers, some open source, some commercial. We could not see a way to extend their functionality into a system that expedited determination of participant compliance without writing a significant amount of original code.

The second category examined were products that implemented portal facilities by combining one or more crawlers, comprehensive information processing facilities, comprehensive information management facilities, and comprehensive presentation facilities.
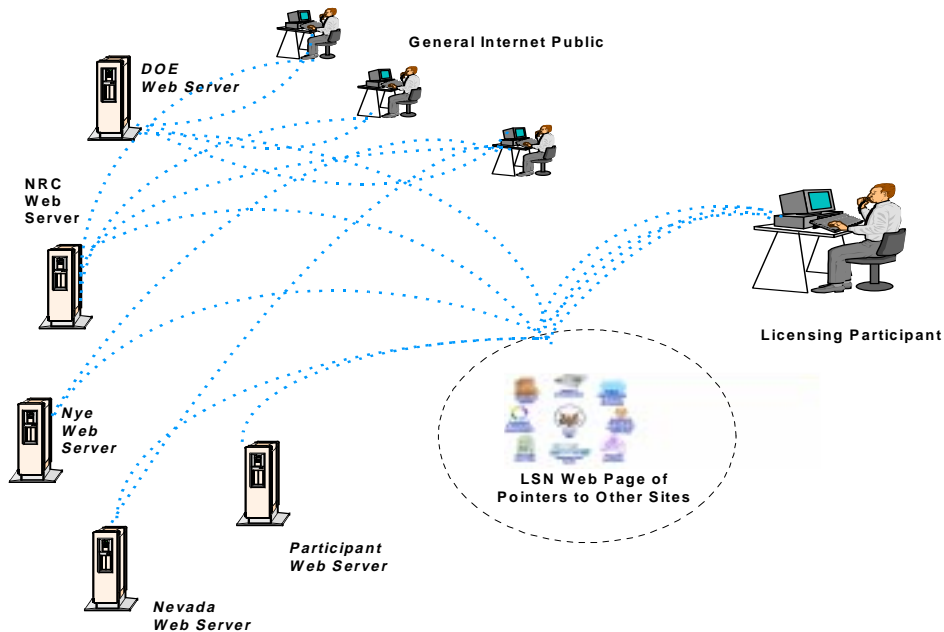
So the design alternative was to specify a bare bones product and develop additional needed functionality or to specify an off-the-shelf product that, perhaps, contained more functionality than needed. Due to the tight time constraints for development, the rigid statutory deadlines for implementation, and the high cost of in-house development we felt it was unacceptably risky, and likely ultimately more costly to do anything other than obtain as much functionality as possible off the shelf. However, we did not attempt to price out these development costs as that would have required much more time, and much more specificity in the system design than were available at the time.

2.3   Technology Investigations of the TWG

2.3.1   October TWG Meetings

The group generally discussed Alternative 1, which was characterized as a relatively non-complex development of a web page that provides links to the various sites.

Alternative 1 Schematic

General Internet Public

DOE Web Server

NRC Web Server

Nye Web Server

Participant Web Server

Nevada Web Server

Licensing Participant

LSN Web Page of Pointers to Other Sites

In this Alternative, questions were raised about the possible need for participants to enhance their sites by the addition of a navigational tool that would return the user back to the portal location in order to move to other collections for searching. Additionally, this approach provided no backup capability for the various participant sites other than what they would provide themselves. The TWG felt that this approach left too many of the perceived functional requirements not addressed. It was also noted that for the typical user, and especially for members of the general public, that the burden of having to learn perhaps 6 or 8 different search engines would become onerous. A final observation is that Alternative 1 gives indirect benefit to some participants, but not to others in that it can only be optimized by those who can afford to pay for intermediaries.

A general consensus was reached that Alternative 1 did not meet requirements for the following reasons:

2.   Too complex for users
3.   Too difficult to navigate
4.   Not possible to aggregate information
5.   User interface not consistent
6.   Not versatile
7.   Does not meet needs of large, complex discovery system
8.   Potentially excludes some participants and "tilts the playing field" for others.

Agreement was reached that Alternative 1 would not be recommended to the full LSNARP.

The group discussed Alternative 2, a medium complexity effort which was characterized as being similar to a central portal page where queries may be launched against individual participant sites, and where the result sets from the individual sites are subsequently merged back together for presentation to the user. It was noted that the distinguishing characteristic of Alternative 2's front end is a "meta search" capability. This is similar to multi-engine or multi-site searches such as are found at http://www.allonesearch.com/ or http://www.dogpile.com/custom/index.html.
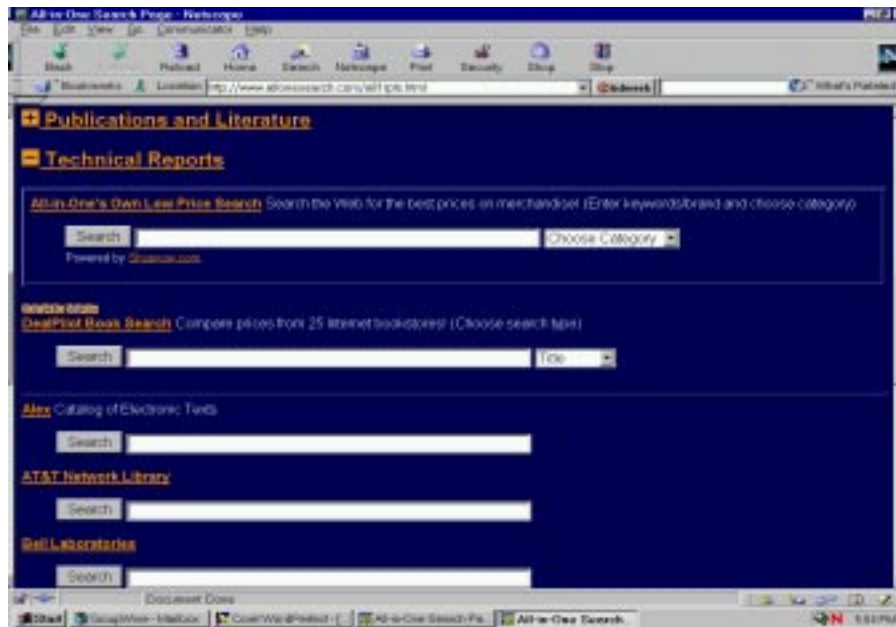
It was noted that Alternative 2 may have difficulty in maintaining any relevancy ranking as the portal site attempted to merge results sets back because each participant's site may use different software and rely on different methodologies to determine relevance. Not merging the result sets may lead to multiple partitions of returns, one for each participant site. It was also noted that having multiple underlying data files, some being structured headers while others were unstructured text searches, could result in a user having to launch separate searches against all headers, and then another against all text, and that customization may be needed to allow searchers the ability to use both header and text attributes in a single search.

Without this integration, it was noted that searches against these different types of source collections (header or text) could result in different and perhaps inconsistent results being generated. Dr. Nartker expressed the opinion that the sheer volume of documentary material would make a meta search difficult and discussed information retrieval techniques to aid in searching such as thesaurus expansion. He then pointed out that thesaurus expansion would not aid a meta search capability because it increases the size of the result set. He expressed the opinion that query refinement and customization was a necessary tool for accurate searching.

Observations included:

1) that interleaving result sets while preserving the relative position of each document's "relevance" will not be easy;

2) that HTML forms query must be supported by each of the underlying sites (and this could be problematic to those participants on a leased site);

3) that the use of multiple search engines detracts from the consistency of retrieval results;

4) that it reduces the overall capability to a level on par with the least capable software searching provided by any single participant (e.g., some of the sites may not support phrase searches, proximity searching, or combinations of boolean, making the whole system rely on just keywords and resulting in the same 100,000 DOE records showing up on every hit list);

5) that thesauri may not be supported; and,

6) that increasing the required level of sophistication to meet basic functions will levy requirements on participants to provide some search engine capabilities at their site.

In this discussion it was noted that the "lowest common denominator" effect may actually increase cost by requiring additional query tools and strategies, additional user assistance and documentation, increase the requirement for vocabulary management, and require significant customization. It was noted that the greatest risk (of obtaining inappropriate query results) was going to be to the lease skilled users. In a brief analysis of the cost implications of this strategy, it was noted that while it appeared initially to be a less costly approach to implementing the LSN, that by the time that the required additional features were added, it would approach or exceed the cost of simply purchasing the portal approach presented in Alternative 3. It was noted that this may be a "good enough" approach to supporting some of the core requirements of the LSN if the implementation becomes cost constrained. It was also noted that the adoption of Alternative 2 would almost certainly extend the implementation schedule to address the issue of inter-operation and integration of participant search engines with the web site.

Prior to the next discussion, NRC provided a description of one (of many) portal software products he had to opportunity to study prior to the ARP meeting. In that software product, the portal has its own underlying SQL database and full text indexes built from data extracted from target sites. The software he saw had an additional feature of building in a "data dictionary" that kept track of the different field naming conventions encountered in each target collection. This approach allows the portal site to present a single user interface to do search and retrievals. It also allows the participants' sites to act as a backup should the portal site become inoperative (by a user going directly to the participant's homepage), and, allows the portal to continue to identify (but not retrieve) the existence of a document even if the participant's site is temporarily inaccessible. He noted that he attended a Delphi Consulting Group seminar on portals and the business is dominated by perhaps only 4-6 companies with current, deployed, and competitive software.

The concept of a portal was discussed and the roles of its different elements outlined. Its utility as a central caching and replication mechanism was covered. "Gadgets" and "connectors" as middle-ware were defined and their role in a portal's operation explained.

Note to TWG members: Here are two sources mentioning alternative products:

> ". . . most corporate portal vendors we've talked to mention that one of their primary competitors is Plumtree. This means that Plumtree right now is the company to beat in this space. Vendors who offer similar types of content management capabilities are startups Glyphica, KnowledgeTrack, and 2Bridge. . .Viador. . . Sequoia and DataChannel. . . SAP, PeopleSoft, Lawson. . .Netscape and Yahoo!. . ."[4]

> ". . . Pointcast. . .Dataware. . .OpenText Livelink. . .Viador. . .Verity's Search97 / Agent Server / Knowledge Organizer. . ."[5]

The group discussed Alternative 3, a significant complexity effort which is represented by a home page supported by its own databases and indexes compiled as a result of software "crawling" each of the participants' sites. In this approach, typified by http://www.tis.eh.doe.gov/portal/ each participant's web-accessible (outside the firewall) collection may use any number of software management systems for structured data (bibliographic) and unstructured data (text and images) under the operational control of the participant.

WEB SERVERS

DOE Participant Nye Nevada NRC

LSN Index & Centralized Portal

---
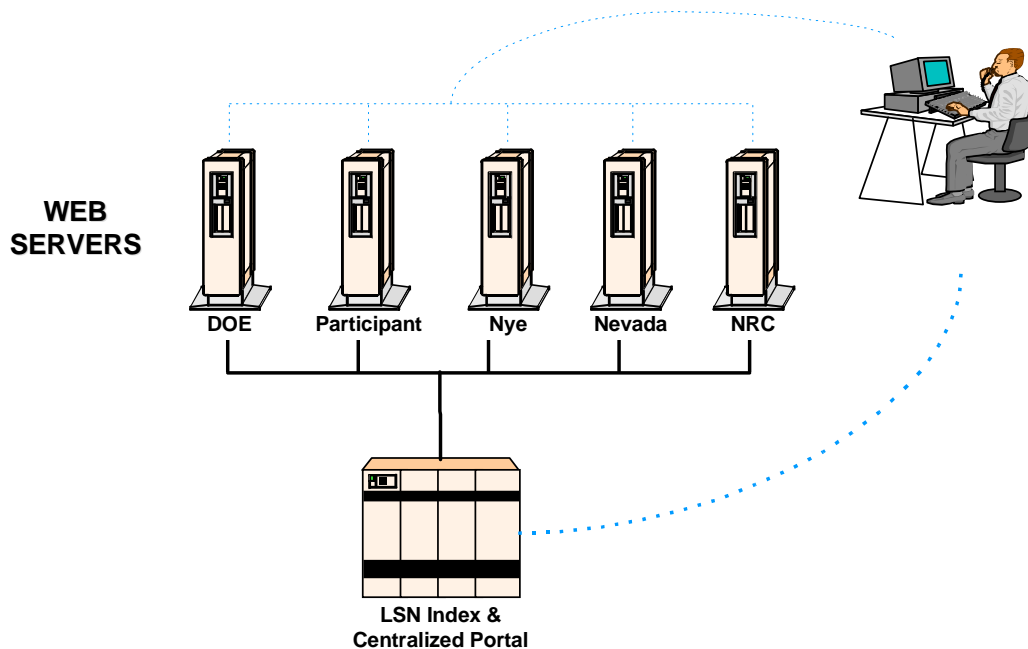
[4]Patricia Seybold Group. "Plumtree Blossoms: New Version Fulfills Enterprise Portal Requirements", in Information Assets: Transforming Information into Profits, June 23, 1999. P.9.

[5]Molly Lyman of Project Performance Corporation.

The LSN portal software scans through these collections and builds its own index to structured data or its own index to text terms found at one of the target sites. Options within Alternative 3 include making decisions about the level to which the system is developed. With more memory, the system can cache the most frequently used files (text or image) right on the portal machine in order to speed response time, but this increases the amount of memory that needs to be stored. The portal could be used to store other media types, such as full motion video or audio files. A decision is required about how much replication would be needed (how much is enough) and what type of replication would be best if the participant sites are not relied upon as the equivalent of "hot site" backup. It was noted that the integration of so much functionality within a single entity would significantly increase its importance and would require a higher standard for availability and reliability.

The issue of priority access led to a discussion of whether the participants' servers' URLs could be hidden so that all users went through the LSN portal in order to access the collections. This would be the only way that service could be prioritized to the participants during the hearing process. DOE indicated that they would support that, but it was unknown as to whether the smaller parties could or would be willing/able to support that approach, especially if they used commercial services. However, with URLs hidden, if any participant's site goes down, then there is no alternative to the portal. This may or may not be a problem, since the LSNA has identified that it is the portal availability and the docket machine host availability that are counted towards "system availability" for meeting a 3 year hearing process. Conversely, if the portal site is not available, then none of the other systems are available, again, because their URLs are hidden. NRC noted that priority access was not supported in one brand of portal software they looked at.

In this approach, it was noted that the portal software gives insight into the IP address of the other sites it is targeting. The group explored the concept of using a VPN (Virtual Private Network) approach in order to establish dedicated bandwidth between participant locations and the portal.   In this approach, security access policies between the participant are established to allow a communications tunnel between sites to be established by use of a second firewall "outside" each site and then using that firewall's software to control the communication channels. The ability of participant using third-party commercial suppliers to implement this is problematic. Much discussion focused on the issue of bandwidth that must be provided between the LSN portal site and each of the participant sites. In addition to bandwidth being an issue (especially during the process of the participant site being "webcrawled"), a sensitivity analysis on the size of the collections, the server platform being used (or, their ISP's capability), might be a worthwhile activity.

Dr. Nartker of UNLV made note of the capabilities of some of the software products they have been evaluating for DOE/YMP, with special emphasis on the Excalibur™ software package. Excalibur provides a capability to establish a uniform software base across multiple sites and it then handles the process of running a distributed query against each site in the enterprise network. It was noted, however, that this would require that all participants license the same software - which would require the LSNA to issue mandates for use well beyond what the LSNA is currently prepared to propose or request. Additionally, this would require all participants to purchase, install, and populate within about a 20 month window and this was deemed non-viable.

Given the variability between participant managed sites and participants who are hosted at an ISP or IVP, it was noted that the LSNA should consider developing classes of standards and guidelines, especially in the areas of security, backup, and recovery. This discussion led to a request that perhaps there should also be classes of standards applied to the other areas of the standards of performance.

DOE representatives proposed a variant on Alternative 3, in which participants would send their documents to the portal site and allow the portal site to act as the LSN host machine for those collections since the portal software was going to build indexes to structured and unstructured text anyhow.

Transmission of data could be accomplished by high density transfer media such as DVDs. They stated that their total collection for this purpose would be about 200GB in size and consist of approximately 200,000 documents. It was noted that configuration management with the DOE scenario could be an issue, and that the DOE scenario moves responsibility for ultimate provision of DOE materials from the DOE to the NRC. However, the DOE proposal would not affect the "front end" aspects of the system.

DOE representatives were asked to develop a writeup of a fourth proposed alternative - Alternative 4. In this alternative, a tightly controlled site holding both NRC and DOE licensing documents is established at the NRC.
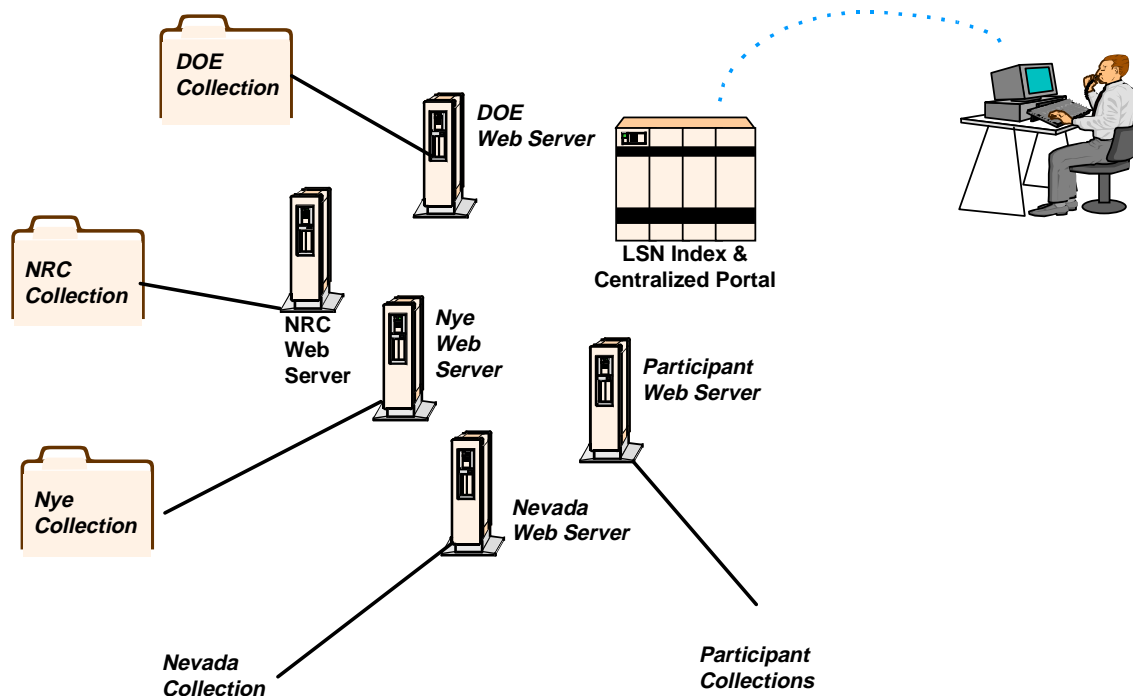
It expands the capabilities of the other proposed solutions in that both DOE and NRC licensing documents would be held in local storage and the remaining participants documents would be replicated and cached as needed. DOE's documents and changes to them would be submitted via a certified transmittal on a preset media and format such as DVD, DLT tape, etc.

This approach provides for tightly controlled access. It requires increasing the hardware required to support Alternative 3. Configuration management issues would need to be resolved before implementation. The policies for and method of certified document transmittal would have to be worked out and tested. In this approach, the primary responsibility for document availability to the public would be shifted to the LSNA.

In the discussion of this approach, it was noted that it is essentially the same architecture as in Alternative 3, and had the benefit of providing a single unified search screen, etc. However, it differs from Alternative 3 in that participant sites may be crawled, or, optionally, that participants such as DOE and NRC could deliver load tapes/CDs to the portal from the participant's internal collections. This idea was iterated and it was noted that those two large collections could be located on the same platform (or, in a cluster configuration) as the portal machine in order to maximize performance. Following that logic, it was noted that a three platform cluster could link a platform with DOE materials, a platform with NRC materials, and, the platform with the portal and also the permanently cached collections of smaller participants. NRC observed that this is not much different than the old LSS except that it is "web-ified". NRC also noted that this may be perceived as NRC providing a capability that is required of the participants by the Rule, which they could do themselves, and therefore has the same effect as providing intervenor funding.

A fine point of distinction between Alternative 3 and Alternative 4 is that while a portal may add value to participant sites in Alternative 3, it should not replace what a party is obligated to do as could be the case in Alternative 4. However, it was agreed that the technical merits of this alternative should continue to be explored by the TWG.

Clustering platforms in close proximity to enhance performance raised questions of system administration. It was noted that depending on where the cluster was located, participants may need to make staff available to support operations at the cluster location rather than try to perform system administration locally. If this is the case, the portal platform and an NRC collection server should be located in LV in order to be closer to the DOE collection, or, the DOE server should be established near NRC and operated out of DOE HQ.

For both Alternatives 3 and 4, there was a following discussion on software that participants may be using that might require the portal site to have additional interfaces developed. A cost sensitivity analysis during the authorization phase of the project would identify the cost of developing interfaces not supported by a portal.

There was also a discussion related to participants having the option in both Alternatives 3 and 4 to either build their own systems or to utilize an ASP (Application Service Provider).

It was agreed that the DOE proposal would have little effect on the audit compliance aspects of the LSN, or no user access to the LSN. It was agreed that the central LSN site was composed of separate functional parts:

–   The baseline audit compliance function - this subsystem is considered to be the responsibility of the LSNA to design since it has no requirement for participant input.

–   The front end with which users interact - this subsystem was discussed in depth the previous day with agreement that a portal provided an acceptable level of functionality.

–   The back end document storage subsystems - this subsystem still has alternatives under consideration. The original alternatives assumed separate sites for participants each publishing their own documentary collections with, perhaps, some participants sharing resources. The DOE proposed an alternative that assembles all or the bulk of the document collection in a single repository with the portal providing access to it.

The group then went through the remaining standards of performance topics/issues to compare and contrast Alternative 3 and Alternative 4.

Integration and Interaction - With regard to integration and interaction between the portal site and the participants' external collections, there seems to be little distinction between the two Alternatives. It was felt that it may be a little easier under Alternative 4 to integrate communications.

Server performance - It was noted that server performance specifications need to be developed.

Text accuracy standards - Dr. Nartker was asked to describe most recent findings. In general, the re-key threshhold has for a long time been held as $\leq$ 95% accuracy (Bradford & Dickey). The best three OCR products on the market, presuming that you are doing manual zoning, now all are capable of $\geq$98% accuracy on office-quality paper source documents. Tests on documents over 10 pages in length indicate that there is not any significant impact effect on either precision or recall. It was noted that dirty data can generate text file index clutter up to five times greater than with relatively clean data; dirty indexes could affect the user's confidence in the retrievability of a document and it could affect relevancy ranking if the term occurrence is the methodology used to generate a relevancy ranking on short documents. It was also noted that in later tests, it was demonstrated that text accuracy did not significantly affect precision or recall in the retrieval of documents under 10 pages in length, either. Scanning from film is not as good as scanning from paper. Xerox™ OCR is best at decolumnizing scanned tables.

The group agreed that all participants would need to adhere to standards (to be developed) for data representation, packaging, and indexing. The LSNA noted that the 1992 bibliographic header list has to be examined and revised with an eye to adjusting to the web environment and possible simplification.

Documentation - In both Alternative 3 and Alternative 4, documentation burdens are similar, focusing mostly on configuration management and exchange standards, although configuration management documentation on ISP or ASP hosts will not be a realistic expectation.

Performance statistics and documentation - In both Alternative 3 and Alternative 4, participant server and portal server statistics would represent the same level of complexity to an

audit server and its software. It was noted that in a clustered configuration (Alternative 4) that the performance statistics may be difficult to segregate because the servers are coupled.

Acceptable formats - There was discussion as to the acceptable formats versus what some participants were already using. NRC's docket environment will require TIFF or PDF submissions. DOE is using TIFF, JPG encoded TIFF, ascii, PDF, and HTML.

Document management and control - It was recognized that both Alternative 3 and Alternative 4 will require the TWG to devise a solution to participant number and records packaging. The issue of NRC/Portal accession numbers and participant accession numbers and how to link them on a unified site was discussed. It was noted that this may require custom code.

Software licensing - Alternative 3 will impose licensing requirements (to varying degrees) on all participants who host their own sites, or, the cost of hosting on an ISP or ASP host machine. Alternative 4 focuses the cost of licenses almost exclusively on the portal location and would therefore require a cost accounting/billing system to be put in place by the LSNA in order to ensure that each participant pays their share-cost. It is problematic to get these proceeds back into the NWF since the only mechanism that the NWPA-AA provides is the 1 mil per kilowatt hour levy against consumers of reactor generated power. Alternative 3 adds license costs over those incurred in Alternative 4 because of the added costs that would be needed to secure the VPN channels.

Search engine performance standards - Under both Alternatives, the portal software should react with similar performance based on the platform horsepower. However, it was noted that under Alternative 3, individual retrievals of text and image files from the participants' file servers might be slower because of the number of calls being made back and forth between the portal and the sites. In either case, a realistic performance metric needs to be developed that considers the impact of the search engines hitting against some collections with only scores of pages while other collections could have well in excess of a million pages of material. Under Alternative 3, the performance standards of the participant servers must be viewed in the context of those machines possibly being the backup resource should the portal site not be operational.

Security - in both Alternative 3 and Alternative 4, physical security will have to be levied on the participants to ensure that "write-protection" is available to the server on which their collection resides. It was evident that the consensus was that no reduction of standards in this area should be considered. Digital signature certificates need to be secured for all electronic document submittal transactions to the docket (this will be provided by NRC LRAA).

Data maintenance - in Alternative 3, this is clearly provided by the participants on their own collections and by the LSNA on the portal indexes. In Alternative 4, the entire burden falls upon the LSNA.

Training - The issue of training was discussed with the consensus being that there is little difference between Alternative 3 and Alternative 4 as far as training was concerned.

2.3.2   Preparation for December TWG Meeting

At the conclusion of the October meetings, the NRC contractor performed a summary analysis comparing the attributes of Alternatives 3 and 4. It built upon the work done by the TWG during the week of 12 October 1999 in which the conceptual technical alternatives described in the LSN Administrator's (LSNA) Compliance Assessment Program (CAP) Guidelines (presented to the LSNARP on 13 October 1999) were evaluated.   Its findings are as follows:

Background The two conceptual design alternatives share many characteristics but differ in a fundamental way. The LSN can be regarded as consisting of three functional components. Specifically, these are:

- A component that aids the LSNA in auditing participant compliance with the LSN Rule.

- A component that presents LSN information to participants, other interested parties, and the general public.

- A component that stores LSN documentary information for the use of components one and two.

The two design alternatives agreed upon by the TWG differ only in the conceptual design of the third component, specifically in how and where LSN materials are stored. The design of the first and second component will not be materially affected by the alternative selected for the third component. The detailed descriptions of components one and two apply to both alternatives currently under consideration. However, there will be differences in the details of implementation and operation.

The sections below will address each of the components in turn and the different alternatives for component three. A separate section will detail how components one and two interface differently with the alternatives for component three.

Compliance Component This component was referred to as the "baseline" system in the LSNA CAP Guidelines document referenced above. It is a "front-end" component (one with which end-users interact) with a small set of users who require specific information at specific times. It is intended to address the in-house needs of the LSNA.

The purpose of this component is two-fold. First, this component ensures that the LSN is functioning as intended and provides assuredness of this functioning to the intended user base. Second, it provides the necessary reports on LSN functionality to enable the LSNA to ascertain that participants are in compliance with the LSN Rule and to aid in determining whether remedial action is required.

The primary method of following the operation and evolution of the LSN is through a reporting mechanism. Reports will be generated automatically by the system on a periodic basis, both when exceptional conditions arise, and on-demand. The full array of required reports is yet to be determined. However, the following have been identified at this time:

- A listing of changes in participant document collections, i.e. additions, deletions, modifications.

- A report on the "health" of the LSN, component and sub-component uptime and performance data (e.g. web server hits, average response times, number of users, etc.)

Exception reports will be used should anomalous conditions arise. Candidates for this type of report include:

- When auditing software detects a possible compliance problem in a participant collection.

- When a component of the LSN itself is determined to be malfunctioning, i.e. due to a computer or network error.

- When a security exception is noted.

It is anticipated that reports may need to be generated from time-to-time to respond to an exception or to "drill down" to garner additional data on a perceived compliance problem. A facility will be provided to expedite this process. It is anticipated that HTML forms will be designed to allow individuals to design and generate most reports on demand. However, it is likely that some reports may need to be developed by systems maintenance personnel from time-to-time.

Certain functions of the system are best accessed through a WWW browser (e.g. Netscape, Internet Explorer, etc.) through the standard HTTP/HTML mechanism bolstered by CGI programs that interact with the data stores. Most commercial and open source network management software currently employ a web-based interface. Specifically, those aspects of the monitoring function that change rapidly can best be monitored through a browser. Examples of these are troubleshooting on-going problems and ascertaining the status of a particular sub-component at a particular time. Historical trends will be maintained both in HTML tables and graphically.

A web browser is also anticipated as the usual interface to generate reports on demand, with an HTML forms interface providing the report and data selection, as well as the formatting function.

It is anticipated that certain reports, especially periodic reports "for the record," will be automatically printed and physically delivered to their intended recipient. Interactive access to the system will be required to produce on-demand reports that have not been anticipated in the design of the web-based, on-demand facility.

File system access is required for ready availability of system logs and other source data for off-line processing and archival.

Data retrieval element to support auditing. This element will consist of one or more programs which will routinely "rove" participant sites, fetching participant data

(documents, statistics, and other) and storing this data pending processing. The exact nature of the data retrieval element will depend on the details of the alternative selected for the storage component, but it is analogous to a "web spider." A web spider, when presented with a starting URL, will traverse all hyperlinks within the body of documents "under" the URL. Through this methodology, it is possible to retrieve and replicate the entire static structure of a web site for further processing.

A data storage element is responsible for storing both data to be processed and the results of that processing. Both file system storage and database storage will be accommodated. The database will be a network-capable SQL relational database that will provide structured data to both front ends.

A data processing element of the system will process the data retrieved, store the results of the processing, and generate the required reports. It is augmented with data reporting tools which consists of several programs that process report outputs into formats appropriate for the delivery mechanisms described above, and assist a user in specification of on-demand reports.

The system requires operational control capabilities that provide a level of assuredness that the systems the LSN is housed on are functioning as required. There are several main sub-elements:

- Security mechanisms. Security sub-elements include a firewall or firewall software, secure remote administration software, and intrusion detection software.

- Network monitoring and management. This sub-element monitors hardware and software and reports outages or sub-optimal operation. It also gathers low-level statistics on network operation for trend and throughput analysis.

- Physical plant and reliability mechanisms. This sub-element provides appropriate environmental and power conditioning and implements disaster recovery, e.g. a backup capability.

A single computer system of the workstation class was conceptually deemed adequate for this functionality. The security sub-element mandates that the system be separate from and more restricted than the computer system (described below) that provides general access. The system should be equipped with the standard components, a graphical display, and a device appropriate for backup. Examples of this type of system include an i386-architecture workstation (e.g. Pentium III "PC") running open-source Unix (e.g. FreeBSD or Linux), or Microsoft NT, a Sun workstation running Solaris, or a Compaq/DEC Alpha running VMS. The primary selection criteria for specific hardware and operating system should be based on security objectives, with specific functionality a secondary (but important) consideration. The following software components will be required: a web server (e.g. Apache, Netscape Enterprise, MS IIS), a database with accompanying report generation software (e.g. PostgreSQL, Oracle, MS SQL Server), firewall software (e.g. IPFW, ipfilter, Firewall-1), network monitoring and management software (e.g. Big Brother, SunNet Manager, HP OpenView), and a web spider (e.g.

MoMspider, BRS/Search, Fulcrum Search Server). Note that the web server, database, and web spider are also part of the presentation component. The same software can be used for both purposes. In addition, it is anticipated that this component will require some custom software, scripts and CGI's rather than full-blown applications. For obtaining network usage statistics and performing monitoring activities, the compliance component will also require SNMP (Simple Network Management Protocol) access to participant web servers and network interface equipment.

Search and Retrieval Component This component is a "front-end" with a large set of users who require access to a wide range of information at arbitrary times. It is intended to fulfill the requirement to provide information to interested parties through WWW technology. It will be a WWW presentation interface with additional sub-components that consist of:

- Introductory and overview documentation.

- Training / tutorial materials on how to use the site to obtain LSN -related information, and the other aspects of the site, and how to submit to the docket.

- Portal software that allows user customization of user interfaces and user document search and access strategies.

- A search facility that allows LSN-wide searching of participant materials, including per-user custom searching strategies.

- Publication of statistical information on LSN participant sites, including site content and performance.

- Aggregation and publication of overall LSN access and usage statistics, e.g., number of hits.

- A web-based interactive forum in which interested parties can discuss or exchange information regarding LSN matters.

- Help-desk assistance for participants and public users with escalation.

- A LISTSERV (e-mail list manager) to allow participants to easily send electronic mail to all interested parties. A number of mailing lists will be created as needed for discussion of specific subjects, including a list with the e-mail addresses of all participants for notification purposes. The LISTSERV software will allow each participant to manage their own subscriptions to interest lists and archive messages to the lists. It is not intended to provide a public LISTSERV function.

The intended user base includes all participants and potential participants, the LSNA and his designees, the press, and the general public.

Web browsers will be the predominant access method to this component. It is anticipated that this will be the sole access method for the majority of users. Browsers

will be used to gain access to general information, participant documentary collections, and to discussion forums.

A single computer system of the server class will be required for this functionality. Examples are as in the previous section but this component will require more processing power and capacity, i.e. a faster CPU or multi-CPU machine, more RAM, bigger disk storage, etc. The primary selection criteria for the hardware is that it should be supported by the portal software selected (the most critical software component).

The following software components will be required: a web server (e.g. Apache, Netscape Enterprise, MS IIS), a database with accompanying report generation software (e.g. PostgreSQL, Oracle, MS SQL Server), firewall software (e.g. IPFW, ipfilter, Firewall-1), a web forum (e.g. UltimateBulletinBoard, WWWboard), and a LISTSERV (e.g. MailMan, majordomo, LISTPROC), and portal software (e.g. Plumtree, Excalibur, Knowledge Center). Note that the web server, database, and web spider are also part of the compliance component. The same software can be used for both purposes.

The storage component represents the "back-end" functionality serving the needs of the front-end components rather than the end-users directly. The data it contains consists of the documents required to be published by participants in accordance with the LSN Rule and accompanying required information. Two alternatives have been identified for providing this functionality.

Storage Alternative 3 - dispersed document storage

This alternative proposes that each participant, assemble, prepare, and publish their own collections of documents on a WWW server. Components one and two will access these collections as WWW clients and perform the necessary operations routinely through participant sites.

This component is the "back-end" that will provide data to the front-end components described above. Participants will make their documentary collections available on a web server located at the site of their choosing and attached to the Internet. Participants are free to establish their own web server, collaborate on a community web server, procure commercial web service, or employ any other provisioning method.

It is anticipated that participants may choose to make their document collections (and ancillary information) generally accessible on the WWW (i.e., other than through the LSN portal site). However, any documents intended to be filed in the licensing process will have to be obtained or cross-referenced through the LSN portal site to ensure the uniqueness, consistency, and traceability of document identification (accession) numbers.

Web access is the primary method by which participant materials are accessed. Access will be interactive (e.g. when a home site front-end user requests a particular document the home site front end will fetch it from the participant's repository) and by batch (e.g. the portal will fetch all materials on the web site, index them, and retain only the

references to the documents for subsequent presentation in response to end-user queries).

Participants are required to make available all documents subject to discovery in standard, LSNA-approved formats on a web site. This consists of the following procedures.

Documents are to be converted to a format that includes an image representation (TIFF/CCITT or TIFF/JPEG), a searchable text file, and a bibliographic header containing metadata about the document. In many cases, this will require scanning and OCR conversion of a paper document. However, if a document exists in electronic format, it may be preferred to perform a more accurate conversion with appropriate software.

The LSNA may allow participants to provide their documentary collections in alternative page-representation formats such as PDF and proprietary word processor formats like Microsoft Word. This will depend on whether the data retrieval software selected for the front-end components is capable of indexing, searching, and otherwise processing these formats. The requirement to provide a bibliographic header for each document will remain regardless of the documents' formats. The bibliographic header is subject to the same retrieval requirements as the source document, e.g. provided as a searchable text file by the web server, as HTTP headers, or from within a database.

Document preparation is potentially the most labor-intensive and costly aspect of building the LSN due to the large number of documents included. Therefore, the burden on a participant is more closely correlated to the number of documents they must prepare than any other factor.

In this alternative, participants will place their documents on the web server of their choice through whatever file transfer mechanism is supported by the web server. This web server must be connected full-time to the Internet through a communications circuit of adequate speed (to be determined by the LSNA) and have a unique IP address and domain name. The domain name and root URL for the documentary collection, and a list of documents, must be provided to the LSNA.

For consistency in retrieval by the front-ends described above, participants may be required to follow a standard format in layout of the web pages that provide access to the documents themselves and accompanying bibliographic header information. Note that many web servers provide a standard way to publish meta information on web-served documents (e.g. by including this information in a file of the same name as the source document in a meta sub-directory). Use of this function may be required by the data retrieval elements of the front-end components.

Participant Hardware and software required It is difficult to determine the exact hardware and software components due to the possibility of collaboration and the differences in the size of the documentary collections of the participants. Foreseeable alternatives for setting up a web server include a dedicated resource at the participant's site, sharing a server with other participants or non-LSN-related web sites, "co-location"

of a participant-owned machine at an IPP (Internet Presence Provider) or outsourcing the entire site to an IPP. Each of these alternatives have a wide range of cost, convenience, assuredness, and administrative issues associated with them.

If a strategy of providing a dedicated web server is adopted, the size of this machine will, again, depend on the size of the document collection the participant is required to make available.

Participants with an extremely small document collection will probably choose to lease web space on an IPP machine or "piggy-back" on another participant's site rather than implement their own web server. The cost of this facility depends on the amount of data published, the bandwidth the site requires, and other metrics. Typical costs for web sites that are appropriate for small participants range from free (of incremental cost over maintaining a basic Internet-access capability) to several hundreds of dollars per month.

For those who choose to operate and maintain their own dedicated resource, a fairly modest machine may be fully satisfactory. An example of this would be an i386 architecture "PC" (e.g. 166MHz Pentium, 128MB RAM, 4GB disk) running an open-source Unix-like operating system (FreeBSD or Linux) and the open-source Apache web server. The total cost (hardware and software) of such a machine at current (4Q99) market prices is under $1,000, and it would accommodate as many as 10,000 documents (at an estimated 250KB per document). Note that operational costs may not be so trivial, especially the disaster recovery aspects (regular backups with off-site storage), and data communications costs. However, resources for these requirements may already exist and participants who choose to share a web server may be able to equitably spread these costs among themselves.

Participants with larger document collections will, naturally, require a more capacious computer system up to or including the one described in storage alternative 2. Note that operational costs will scale as well.

Storage Alternative 4 - aggregated document storage

This alternative proposes that all or a subset of participants' documentary collections be stored at a facility "topologically close" to the facilities that provide the front-end functionality. In networking terms, this equates to a LAN connection between the front-end presentation tools and the back-end data storage in a "LSN campus" of computers.

Specifically, the original proposers of this alternative expressed the desire to co-locate the DOE's documentary collection with the NRC's. However, since this represents over 80% of the total documents within the entire LSN, there is no reason not to logically extend this concept to the total documentary set of all participants.

From one perspective, this alternative can be regarded as comparable to the original Licensing Support System (LSS) in where and how the documentary collection is stored. However, it differs greatly in several ways:

• how the collection is accessed by participants (as described above),

- how the funding allocations are administered,

- how to isolate the function (or malfunction) of one participant's components from another participant's,

- how to assign the responsibility for facility administration and support (in an environment where contending parties share a resource), and

- the procedure established for ensuring participant responsibility for the accuracy and completeness of its documentary collection in the event of an error on the part of its third-party database administrator or the mis-configuration or mis-operation of a shared resource.

SNMP access to all computer systems used within the "LSN campus" will be required for monitoring and management purposes.

File system access (local disk or network disk) is anticipated to be the primary access method to LSN campus-stored document collections. It has certain efficiencies over web access.

Each participant will still have to prepare the documents they are responsible for providing as described above. However, the different access methods may alter the allowable data representation formats and the method of provision of the bibliographic headers (e.g. by "live" access to an SQL database rather than as a text file accompanying the source document).

To preserve the ability of the LSNA to perform independent review and make impartial decisions, a way will have to be found to ensure that the LSNA does not assume ultimate responsibility for ensuring that participant collections are available, this responsibility ultimately rests with each participant regarding their own collections. Participants will provide a resource or agree to share a resource who will place documents into their area on the shared storage facility.

Allowable methods of conveyance from a participant's preparatory facilities to the shared storage are not determined at this time. Options may include magnetic media (tape or disk), optical media (CD-ROM or DVD), or bulk transfer over a network. Note that all allowed options must be associated with a mechanism that provides assuredness that the document actually entered into the designated storage area is a certified copy of the participant's original document.

It is reasonable to expect that the LSNA will designate separate storage areas for each participant's collection (rather than commingling collections). These separate storage areas may actually be separate machines provided and maintained on the campus LAN by the participants themselves.

A storage component that houses, potentially, the entire LSN pre-discovery documentary collection is quite massive (between 200GB and 4TB by various

estimates). Such size will require a very robust computer system or multiple computer systems of the server or enterprise server class.

Software included with the operating system will provide file system access over the campus LAN to the front-end components. This functionality is generally included in both open source and commercial server operating systems (or is readily available) to support the most common network file system sharing schemes (NFS - Network File System, SMB -Server Message Block, NCP - NetWare Core Protocol, etc.)

Since the database required by the front-end components has a large data-storage requirement, and the standard method of access is over the network, performance issues may dictate that the software implementing the database server functionality run on the storage server instead of the machine implementing the front-end. However, this is the same software component described above, not another database implementation.

Comparison between Alternative 3 and Alternative 4 The key differences between the two alternatives for the storage component are in the areas of:

• 	component integration,

• 	participant procedures for document preparation and publication, and

• 	funding of implementation.

In alternative 3, the various front-ends and back-ends are loosely-coupled (a distributed system). In alternative 4 they are, from the point-of-view of an external observer, effectively a single system (the LSN campus). The usual trade-offs for such alternatives apply.

Alternative 3 - Distributed system In the case of distributed systems:

– 	Independent operation of various sub-components allows partial functionality in the case of system and or network outages. For example, if the portal/presentation component becomes unavailable, access to participant documentary collections is still available directly through participant web sites. In contrast, a single or tightly-coupled system will most likely be completely unavailable if a single sub-component fails.

– 	The system can be more easily reconfigured or extended without disruption to the system as a whole. Since the design emphasis is on the interfaces between distributed stand-alone systems, another stand-alone system (with the correct interface) can be "plugged-in" as an additional component. For example, midway through LSN implementation, a participant may find that the initial server selected for implementation cannot handle its entire documentary collection. In the distributed alternative they can obtain another server rather than start again from scratch with a larger machine.

–   Additional functionality for participant sites is possible. Participants may choose to publish information on their LSN web site that is not part of the LSN (in that it does not relate to the high-level nuclear waste repository licensing procedure). In alternative 2, the LSNA will likely choose to host only LSN-related materials.

–   Individual participants, or collaborations of participants, assume all responsibility for publication of their documents and have a well-defined point at which this publication can be assessed for compliance, i.e. their web site interface. In alternative 2, it may be difficult to ascertain which party is responsible for perceived non-compliance. Participants will have to place a large measure of trust in the entity selected to manage and maintain the LSN campus.

–   In the current computer marketplace, indications are that the aggregate cost of the distributed system would be lower. Specifically, this is due to the availability of commodity computers at extremely low prices. The larger, data-center machines required by alternative 2 generally cost significantly more than the equivalent computing power in commodity machines.

–   Spreading the resources available for implementation over multiple sites usually shortens the roll-out period. The "many hands make light work" principle applies (even though there may actually be more work in total). In the case of a single large system, the large burden of implementation placed on a single staff can result in implementation delays.

Alternative 4 - the LSN campus In the case of an LSN campus:

–   Aggregate performance will likely improve, and will certainly be more predictable. In contrast, the performance of distributed systems is generally much more variable.

–   The tightly-coupled nature of the aggregated facilities will allow efficiencies in communications to occur. For example, it will allow high-speed LAN connections between components rather than slower long distance telecommunications circuits. This has implications in custom software development and fine-tuning performance in off-the-shelf software.

–   The reduced number of sub-components will increase average reliability (at the cost of increasing the consequences of a system-wide outage) and the security of the system will be more easily assured.

–   Technical expertise required by participants is lessened (but greatly increased at the LSN campus site). Participants will not have to be webmasters or acquire webmaster services. In a distributed Alternative, web site maintenance capability will have to be in-house or acquired.

–   The system, as a whole, will be easier to manage and maintain, and will require fewer total resources for this function, than distributed, stand-alone systems. Additionally, the responsibility for management and maintenance will be more

clearly defined. However, the level of maintenance and management of the LSN campus system, and the expertise required to accomplish it, will increase in direct proportion to its size.

It is not clear how a shared storage component will be funded. In the case of the original design concept (represented by alternative 1), each participant is responsible for funding publication of their own collection. Aggregating some or all of the collections may make appropriate allocation of costs difficult. The usual issues in any compensation situation arise, e.g., what happens when a participant doesn't pay what it owes or when a participant loses standing but still owes a contribution? The risk of unanticipated expenditures is shifted from one participant for their own requirements to all participants and, primarily, the NRC as LSN campus host.

2.3.3   December TWG Meeting

General discussion of Alternatives 3 and 4 developed at the October TWG meeting focused on the fact that in any system that will link existing repositories, bandwidth will be the over-arching issue and that any implementation strategy will need to address this. In the two viable Alternatives, it was noted that servers can be scaled to size and enhanced to address performance issues. Additionally, it was noted that bandwidth sizing is optimized at a point just above the usage spikes. In the discussion, it was also noted that the bandwidth issue was most likely to be impacted in trying to deal with image handling.

These observations led into a general discussion of the system's architecture attributes that would most directly bear on its performance in handling requests for large files. Having large text and image files reside on participant maintained storage devices provides a "multi-pathway" capability, thereby spreading out bandwidth impacts to some degree although with 85% of the documents being at the DOE site the impacts may still be felt, thus placing a larger burden on DOE's bandwidth capacity. Multi-pathway is a predominant feature of the original Alternative 1 and Alternative 2 which were, for the most part, discounted at the previous TWG meetings.

Conversely, developing the system in a central campus means that only one feed will need the higher bandwidth, minimizing the set of connections needed, localizing the area, and requiring dedicated lines. The campus approach is simpler to design in a way that will ensure performance, has bounded costs, and is more manageable for backup, recovery, load-balancing, etc. It was noted that caching is what creates a localized effect, and enhanced performance is not based on where documents are located. This led to the proposition that another architecture could be considered, a distributed portal approach which retains a complete cache of each participant's holdings.

In that approach, the cache needs to be at a high-speed location, right at the entrance to "big bandwidth". Approaches that do not heavily utilize a cache require proxies over to a participant operated storage device and then use the multi-path approach to directly delivering files to the requestor.

WEB SERVERS

DOE | Participant | Nye | Nevada | NRC

LSN Index & Centralized Portal

Text & Image Cache
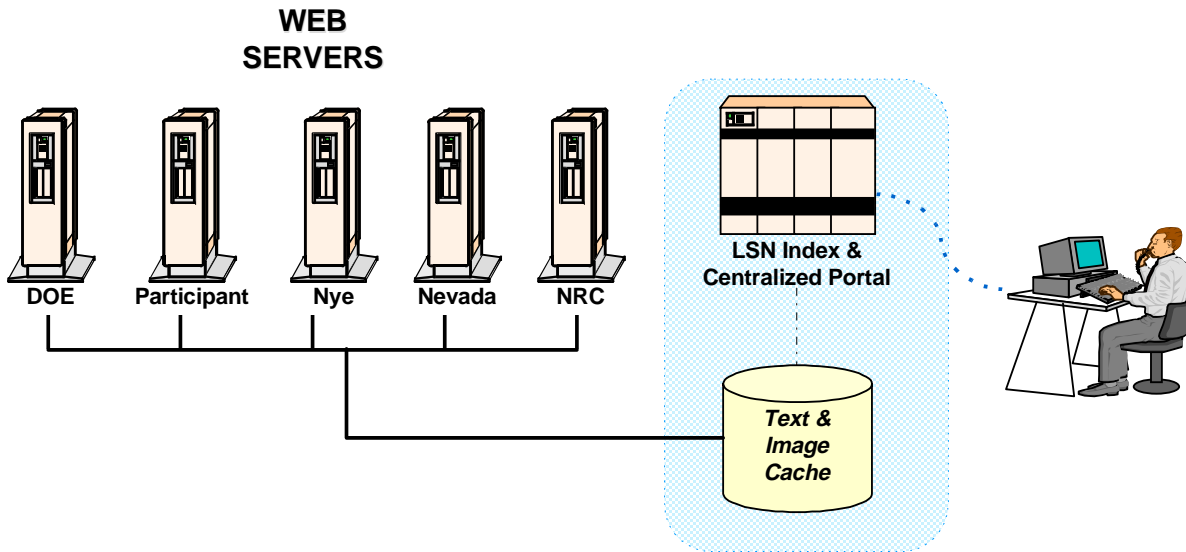
The general distinctions were then categorized as being:

- **A Comprehensive Distributed Portal with Participant Remote Storage** (Alternative 3) typified by a remote portal with software that only maintains indexes, and by participant sites in which the participants each maintain their collection. Their several collections represent the single source of document, header, and image files (except for backup).

- **A Comprehensive Campus Portal** (Alternative 4) typified by a centralized portal with software that only maintains indexes, and by participant maintained file storage and backup devices that are proximate to the portal device. [6] Participants provide for and maintain their collection, and their several collections represent the single source of document, header, and image files (except for backup).

---

[6] "Close" does not mean geography. Close is defined by the nature of the communications between the machines, specifically that it is quick (high bandwidth and low latency), predictable, and private. For our purposes high bandwidth is somewhere over 25Mbps and low latency is less than 5 ms average with less than .5 ms std. deviation.

It is certainly possible to achieve these performance figures with geographically dispersed systems through the use of appropriate technologies (e.g. 100baseF, a FDDI ring, or DS3 telco circuit) but the latency requirement limits the total distance that can be spanned (to about 100 miles) and the type of circuits that can be used (e.g. no satellite circuits need apply).

In the specific instance of Summerlin to UNLV, this could pretty easily be accomplished with a DS3 or ATM circuit leased from SW Bell (or whoever). This is not cheap, a SWAG is $10-20K per month plus $150K-$300K equipment at each end.

If line-of-sight can be established, it would be possible to use microwave or laser equipment at each end with no recurring costs (uwave = $50K-100K, laser = $15K-$50K). Of course, LOS technologies are subject to weather disruptions but that is probably not too much of a problem in LV. Air rights may have to be secured to avoid disruptions from construction.

It is not realistic to expect to be able to pull a single fiber cable between the sites. Any other hard-wire approach depends on the nature of the rights of way that can be secured and the specific physical topology of interconnections that would result. You would need repeater equipment at each interconnection.

Traditionally, use of the term "campus" has indicated that a single entity controls the physical plant that the gear and interconnections occupy. This includes the ability to trench and install cable. What you are describing is a "multi-campus" situation.

- **A Comprehensive Distributed Portal with Enhanced Central Storage** (Alternative 5) typified by a remote portal with software that maintains indexes and a cached copy of all document, header, and image files. Participant collections are downloaded and the portal caches a copy of the participants' files and thereafter uses the cached items exclusively for general search and retrieval.

DOE representatives then introduced a discussion in which the essential technical attributes of the LSN system were identified. These included the ability to 1) provide a high degree of control that can be exercised by the LSNA; 2) ensure timely availability of the system to support the licensing process; and 3) deliver the highest performance at the least cost. NRC noted that these technical attributes reflect the basic mission of the LSNA: 1) to deliver a web-based system that makes all documents equally available in a uniform way, 2) do so in an environment that can be independently audited for compliance, and 3) which provides the tools for ensuring that the system overall performs with acceptable responsiveness.

The group discussed backup/redundancy and noted that the presence of an enhanced central storage facility would lessen the participants' requirement to implement rigorous backup and disaster-recovery procedures (since the central storage facility would be an implicit backup). However, this does not alleviate them of their responsibility to provide and preserve the "true copy" of a document.

The group discussed performance enhancement and noted that this is easier to accomplish via a campus approach, especially if the portal server is modular and multi-processor based.

The group noted that centralized cache storage in a campus location provides the best control, the cheapest overall storage-per-document, and was more predictable.

The group noted that the Alternative where the portal is remote and the participants maintain their own collection storage servers will cost more to fix if there are performance problems which should be anticipated especially in large text file and image file transfers to users.

The group then discussed the issue of caching: the distinction between what it will take for the LSNA to ensure system performance and responsiveness viz NRC providing a capability which the participants are required to deliver. There is still also an open issue of certification of records for use in hearing and other legal proceedings which must be done by the submitter - and the fact that the chain of custody goes through the portal site (and the LSNA) in any option where the portal caches everything and that is the file being relied upon. It was decided this is an issue for the full LSNARP to consider.

The group finally discussed overall cost elements in the life cycle and noted that while it may be cheaper to ease in the door with the Alternatives that do not rely on centralized cache/memory, that in the long term the solutions where participants maintain decentralized data stores may prove much more labor intensive on an ongoing basis to ensure system control and performance. DOE representatives noted that the cost of memory in the terabyte range has gotten down to the $300-400K range. NRC noted that a recent RAID implementation in that class cost in the $700-800K range; but all agreed that memory/storage costs were declining and could be expected to continue doing so when equipment purchases occur next year. Dr.

Nartker observed that delivering and sustaining performance will be the biggest technical problem confronting the operational phase of the LSN. DOE noted that it was easier to initially over-engineer the system rather than to try to remediate performance on a system that is architecturally constrained.

At this point the group decided to start developing a presentation chart which could be used in presenting the issues and recommendations to the full LSNARP. It was decided that the two Alternatives discarded at the initial TWG meetings should be included in this chart so that the TWG's evaluations could be documented with the same detail as those options still in consideration. See the charts on the pages following.

Discussion closed on the issue of functional requirements and the difficulties that were being encountered. It was noted that something would have to be done because they will be needed for procurement and also for acceptance testing. DOE representatives made an observation that detailed capability requirements such as print, deliver paper, storing canned queries, etc., were, of course, causing problems because the nature of the system is now connecting diverse collections and we are looking at the technologies to do that which are commercial-off-the-shelf (COTS). [E.g., we're purchasing a method to connect existing collections so the FR's need to reflect that as opposed to reflecting the attributes of a licensing methodology management system. The one could be COTS but the other is definitely custom. If we have FR's for a licensing software environment, when we try to do test and acceptance against the COTS portal, we will have disconnects and failed requirements, or, we will walk into a commitment for high degrees of customization that may preclude any COTS portals. So, the revelation was that NRC will have to spec to meeting a different mission and will rework the FR's.] NRC will rework the functional requirements to reflect the mission of providing connectivity and performance in a web environment, rather than focusing on the attributes of a legal support environment.