From: Dan Graser
To: Abby Johnson, Alan Kall, Andrew Remus, Bob Wells, Bollwerk, G. Paul, Claudia Newberry, Debra Kolkman, Dennis Bechtel, Eve Culverwell(...)
Date: Fri, Nov 5, 1999 5:21 PM
Subject: LSN ARP Technical Working Group Meetings Minutes

Please let me know if you are unable to open the

.wpd = Wordperfect
.doc = Word '97

versions and I can try something different if need be.

Mail Envelope Properties (38235877.4A1 : 1 : 18974)

Subject: LSN ARP Technical Working Group Meetings Minutes
Creation Date: Fri, Nov 5, 1999 5:21 PM
From: Dan Graser

Created By: DJG2.TWF2_PO.TWFN_DO

Recipients	**Action**	**Date & Time**
internet	Transferred	11/05 5:22 PM
"Bob.Wells@rw.doe.gov" (Bob Wells)		
"Clark.Ray@EPA.GOV" (Ray Clark)		
acj.carson-city.nv.us	Transferred	11/05 5:22 PM
abby (Abby Johnson)		
anv.net	Transferred	11/05 5:22 PM
JudyTF (Judy Treichel)		
aol.com	Transferred	11/05 5:22 PM
hoylej (John Hoyle)		
MalMurphy (Malachy Murphy)		
nvtapper (Les Bradshaw)		
Tiffanih (Nick Stellavato)		
tuftam (Tammy Manzini)		
caliente.igate.com	Transferred	11/05 5:22 PM
jcciac (Eve Culverwell)		
co.clark.nv.us	Transferred	11/05 5:22 PM
dax (Dennis Bechtel)		
evt (internet:evt@co.clark.nv.us)		
cs.unlv.edu	Transferred	11/05 5:22 PM
taghva (Internet:taghva@cs.unlv.edu)		
tom (INTERNet:tom@cs.unlv.edu)		
eurekanv.org	Transferred	11/05 5:22 PM
lfiorenzi (Leonard Fiorenzi)		
gfoster.com	Transferred	11/05 5:22 PM
gfoster (internet:gfoster@gfoster.com)		
govmail.state.nv.us	Transferred	11/05 5:22 PM
madams (Marta Adams)		

ssteve (Steve Frishman)		
idsely.com wpnucwst (Debra Kolkman)	Transferred	11/05 5:22 PM
labat.com Joseph_speicher (internet:Joseph_speicher tony_neville (INTERNet:tony_neville@labat	Transferred	11/05 5:22 PM
NCAI.org Robert_Holden (Robert Holden)	Transferred	11/05 5:22 PM
nei.org spk (Steven Kraft)	Transferred	11/05 5:22 PM
Notes.YMP.gov Claudia_Newbury (Claudia Newberry) Jill_Schrecongost (Jill Schrecongost) John_Gandi (John Gandi) Lew_Robertson (internet:Lew_Robertson@not	Transferred	11/05 5:22 PM
phonewave.net cccomp (Alan Kall)	Transferred	11/05 5:22 PM
sierra.net escorop (Tony Cain)	Transferred	11/05 5:22 PM
smtp.winston.com STrubatc (internet:STrubatc@smtp.winston.	Transferred	11/05 5:22 PM
telis.org inyoplanning (Andrew Remus)	Transferred	11/05 5:22 PM
chris berlien (internet:chris.berlien@terraspec	Transferred	11/05 5:22 PM
elaine ezra (internet:elaine.ezra@terraspectra.c	Transferred	11/05 5:22 PM
threeputt.hawthorne.nv.us wallace (Jackie Wallace)	Transferred	11/05 5:22 PM
winston.com strubatc (Sheldon Trubatch)	Transferred	11/05 5:22 PM

ymp.gov	Transferred	11/05 5:22 PM
David_hunt (internet:David_hunt@ymp.gov)		
harry_leake (internet:harry_leake@ymp.gov		
sam_hobbs (internet:sam_hobbs@ymp.gov)		
twf1_po.TWFN_DO	Delivered	11/05 5:22 PM
GPB (G Paul Bollwerk)	Opened	11/07 1:23 PM
JGW (Jack Whetstine)	Opened	11/09 8:39 AM
JXH5 (John Hoyle)	Opened	11/09 8:05 AM
TSM2 (Thomas Moore)	Opened	11/06 10:47 AM
twf2_po.TWFN_DO	Delivered	11/05 5:21 PM
HJS (Harvey Spiro)	Opened	11/08 2:34 PM

Post Office	**Delivered**	**Route**
internet		
acj.carson-city.nv.us		internet
anv.net		internet
aol.com		internet
caliente.igate.com		internet
co.clark.nv.us		internet
cs.unlv.edu		Internet
eurekanv.org		internet
gfoster.com		internet
govmail.state.nv.us		internet
idsely.com		internet
labat.com		internet
NCAI.org		internet
nei.org		internet
Notes.YMP.gov		internet
phonewave.net		internet
sierra.net		internet
smtp.winston.com		internet
telis.org		internet
chris		
	internet:terraspectra.com	
elaine		
	internet:terraspectra.com	
threeputt.hawthorne.nv.us		internet
winston.com		internet
ymp.gov		internet
twf1_po.TWFN_DO	11/05 5:22 PM	
twf2_po.TWFN_DO	11/05 5:21 PM	

Files	**Size**	**Date & Time**

twg_minutes.wpd	72744	Friday, November 5, 1999 5:11 PM
twg_minute.doc	70458	Friday, November 5, 1999 5:12 PM
MESSAGE	796	Friday, November 5, 1999 5:21 PM

Options

Auto Delete:	No
Expiration Date:	None
Notify Recipients:	No
Priority:	Standard
Reply Requested:	No
Return Notification:	None
Concealed Subject:	No
Security:	Standard
To Be Delivered:	Immediate
Status Tracking:	Delivered & Opened

Licensing Support Network Advisory Review Panel
Technical Working Group Meetings
October 12, 14, and 15

October 12, 1999

Attending:

Dan Graser	NRC/ASLBP	(301)415-7401	djg2@nrc.gov
Glen Foster	NRC/Labat	(703)598-3759	gfoster@gfoster.com
John Gandi	DOE/YMP	(702)794-1313	john_gandi@ymp.gov
E. v. Tiesenhausen	Clark Co	(702)455-5184	evt@co.clark.nv.us
Thomas Moore	NRC/ASLBP	(301)415-7465	tsm2@nrc.gov
Paul Bollwerk	NRC/ASLBP	(301)415-7454	gpb@nrc.gov
Jack Whetstine	NRC/ASLBP	(301)415-7391	jgw@nrc.gov
Chris Berlien	Nye Co	(702)795-8254	chris.berlien@terraspectra.com
Elaine Ezra	Nye Co	(702)795-8254	elaine.ezra@terraspectra.com
Tony Neville	NRC/Labat	(703)506-1400x506	tony_neville@labat.com
Joe Speicher	NRC/Labat	(703)506-1400x835	joseph_speicher@labat.com
Sam Hobbs	M&O/YMP	(702)295-5472	sam_hobbs@ymp.gov
Tom Nartker	UNLV-ISRI	(702)895-0848	tom@isri.unlv.edu
Harvey Spiro	NRC/OCIO	(301)415-5862	hjs@nrc.gov

At the kickoff meeting of the technical working group, each TWG attendee was provided a copy of the binder containing the handouts that would be used at the ARP meeting on 10/13/1999. Dan Graser of the NRC listed the following agenda:

A. TWG Ground Rules, Charter & Objectives
B. Survey Results
C. LSN Project Schedule and Gantt Chart
D. Three General Scenarios (correspond to level of integration)
E. Plan for two working days after the ARP meeting

He noted that the TWG operated as an extension of the full LSNARP as per the charter included in the binder. (Binder Tab D) The TWG meetings are informal and anyone who has an interest in the technical aspects of the system is invited to attend or have a representative present. The TWG performs any investigation, research, or analysis as is directed by the full ARP, and provides various products, analyses, presentations, etc., back to the ARP for their consideration and possible action. He noted that additional TWG meetings will likely occur. These will be held in Las Vegas to minimize the scheduling requirements of the majority of the participants.

There was a brief overview of results received to date (Binder Tab E) to the LSNA's survey of the participant/potential participant/AULG's current Internet availability and future plans. Dan Graser noted that in general those responding seemed to demonstrate a high degree of sophistication and understanding of Internet technologies.

It was stated that the volume of documentary material does not seem to be as difficult to manage as before, partly because of advances in technology and partly because changes in the

rule affect the scope of the collections. Current projections of the amounts of documentary material to be handled in the LSN will probably have to be modified as the design progresses.

DOE noted that the overall licensing strategy that DOE intends to follow has not yet been determined. In light of this it was noted that the Topical Guidelines would likely be modified to more closely coincide with DOE's licensing approach.

There was a brief overview of the LSN Project Schedule based on the Gantt Chart (Binder Tab F). Dan Graser noted that this was a strawman and that many of the tasks and time frames following the recommendation of a solution to NRC's Executive Council were speculative, and would have to be revised in order to reflect any recommended/selected solution to meet LSN functionality. He explained the NRC Capital Planning and Investment Control (CPIC) process, and its impact on the preliminary project schedule. DOE offered to assist in the CPIC process with documentation, justification, or other assistance. Dan explained that the work of the TWG comprised much of the required CPIC documentation. He noted that the project timeline demonstrated that work needs to begin now if there is any hope of meeting a July 2001 readiness date. He also noted that the schedule indicated the need for access to some participant collection materials in order to test the system connectivity and performance and that those documents would need to be available before the actual date of availability required in the rule.

DOE indicated that they were exploring the possibility of significant modifications to their DBMS that housed their bibliographic headers. This was discussed with the ramifications of changes in "mid-stream" being pointed out. Dan Graser stated that he was receptive to necessary changes but that the schedule should be considered in making them.
There was a high level characterization of three scenarios that would be briefed to the ARP in full session and NRC indicated that the task of the TWG was to explore the technical feasibility of those approaches and additionally to iterate those solutions or propose other approaches that would then be fleshed out, priced out, and presented in an analysis to the full ARP.

There was a general discussion about how to go about reviewing the three strawman solutions when the TWG began its meetings on Thursday. Additional background materials were handed out to the TWG members, including:

- Screen shots of a DOE/ES&H portal site at http://www.tis.eh.doe.gov/portal
- Graphic showing participant commitment/expanding functionality of 3 systems
- Matrix of 3 systems (coverage, website functionality, software functionality, hardware functionality, and communications functionality of each approach)
- LSN Standards of Performance Issues
- Compilation of 1995 LSS functional requirements (Level 1 & 2) with preliminary commentary

The electronic information exchange (EIE) submission process was explained with a discussion of the issues surrounding secure document transfer. It was noted that the docket submission would use NRC EIE standard practices. It was also noted that EIE scheduling considerations were not specifically addressed in the project schedule.

Several additional areas that the TWG will address were outlined:

- Functional Requirements
  - How they will change

- How they will apply - must support mission
- Who is responsible for determining them - LSNA; but presiding officer can adjust if necessary
- Necessity of functional requirements being in place before LSN design goes forward

- Bibliographic Headers
- Document Packages solution

October 14, 1999

Attending:

Dan Graser	NRC/ASLBP	(301)415-7401	djg2@nrc.gov
Glen Foster	NRC/Labat	(703)598-3759	gfoster@gfoster.com
Lew Robertson	MTS/YMP	(702)794-5077	lew_robertson@ymp.gov
Tom Nartker	UNLV-ISRI	(702)895-0848	tom@isri.unlv.edu
Harry Leake	M&O/YMP	(702)295-5531	harry_leake@ymp.gov
Kazem Taghva	UNLV-ISRI	(702)895-0873	taghva@cs.unlv.edu
Harvey Spiro	NRC/OCIO	(301)415-5862	hjs@nrc.gov
Thomas Moore	NRC/ASLBP	(301)415-7465	tsm2@nrc.gov
Paul Bollwerk	NRC/ASLBP	(301)415-7454	gpb@nrc.gov
Chris Berlien	Nye Co/TSG	(702)795-8254	chris.Berlien@terraspectra.com
Sam Hobbs	M&O/YMP	(702)295-5472	sam_hobbs@ymp.gov
David Hunt	MTS/YMP	(702)794-5571	david_hunt@ymp.gov
John Gandi	DOE/YMP	(702)794-1313	john_gandi@ymp.gov
Dennis Bechtel	Clark Co	(702)455-5178	dax@co.clark.nv.us

Glen Foster of Labat-Anderson (supporting the LSN Administrator) began the session with a walk through technical description of each of the alternative scenarios and emphasized that all the approaches considered will have to be able to be "plugged into" by NRC software used to audit the content and performance of individual participant sites.

The group generally discussed scenario A, which was characterized as a relatively non-complex development of a web page that provides links to the various sites. In this scenario, 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 scenario A 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 scenario A did not meet requirements for the following reasons:

- Too complex for users
- Too difficult to navigate
- Not possible to aggregate information

- User interface not consistent
- Not versatile
- Does not meet needs of large, complex discovery system
- Potentially excludes some participants and "tilts the playing field" for others.

Agreement was reached that scenario A would not be recommended to the full LSNARP.

The group discussed scenario B, 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 scenario B'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 this scenario 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 scenario C. 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 scenario B 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, Dan Graser 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!. . ."[1]

> ". . . Pointcast. . .Dataware. . .OpenText Livelink. . .Viador. . .Verity's Search97 / Agent Server / Knowledge Organizer. . ."[2]

The group discussed scenario C, 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. 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 scenario 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

---

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

[2]Molly Lyman of Project Performance Corporation.

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 the 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 scenario C, 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 - scenario D.

October 15, 1999

Attending:

Dan Graser	NRC/ASLBP	(301)415-7401	djg2@nrc.gov
Glen Foster	NRC/Labat	(703)598-3759	gfoster@gfoster.com
Lew Robertson	MTS/YMP	(702)794-5077	lew_robertson@ymp.gov
Tom Nartker	UNLV-ISRI	(702)895-0848	tom@isri.unlv.edu
Harry Leake	M&O/YMP	(702)295-5531	harry_leake@ymp.gov
Sam Hobbs	M&O/YMP	(702)295-5472	sam_hobbs@ymp.gov
David Hunt	MTS/YMP	(702)794-5571	david_hunt@ymp.gov
John Gandi	DOE/YMP	(702)794-1313	john_gandi@ymp.gov

DOE representatives delivered a summary writeup of a fourth proposed alternative - scenario D. 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 Scenario C. 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 scenario C and had the benefit of providing a single unified search screen, etc. However, it differs from scenario C 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 scenario C and scenario D is that while a portal may add value to participant sites in scenario C, it should not replace what a party is obligated to do

as could be the case in scenario D. 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 scenarios C & D, 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 scenarios C and D 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:

1. 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.

2. 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.

3. 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 scenario C and scenario D.

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 scenarios. It was felt that it may be a little easier under scenario D to integrate communications because

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 scenarios C and D, 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 scenarios C and D, 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 (scenario D) 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 scenarios C and D 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 - Option C 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. Option D 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. Scenario C adds license costs over those incurred in scenario D because of the added costs that would be needed to secure the VPN channels.

Search engine performance standards - Under both scenarios, the portal software should react with similar performance based on the platform horsepower. However, it was noted that under scenario C, 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 scenario C, 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 scenarios, 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 scenario C, this is clearly provided by the participants on their own collections and by the LSNA on the portal indexes. In scenario D, the entire burden falls upon the LSNA.

Training - The issue of training was discussed with the consensus being that there is little difference between scenario C and scenario D as far as training was concerned.

ACTION ITEMS

Action items and assigned responsibilties are as follows:

1. Develop a strawman revised version of Functional and Performance Requirements (including scenario C & D server performance specifications) for a Web-based LSN system (NRC).

2. Develop recommendations on changes to bibliographic headers (NRC)

3. Develop more detailed descriptions for the two viable alternatives (NRC-Labat)

4. Develop ballpark pricing estimates for the two viable alternatives (NRC-Labat)

5. Identify portal software vendors. Identify if any of them operate on non-NT systems (e.g, UNIX?) (NRC)

6. Explore tools used for corporate data mining and find out if any of them have multi-repository and web-based products. (NRC)

7. Contact DOE/ES&H to determine if performance statistics are kept on their portal site. (NRC)

8. TWG needs to address records packages and participant document numbering strategies.

The following items will be prepared for the ARP some time after the TWG reviews the above material and provides additional input.

- A technical alternative decision tree
- A chart showing salient factors of and differences between the alternatives
- Cost profiles for each alternative

It was agreed that NRC would take the lead in drafting meeting minutes, that the draft would be circulated to all TWG members for additional comments or input, NRC would finalize the minutes and then distribute the minutes to the entire TWG and ARP mailing lists.

GLOSSARY

ASP An application service provider (ASP) is a company that offers individuals or enterprises access over the Internet to application programs and related services that would otherwise have to be located in their own personal or enterprise computers. Sometimes referred to as "apps-on-tap," ASP services are expected to become an important alternative, especially for smaller companies with low budgets for information technology. Early applications tend to be generalized and include:

- Remote access serving for the users of an enterprise
- An off-premises local area network (LAN) to which mobile users can be connected, with a common file server
- Specialized applications that would be expensive to install and maintain within your own company or on your own computer

Hewlett-Packard, SAP, and Qwest have formed one of the first major alliances for providing ASP services. They plan to make SAP's popular R/3 applications available at "cybercenters" that will serve the applications to other companies. Microsoft is allowing some companies to offer its BackOffice products, including SQL Server, Exchange and Windows NT Server on a rental, pay-as-you-use basis.

While ASPs are forecast to provide applications and services to small enterprises and individuals on a pay-per-use or yearly license basis, larger corporations are essentially providing their own ASP service in-house, moving applications off personal computers (referred to as thin clients) and putting them on a special kind of application server that is designed to handle the stripped-down kind of thin client workstation. This allows an enterprise to reassert the central control over application cost and usage that corporations formerly had in the period prior to the advent of the PC. Microsoft's Terminal Server and Citrix's WinFrame products are leading thin-client application server products.

DVD DVD (digital versatile disk) is an optical disk technology that is expected to rapidly replace the CD-ROM disk (as well as the audio compact disc) over the next few years. The digital versatile disk (DVD) holds 4.7 gigabytes of information on one of its two sides, or enough for a 133-minute movie. With two layers on each of its two sides, it will hold up to 17 gigabytes of video, audio, or other information. (Compare this to the current CD-ROM disk of the same physical size, holding 600 megabytes. The DVD can hold more than 28 times as much information!)

DVD-Video is the usual name for the DVD format designed for full-length movies and is a box that will work with your television set. DVD-ROM is the name of the player that will (sooner or later) replace your computer's CD-ROM. It will play regular CD-ROM disks as well as DVD-ROM disks. DVD-RAM is the writeable version. DVD-Audio is a player designed to replace your compact disc player.

DVD uses the MPEG-2 file and compression standard. MPEG-2 images have four times the resolution of MPEG-1 images and can be delivered at 60 interlaced fields per second where two fields constitute one image frame. (MPEG-1 can deliver 30

noninterlaced frames per second.) Audio quality on DVD is comparable to that of current audio compact disks.

HTML HTML (Hypertext Markup Language) is the set of "markup" symbols or codes inserted in a file intended for display on a World Wide Web browser. The markup tells the Web browser how to display a Web page's words and images for the user. The individual markup codes are referred to as elements (but many people also refer to them as tags).

HTML is a standard recommended by the World Wide Web Consortium (W3C) and adhered to by the major browsers, Microsoft's Internet Explorer and Netscape's Navigator, which also provide some additional non-standard codes. The current version of HTML is HTML 4. However, both Internet Explorer and Netscape implement some features differently and provide non-standard extensions. Web developers using the more advanced features of HTML 4 may have to design pages for both browsers and send out the appropriate version to a user. Significant features in HTML 4 are sometimes described in general as dynamic HTML. What is sometimes referred to as HTML 5 is an extensible form of HTML called XHTML.

HTML Forms Web forms let a reader return information to a Web server for some action. For example, suppose you collect names and email addresses so you can email some information to people who request it. For each person who enters his or her name and address, you need some information to be sent and the respondent's particulars added to a data base.

This processing of incoming data is usually handled by a script or program written in Perl or another language that manipulates text, files, and information. If you cannot write a program or script for your incoming information, you need to find someone who can do this for you.

The forms themselves are not hard to code. They follow the same constructs as other HTML tags. What could be difficult is the program or script that takes the information submitted in a form and processes it. Because of the need for specialized scripts to handle the incoming form information, fill-out forms are not discussed in this primer.

ISP An ISP (Internet service provider) is a company that provides individuals and other companies access to the Internet and other related services such as Web site building and hosting. An ISP has the equipment and the telecommunication line access required to have points-of-presence on the Internet for the geographic area served. The larger ISPs have their own high-speed leased lines so that they are less dependent on the telecommunication providers and can provide better service to their customers. Among the largest national and regional ISPs are AT&T WorldNet, IBM Global Network, MCI, Netcom, UUNet, and PSINet.

They also include thousands of local providers. In addition, Internet users can also get access through online service providers (OSPs) such as America Online and Compuserve.

The larger ISPs interconnect with each other through MAEs (ISP switching centers run by MCI WorldCom) or similar centers. The arrangements they make to exchange traffic

are known as peering agreements. There are several very comprehensive lists of ISPs world-wide available on the Web.

An ISP is also sometimes referred to as an IAP (Internet access provider). ISP is sometimes used as an abbreviation for independent service provider to distinguish a service provider that is an independent, separate company from a telephone company.

URL A URL (Uniform Resource Locator) (pronounced YU-AHR-EHL or, in some quarters, UHRL) is the address of a file (resource) accessible on the Internet. The type of resource depends on the Internet application protocol. Using the World Wide Web's protocol, the Hypertext Transfer Protocol (HTTP) , the resource can be an HTML page (like the one you're reading), an image file, a program such as a CGI application or Java applet, or any other file supported by HTTP. The URL contains the name of the protocol required to access the resource, a domain name that identifies a specific computer on the Internet, and a hierarchical description of a file location on the computer.

On the Web (which uses the Hypertext Transfer Protocol), an example of a URL is:

http://www.mhrcc.org/kingston

which describes a Web page to be accessed with an HTTP (Web browser) application that is located on a computer named www.mhrcc.org. The specific file is in the directory named /kingston and is the default page in that directory (which, on this computer, happens to be named index.html).

An HTTP URL can be for any Web page, not just a home page, or any individual file. For example, this URL would bring you the whatis.com logo image:

http://whatis.com/whatisAnim2.gif

A URL for a program such as a forms-handling CGI script written in Perl might look like this:

http://whatis.com/cgi-bin/comments.pl

A URL for a file meant to be downloaded would require that the "ftp" protocol be specified like this one:

ftp://www.somecompany.com/whitepapers/widgets.ps

A URL is a type of URI (Uniform Resource Identifier).

VPN A virtual private network (VPN) is a private data network that makes use of the public telecommunication infrastructure, maintaining privacy through the use of a tunneling protocol and security procedures. A virtual private network can be contrasted with a system of owned or leased lines that can only be used by one company. The idea of the VPN is to give the company the same capabilities at much lower cost by using the shared public infrastructure rather than a private one. Phone companies have provided secure shared resources for voice messages. A virtual private network makes it possible

to have the same secure sharing of public resources for data. Companies today are looking at using a private virtual network for both extranets and wide-area intranets.

Using a virtual private network involves encrypting data before sending it through the public network and decrypting it at the receiving end. An additional level of security involves encrypting not only the data but also the originating and receiving network addresses. Microsoft, 3Com, and several other companies have proposed a standard protocol, the Point-to-Point Tunneling Protocol (PPTP) and Microsoft has built the protocol into its Windows NT server. VPN software such as Microsoft's PPTP support as well as security software would usually be installed on a company's firewall server.