

January 29, 2002

MEMORANDUM TO: Licensing Support Network Advisory Review Panel (LSNARP) Members

FROM: Dr. Andrew Bates /RA/
Chairman, LSNARP

SUBJECT: PROGRESS ON THE LICENSING SUPPORT NETWORK (LSN)
WEBSITE DEVELOPMENT AND RESPONSE TO QUESTIONS
RAISED AT AUGUST 2001 LSNARP MEETINGS

LSN Website Progress

The Nuclear Regulatory Commission (NRC) is pleased to announce that software release 1.0 of the LSN website was accepted on October 18, 2001. As a result of the government's acceptance testing, a number of items have been identified for inclusion in subsequent "patch" releases planned for FY 2002.

The LSN site was initially populated with a test document collection - - approximately 100 documents that had been replicated 10 times - - subsequently removed for security review. A smaller subset will be used for testing on the planned "patch" release. In February 2002 we expect to connect more than 17,000 NRC documents and an initial collection of Department of Energy (DOE) materials as soon as the new release has been tested, the test collection has been flushed from the database, and the participant organizations complete security review of their documents. Until then, however, the website is available for your use and we urge you to take the opportunity to become familiar with its operation.

The website contains a mechanism to submit comments and suggestions directly to the LSN Administrator (LSNA) by use of the "webmaster" e-mail function. If you prefer, you may contact us via phone (301-415-7401), mail, or fax (301-415-5599).

Response to Questions Raised at LSNARP and Technical Working Group (TWG) Meetings

DOE submitted a number of questions in writing at the LSN Advisory Review Panel (LSNARP) meeting held this past August, primarily as a result of its review of the June 8, 2001 version of the LSN Guidelines for Technical Implementation. Additionally, a number of questions were raised in subsequent technical working group meetings and by representatives of the Nuclear Energy Institute. The questions received, and the LSNA's answers and clarifications are provided below. In a few instances, the resolution recommended will involve updates to the LSN Guidelines previously provided to the LSNARP membership. These updates will be forwarded to you under separate cover in the near future.

DOE COMMENTS and QUESTIONS ON LSN GUIDELINES

1. **Is the June 8, 2001 version of the document being provided for review and comment?**

- **If so, when are comments due back to NRC?**
- **If so, who should comments be submitted to and via what medium?**

Response: The LSN Guidelines are operative guidance. They have been provided to the LSNARP to provide guidance and elicit feedback. The LSNA will make revisions as necessary to respond to specific issues or concerns that the participants have with their technical solution. Comments can be sent anytime in connection with any guidance material. E-mail communication is acceptable. Individual guidelines can be updated, as needed.

2. **Page 2-3, 3rd bullet refers to “Electronic Hearing Docket (EHD) and Electronic Information Exchange (EIE) components of the HLW repository proceeding.” How is the OCRWM/LSN expected to interface with EHD and EIE?**

Response: Instructions for use of the EHD and EIE will be provided on the NRC’s homepage and also via the LSN homepage. The OCRWM/LSN (e.g., the DOE document collection server) does not interface with either of these systems.

3. **Page 2-3, 5th bullet states: “Post notices on the central LSN site that contain listings of changes, if any, to each participant’s collection, identified by LSN accession number, with a description of what the change was and why it was necessary.” This is stated as a responsibility of the LSNA.**

- **How will the LSNA get this information?**
- **How will the LSNA post this information to the central LSN site?**

Response: Participants should notify the LSNA via e-mail of items that change (in accordance with Guideline 14). The LSNA will place that data in an area of the LSN website identified to notify other participants. LSNA staff will post the data on a tabular report.

4. **Page 2-3, 2nd bullet in 2.5 LSN Participants, states: “Designate technical points of contact for various functions ...” This paragraph requires each participant site to identify three (3) technical Points-of-Contact.**

- **When is the information needed?**

Response: As soon as available, but no later than two weeks prior to the participant’s collection being operationally connected to the LSN.

- **What information is needed about the technical P-O-C’s?**

Response: An e-mail message from the participant's Point of Contact will suffice. We need enough information to allow a system user to reach the point of contact via either e-mail, telephone, fax, or regular mail.

5. **Page 2-3, 7th bullet in 2.5 LSN Participants, states: "Make available (for inspection and copying) any document not provided in electronic form within five business days after it has been requested." Should that read: "Make available (for inspection and copying) any document not provided in electronic form within five business days after it has been directed by the administrative judge."?**

Response: Yes, the guideline will be changed. We propose changing it to: "Make available (for inspection and copying) any document not provided in electronic form within five days after directed by the Pre-License Application Presiding Officer." Note that, upon direction from NRC's Office of General Counsel, "five business days" has been changed to "five days."

6. **Page 2-3, last bullet, and states: "Comply with all standards for presentation of documentary materials established by the LSNA." Where are the standards documented?**

Response: In 10 C.F.R. Part 2, Subpart J, and in the LSN Baseline Design Requirements in Release 1.0, June 5, 2001.

7. **Page 3-2, 4th paragraph, last sentence states: "Autonomy log files provide statistics on bibliographic headers, documents, and image files add/change/delete transactions as it "crawls" each participant's website at the byte level." Please explain what is meant by "...at the byte level."**

Response: Byte level simply refers to the fact that we will be ensuring integrity of the documents (i.e., they have not been changed without the LSNA's or LSN users' knowledge) by comparing the digital "fingerprint" of participants' documents for changes. Specifically, we will be using MD5, a digital signature algorithm that is used to verify data integrity, to ensure each document has not changed since it was placed on the participant's server.

8. **Page 3-2, 6th paragraph, 2nd and 3rd sentences state: "Participants will have a user ID and password to login on the LSN web portal. When a participant logs in they will be switched to a specific path into the network that will provide processing on a dedicated server." Please clarify. Does the NRC intend to have two parallel LSN access servers?**

Response: Yes.

9. **Page 3-2, last paragraph, discusses "...six primary stores of information." However, the paragraph does not contain any reference to the actual documentary material repositories. Please add a sentence or two that explains specifically where the documentary material repositories actually reside. In**

reality, there are many more than just the "... six primary stores of information" that are part of the Major LSN Repositories.

Response: The LSNA does not believe a change is necessary. Section 3.1, Technical Focus, states that "the LSN ... is not a central repository ... documents from participant sites are not stored on the LSN web portal ..." Detailed information is in the LSN design documents provided to the participants.

10. **Page 3-4, 4th paragraph discusses participant site audits. How will the audit function work for pages and/or documents that are replaced in their entirety because the original image(s) have become corrupted?**

Response: If original images(s), pages, and/or documents become corrupted on a participant server, that is not a reason to replace the document in its entirety. Rather, the participant should use routine database recovery procedures to restore a corrupted database.

11. **Page 4-1, 4. SYSTEM AVAILABILITY, states: "The NRC has determined that a day-for-day extension of the hearing period applies when either the central LSN or the server hosting the licensing hearing docket is unavailable to participants during established access periods." Both of these functions are NRC functions. What processes are planned to be implemented by the NRC that will prevent "downtime" from exceeding the four-hour time period for these servers, thereby extending the time specified (three years from submission of the DOE license application) for issuance of a final decision approving or disapproving issuance of the construction authorization?**

Response: Information about backup and recovery of the LSN is in the design documents provided to the participants; in general, however, it includes the service level agreements we have through our contract with GRCI and AT&T for database availability, which is reflected in the architecture for both the servers and providing web hosting and communications services. The EHD service agreement between NRC's Office of the Secretary (SECY) and Office of the Chief Information Officer (OCIO) are being established at this time and that level of service is an identified performance requirement.

12. **Page 4-1, 4.1 LSN Search and Retrieval Availability, last sentence states: "...users may notice performance degradation during the daily planned backup time between the hours of midnight and 6 a.m. Eastern Standard Time (EST)." That translates to 9 p.m. to 3 a.m. Pacific Standard Time (PST). It is expected that some of the LSN users will intend to work after 9 p.m. PST. Is this a concern?**

Response: Yes, that is why we are providing notification in advance. We hope that participants will adjust their work schedules in recognition of the potential for performance being somewhat slower. The LSN staff will continue to work with the participants to accommodate server maintenance schedules so that downtime is minimized, and potential system administration-related performance degradations are disclosed to the parties so they can adjust their work plans accordingly.

13. **Page 4-1, 4.1 LSN Search and Retrieval Availability, 2nd paragraph, discusses the fact that one or more of the participant LSN websites may be down and unavailable. Please add a sentence that addresses participant LSN website maintenance.**

Response: We have asked that the participants coordinate with us when their website maintenance occurs. As soon as we have that information, we can put in a sentence addressing what impacts, if any, should be expected. Participants should communicate their maintenance requirements any time during the connection process, but not later than two weeks before they make their “live” collections available. This will allow LSN staff to post scheduling information on the LSN web pages covering LSN schedule and system availability.

14. **Page 5-2, 2nd paragraph, states: “If an electronic image is not associated with the bibliographic header, the Comments field in the bibliographic header should be used to document where an image version of a document may be acquired.” OCRWM plans to insert an image of the Special Instruction Sheet (SIS) that would contain the required information.**

Response: This is an acceptable approach, as there is an image associated with the bibliographic header; it is the image of the SIS (rather than the image of the record). We’re concerned with bibliographic headers that have NO image immediately accessible for them, such as an image that requires proprietary software or specialized viewers.

15. **Page 7-1, 7.1 Public Access to the LSN, 3rd sentence states: “Computers located at the NRC and DOE Public Document Rooms will allow free access to the central LSN site for the public, no later than six months prior to DOE’s submitting the license application.” It was DOE’s understanding the public libraries could be substituted for this requirement. Why the duplication?**

Response: The requirement in 10 C.F.R. 2.1007(a)(1) states that “a system to provide electronic access to the Licensing Support Network shall be provided at the headquarters of DOE, and at all DOE Local Public Document Rooms established in the vicinity of the likely candidate site for a geologic repository, beginning in the pre-license application phase.” A computer with access to LSNNET.gov must be provided in all DOE Public Document Rooms (PDR). Having the LSN accessible via the Nevada public library systems was proposed by an LSNARP representative to be investigated as a means to augment public access that is to be provided by PDRs, but would not be in lieu of what the rule requires.

16. **Page 7-1, 7.1 Public Access to the LSN, 4th paragraph states: “Internet users will have the ability to receive documentary material from the participants’ LSN websites without utilizing a proxy from the LSN server. This connection will be able to provide reasonable responsiveness during periods of normal usage.” Implication is that the OCRWM/LSN file server will be a stand-alone web server. That is not true. Please clarify that the documents are only served from the**

OCRWM/LSN file server based on a request from an internet user for a specific reference.

Response: We don't agree that it is reasonable to imply that the OCRWM/LSN or any other participants' servers are standalone web servers given the design documentation provided, the extensive briefings to all the participants on the architecture, and the ongoing interactions we have had since January 2001; therefore, we don't see any need for any additional explanation.

17. **Page 7-1, 7.2 General Users' Limited Access to the LSN, 2nd paragraph discusses the "load-balancing" intended to be done by the NRC at the central LSN site (See page 3-2, 3.3. Hardware and Software Configuration, last paragraph in that section.)**

- **Are there any requirements for the participant LSN sites to support this "load-balancing" activity?**
- **If so, what are the requirements?**

Response: No requirements are levied on participants as a result of the LSN load balancing capability.

18. **Page 8-1, 8.2 Early Submission for Testing, 3rd sentence states: "A significant percentage of each participant's document collection should be made available well before the end of this period so the LSNA can size the system, evaluate system performance characteristics and connectivity, and extrapolate the results into a meaningful performance model." DOE intends to comply with the NRC requirement to provide documentary material during the pre-license application phase.**

- **Please define "A significant percentage of each participant's document collection ..."**

Response: In the case of DOE, this would include at a minimum, the "known" document set currently available on the Web with documents added as progress is made toward LA. We recognize that this is probably the best we can do until the participants know how many documents they will have.

- **Please define "... well before the end of this period ..."**

Response: Pre-License Application Phase, far enough in advance of Certification to ensure successful "Load Testing," etc.

19. **Page 13-1, 13. SUBMITTING TEXT FILES OR IMAGE FILES, 7th paragraph states: "The URL of the electronic file(s) and electronic image(s), as applicable, must be entered into the bibliographic header for all documentary material. The electronic file URL(s) must be entered into the Text URL field, and the electronic image**

URL(s) must be entered into the Image URL field.” DOE intends to place the URL in the XML file, not in the bibliographic header.

Response: That is an acceptable solution, recognizing that the XML file is a file of the bibliographic header data. There are other solutions in which a participant may not be using XML and the general guidance still holds.

20. **Page 14-1, 14. CORRECTED OR REVISED DOCUMENTS, 2nd paragraph, 3rd bullet states: “Other parties or potential parties are notified of the change.” Does this fall under the purview of the LSNA to notify the parties or potential parties?**

Response: Change Notices will be posted on LSN Web Page by LSNA. See response to Question #3 above.

21. **Page 14-2, 1st paragraph, 3rd sentence states: “The notice on the central LSN site will contain listings of changes, if any, to each participant’s collection, identified by LSN accession number, with a description of what the change was, the date of the change, and why it was necessary.”**

- **How is this information going to be collected?**
- **Why does the nature and reason of the change (... description of what the change was ...) and “... why it was necessary”) have to be provided?**
- **How does the NRC intend to display the reason for the change (the Why)?**

Response: It is collected via e-mail notification from the participant to the LSNA. 10 C.F.R. 2.10011(c) requires the LSNA to “[i]dentify any problems regarding the integrity of documentary material certified in accordance with Sec. 2.1009(b) by the participants to be in the LSN” The information is collected so the LSNA has some documentation on why any changes occurred should it ever be necessary for a party to bring an issue before the Pre-License Application Presiding Officer. The reason for the change is included on the table identifying each change. The mode of presentation will initially be a table format; however, should this prove cumbersome, the LSNA will consider alternative implementation methods, such as a picklist that would be included in a later release of the LSN.

22. **Page 14-2, 2nd bullet, 3rd sentence states: “The participant also must notify the LSNA of the specific change(s), including the LSN Accession Number, ...” If the LSN Accession Number is assigned when the site is “crawled” how can the participant “... notify the LSNA ... including the LSN Accession Number ...?” (See Page 18-5, 18.4 LSN Accession Number Assignment, 2nd sentence.)**

Response: The participant posts a document, then the LSN crawls it and assigns that document a unique LSN accession number at the time of the crawl. Whether it is six hours, six days, or six months later, at some point in time, some action is requested by the participant. The document has always had a participant accession number, and has had an LSN accession number for six hours/days/months. If a transaction is to occur, the participant should use both numbers so that there is no question on identification of

the item. If a document is changed prior to “crawling,” the participant can change the document without notifying the LSNA; nobody will have had access to the document. If a document is changed or removed after “crawling,” then there will be an LSN Accession number for that document, and the participant must notify the LSNA of the change, including the LSN Accession Number assigned during the “crawl.” Bibliographic header fields are copyable using standard Windows commands to facilitate identification.

23. **Page 16-6, 2nd paragraph, 2nd sentence states: “It is expected that the Presiding Officer(s) will require that all pages included in the offered exhibit (with exceptions for special handling items) must be in a single PDF file.”**

- **DOE’s images are stored in TIFF file format. Who is to do the conversion to “... single PDF file?”**

Response: The party intending to submit the document as an exhibit will be responsible for meeting docketing requirements associated with their proffered exhibits.

- **When is the conversion expected to be accomplished?**
- **What additional information is to be provided during the conversion process?**
- **What are NRC’s expectations on how PDF images are going to be certified that they match the TIFF images?**

Response: The party requesting admission of a document into the proceeding and, therefore, into the Electronic Hearing Docket (EHD) is responsible for the conversion of that document prior to submitting it. The rule requires that a party posting a document in the LSN must be able to direct a requesting party to where an image of the document may be obtained. The rule does not specify either the media (paper, film, digital) or, if digital, the file format (.tif, .pdf, etc.) that the originating party may use in delivering the “image” to the requesting party. When one party requests an image version of another party’s document, the requestee has no way of knowing whether the requestor will be submitting the document to the docket. The document goes to the requestor, and it is the requestor that is responsible for submitting the materials to the EHD that support its case (no matter who originated the document), not the requestee. The judges will request a bibliographic header (in LSN format) to accompany each document. See the response to Question #37 regarding what steps must be taken to meet the criteria for an authenticated image.

24. **Page 19-1, 19.1.1 Request for Hearing and Designation of Licensing Board(s), 2nd paragraph states: “For the Yucca Mountain case, DOE’s license application must be filed in electronic format capable of being uploaded into the Agency’s EHD, which will reside in ADAMS.”**

- **This requires the LA documentation to be converted to PDF format. Who is to do the conversion?**
- **ADAMS has major data handling problems. Where are the technical specifications of what ADAMS will and will not accept?**

Response: The license applicant is required to submit its license application in the format specified. NRC is in the process of developing the LA guidance that will be provided in writing to DOE.

25. **Page 19-1, 19.1.3 Pleadings, 2nd paragraph states: “For the Yucca Mountain case, pleadings are to be submitted electronically via the NRC’s EIE process.” Are the technical specifications of what the NRC’s EIE process will and will not accept different from the technical requirements for ADAMS?**

Response: An adjudicatory EIE pilot is being developed and it will be internally consistent with ADAMS’s file format handling capabilities.

26. **Page 19-3, 19.1.7 Witness Lists, Pre-File Testimony, and Exhibits, 2nd paragraph states: “For the Yucca Mountain case, pre-filed testimony and exhibits, unless they are classified, safeguards, physical things, etc., must be filed in electronic format capable of being uploaded into the EHD, which will reside in ADAMS.”**

- **This requires this documentation to be converted to PDF format. Who is to do the conversion?**
- **ADAMS has major data handling problems. Where are the technical specifications of what ADAMS will and will not accept?**

Response: Whoever is submitting the documentation to the EHD is responsible for the conversion of that document prior to filing it.

27. **Page 20-2, 1st paragraph, 2nd sentence states: “Therefore, when a participant submits a document from its LSN website to the NRC docket, the participant must ensure that all referenced supporting documentation is properly assembled as a record package (in a parent/child context) and submitted together.” Please define what is meant as a “record package.”**

Response: See the description for package identifier in the baselined design requirements. NRC uses the same concept of “record package” as OCRWM. As specified in section 2.1003, participants are responsible for incorporating all their “documentary material” that meets the requirements defined in section 2.1001, including material that is relevant to, but does not support, their positions in the high-level waste repository proceeding, and any report or studies prepared by or on behalf of a party relevant to the license application or the Topical Guideline issues in Regulatory Guide 3.69, regardless of whether they are cited and/or relied upon by a party. How a party translates its internal record keeping structure into published web-accessible collections that meet the requirements of the rule is left to the party’s discretion. If a party chooses to maintain the package concept that exists in their internal records system, thereby ensuring “that all referenced supporting documentation is properly assembled as a record package and submitted together,” this undoubtedly will help LSN users to understand and navigate in that documentary context.

28. Page 20-2, 1st paragraph, 3rd sentence states: “Relationships between the documents being submitted to the docket must utilize a cross-reference capability contained in the bibliographic header.”

- **Please define what is meant by “...cross-reference capability contained in the bibliographic header.”**
- **Who is going to provide the “...cross reference capability contained in the bibliographic header?”**
- **Where, in the rule, is the basis for this requirement?**

Response: Regarding cross-reference capability, see the descriptions for related record number and related record code in the baselined design requirements. In this regard, the LSN rule gives the LSNARP the responsibility to provide advice on “the format standards for providing electronic access to documentary materials ...” Because these fields are in the bibliographic header fields as originally developed by the predecessor Licensing Support System (LSS) Advisory Review Panel and are reflected in subsequent iterations for use in a web-based LSN, this guidance is simply following advice from the LSNARP. For example, the LSN test collection contains the following bibliographic entry that exemplifies the use of cross-references in bibliographic fields:

LSN Accession # :	DEN000000944	
Organization:	Department of Energy	
Participant Accession #:	MXJ.20001102.0067	
Title:	SATURATED ZONE FLOW AND TRANSPORT PROCESS MODEL REPORT, TDR-NBS-HS-000001, REVISION 00, ICN 02 - APPROVED (C)	
Document Date:	11/02/2000	
Comments:	THIS IS A ONE-OF-A-KIND COLOR GRAPHIC DOCUMENT WHICH CAN BE LOCATED THRU THE RECORDS PROCESSING CENTER / PER SOURCE: ABOVE DATE IS THE EFFECTIVE DATE OF THIS DOCUMENT, NOVEMBER 2, 2000	
Non-Digital Media:		
QA Record Indicator:	Y	
# Of Images:	321	
Descriptors:		
Access Controls:	PUB	
Addressees:	N/A	DOE
Authors:	EDDEBBARH AA	N/A
	BIGGAR N	N/A
	KELLEY MJ	N/A
	DIXON PR	N/A
	SNELL RD	N/A
Document Numbers:	TDR-NBS-HS-000001	
Document Types:	REPORT	
	PMR	
Package Ids:	MOY-001208-23-06	
	MOY-001220-05-01	
	MOY-001215-39	
	MOY-010117-04-01	
Related Records:	MXJ.19991118.0188	PMR-REF
	MXJ.20000329.1187	PMR-REF
	MXJ.20000622.0363	PMR-REF
	MXJ.20000824.0513	PMR-REF
	MXJ.20000602.0052	PMR-REF
	MXJ.20000609.0266	PMR-REF
	MXJ.20000918.0287	PMR-REF

	MXJ.20000825.0122	PMR-REF
	MXJ.20000526.0328	PMR-REF
	MXJ.20000526.0330	PMR-REF
	MXJ.20000526.0338	PMR-REF
	MXJ.20000817.0546	PMR-REF
	MXJ.20000802.0010	PMR-REF
	MXJ.20000830.0340	PMR-REF
	MXJ.20000629.0907	PMR-REF
	MXJ.20001204.0064	XREF-BY
	MXJ.20001215.0185	XREF-BY
	MXJ.20010104.0043	ATT-TO
	MXJ.20010104.0046	ATT-TO
	MXJ.20010123.0123	AMR
Traceabilities:	TDR-NBS-HS-000001	
	DC 26712	
	YDAR 23499	
	SLP60DM3	
	TCR T2001-0009	
Versions:	REVISION 00	
	ICN 02	

Relative to the EHD, NRC's SECY organization defines the requirements for docket submissions. Whoever submits the documents to the EHD is responsible for providing the cross references.

29. **Page 21-1, 3rd full paragraph states: "DOE, however, has indicated that this prohibition does not preclude the so-called "affected units of local government" (AULGs)(i.e., Yucca Mountain-area counties) from using DOE-provided financial assistance to achieve compliance with LSN requirements." The 4th sentence of that paragraph then states: "Although the DOE representative did not directly address the status of "affected Indian tribes" (i.e., those Native American tribes whose reservation encompasses, or whose possessory or usage rights to other land would be substantially and adversely affected by, the repository), the same interpretation may well apply because those entities obtain funding under similarly-worded NWSA section 118(b), 42 U.S.C., Section 10138(b)."**

Since the decision on whether to designate an Indian tribe as "affected" or not falls under the purview of the Department of Interior, DOE suggests the following comment be added as the 5th sentence in that paragraph:

"At this time, the Secretary of Interior has not designated any Indian tribes in the vicinity of the Yucca Mountain Site as 'affected.'"

(See NWSA definitions section for "affected Indian tribe.")

Response: We will add that sentence.

30. **Page B-1, 5th paragraph, 3rd and 4th sentences state: "Having most documents submitted in the native word processing formats or in HTML rendition output from the authoring package (by definition, 100% accurate, and at least 99.95% accurate after transmission) should compensate for portions of the collection that may**

have an accuracy rate as low as 95% accuracy (unedited OCR output) within a document. The blend of clean native text and some non-edited ASCII should result in meeting a target of 98.5% for a collection as a whole.” That is a true statement if in fact the documents that DOE has in its document collection were just being created or, were created just within the past several years or, were created using a single word processor or, were from a single or very small group of authors. However, that is NOT the nature of the DOE document universe. Utilization of native file format is being considered for future submissions and DOE plans to address an infrastructure that would promote utilization of native files if a cost/benefit analysis indicates an acceptable ROI. However, the document sets created under any new procedure or file format requirement would be a very small percentage of the overall DOE document collection. (Native file format has several drawbacks that must be resolved before they can be effectively utilized, i.e., and pagination issues between native file and images.)

Response: 10 C.F.R. Part 2, Subpart J has been in effect since 1989. OCR accuracy concerns were documented by DOE in the 1988-1989 Conceptual Design Studies for the LSS. In the early and mid-1990s, DOE supported major research activities into technical solutions for improving OCR accuracy and in developing a document management system, subsequently not implemented, that would store native word processing files. As a consequence, we would anticipate that it is in a position to meet the 98.5% accuracy target. Moreover, we would note that in its current version, 10 C.F.R. Part 2, Subpart J, requires only text versions of documents and does not require synchronization of text and image versions of documents online.

31. **Page B-4, 2nd paragraph discusses Shannon’s communication theory, testing of the Autonomy software against data streams generated by OCR software, “dirty data,” etc.**

One of the strengths of a concept-based engine is its ability to overcome OCR (or spelling) errors. This is because a concept gets discovered by the occurrence of related words in a given document. Consider the following example from Autonomy’s technical white paper:

“I was walking down the street. The street was dark ... There were no other people in the street ... the street ... I was attacked by a mugger”

According to Autonomy, a word-based retrieval engine will characterize this document as an article about a “street” while Autonomy’s engine based on Shannon’s and Bayesian theory will represent this document as a document about “crime.” The concept “crime” is inferred from several words contained in the document such as “mugger,” “attack,” etc. It is this set of words (or bit strings) which leads to the concept “crime.” A concept-based retrieval engine will not define a concept to be a “misrecognized word” that occurs infrequently.

- **Has the NRC run a test against DOE’s test document collection to determine the impact on search and retrieval?**

- If in fact Autonomy is “concept based” and “... is not a classically structured text retrieval package (i.e., it does not build a word list index of pointers to text occurrence locations), but rather relies upon Bayesian probability theory and Shannon’s communication theory ...” the introduction of “... large quantities of misread characters, ...” should have absolutely no impact on search and retrieval since the Autonomy search engine is NOT doing a word-for-word match. In fact, auto-zoning should improve search and retrieval since text contained in tables, columns, etc., will not be “zoned out” of the document before the OCR conversion takes place.

Response: NRC requested, and DOE has committed to providing, approximately 50 clean test documents and the same 50 documents in “dirty OCR” format. We will begin comparison testing as soon as we have that document set available (estimated for the first quarter of CY2002) to evaluate impacts on search and retrieval. We plan to examine not just precision and recall, but also relevancy ranking and proximity searching.

32. Page B-5, 1st paragraph, 1st and 2nd sentences state: “The issue for the LSN design team, therefore, is to encourage non-unique but important words to naturally rise to the surface in relevancy ranking. That is to say that we want to minimize the potential for OCR’d but non-edited data (full of spurious characters) to cause truly relevant documents from “sinking from the surface” in relevancy ranking because of the occurrence of so many unique characters/words (e.g., the misread characters) that are more “interesting” to Shannon’s model.”

The 4th sentence goes on to state: “In effect, the one-time-occurring error strings are “unique” and, per Shannon, therefore more valuable than a term occurring repeatedly in the document.”

There are several term frequencies used to determine term-to-document relevance. In particular, there is within document frequency which counts how many times a word occurs in a document and the second is collection frequency which counts how many documents in the collection contains a certain word. It is the combination of these two frequencies that lead to the weight (or importance) assigned to a term for any particular document. We believe this is the way Shannon’s theory is applied in Autonomy to determine term importance. A misrecognized word with a low within document frequency plays no role in concept identification.

Document length normalization is another factor that may play a role. In the past, retrieval engines used to take the number of unique index terms or maximum term frequency to normalize the length of a document.

ISRI found that this kind of normalization can cause problems in ranking in some extreme cases. Modern engines now apply actual document byte size to

normalize document length in order to avoid problems with the errors caused by OCR, speech recognition, and other recognition systems that may generate noise.

- **Post-processing software has been acquired from UNLV's ISRI that will allow the removal of a majority of the "dirty data" that it is believed the LSN design team is alluding to.**

Response: Given our concern about the impact of "dirty data" on search and retrieval from participant collections generally, we provided this discussion as background on the rationale for our OCR guidance. We will modify it if tests show that dirty data does not affect relevancy ranking and like search results. In any event, we are still concerned that dirty data constrains reuse or citation of text using cut-and-paste technologies.

- 33. Page B-5, B.3 Analysis, 1st paragraph, 1st sentence states: "The LSN will be implemented using software that might be sensitive to large volumes of inaccurate data." Again, if Autonomy is "concept based" and "...is not a classically structured text retrieval package (i.e., it does not build a word list index of pointers to text occurrence locations), but rather relies upon Bayesian probability theory and Shannon's communication theory ..." the introduction of "... large quantities of misread characters, ..." should have absolutely no impact on search and retrieval since the Autonomy search engine is NOT doing a word-for-word match.**

Response: As we noted in the response to Question #32, we will modify the guidance discussion if tests show that dirty data does not affect relevancy ranking and like search results.

- 34. Page B-5, last bullet states: "Eliminate, as much as possible, the occurrence of "dirty data" and adhere to the 98.5% overall accuracy standard for a participant's entire collection rather than lowering the standard to 95% as was originally proposed in the April 2000 draft of the LSN Functional Requirements." The 98.5% overall accuracy rate is NOT consistent with the requirements of Page 15-1, 2nd bullet that states: "For OCRd documents, the target is 99.5% character accuracy, with a 98.5% character accuracy target for each individual page to ensure the overall consistency of each page." Please clarify.**

Response: The following is an example based on a ten-page document with a goal to achieve 99.5% character accuracy on the whole document. If nine pages are relatively pristine, but one page is significantly misread by the OCR engine and its accuracy is 14% this renders the document unacceptably "dirty" to the extent that the document, overall, fails to meet the 99.5% mark. The purpose of our discussion is to suggest that pages with high volumes of errors be recognized as having the potential to pull down the overall accuracy of the document. If no individual page is below a 98.5% accuracy rate, then the document sponsor stands a better chance of the document meeting the 99.5% standard. Moreover, this example suggests that fixing one egregious page is a better investment than churning through a high volume of pages already at 98.5% and trying to bring them up by 1%.

Applying that the same perspective across a whole collection of documents, if the target for the overall collection is 98.5% and there are a significant number of very large documents with very low OCR accuracy overall, it will take a lot of clean documents and pages to compensate. Alternatively, a party could have a large number of documents generated from native word processing and rendered directly into HTML that would result in big blocks of 99.99% accuracy, which might then “carry” some of the dirtier portions of the collection.

35. **Page B-5, B.4 Text Accuracy Evaluation, last sentence states: “Although an argument can be made that introduced characters detract nothing from the subset of all other characters and words properly identified and therefore should not be counted against achieving the accuracy target, it has not been demonstrated that spurious characters have no effect on proximity searches.”**

Proximity searching is based on the hypothesis that when words occur within a specific distance from each other repeatedly, then the document is relevant to the concept the searcher is looking for. Although it is plausible that an occurrence of a phrase may be missed due to OCR error, it is very unlikely that all occurrences are corrupt due to errors. Again, document concepts are not developed from unique or single occurrence of words or phrases. DOE is proposing tests be conducted that will in fact demonstrate or disprove the effect “dirty data” may have on proximity searches.

Response: The LSNA expects to conduct tests the first quarter of CY2002.

36. **Page B-7, last sentence states: “Until then, participants should assume that introduced, non-cleaned up spurious characters are counted.” Post-processing software has been acquired from UNLV’s ISRI that will allow the removal of a majority of the “dirty data” to minimize the impact on search and retrieval that the LSN design team is looking for.**

Response: Assuming this software works, we hope it can be made available to other LSN participants.

COMMENTS ARISING FROM LSNARP MEETING (Wednesday):

37. **During the LSNARP meeting, the term “authenticated image” was used in a context indicating that possibly the LSN would deliver both authenticated and non-authenticated images.**

- **Please define the term “authenticated image.”**

Response: The term “authenticated image,” or more precisely “authenticated image copy,” is used in 10 C.F.R. § 2.1003(a)(1). This provision states that at the time the NRC, DOE, and each potential party, interested governmental participant, or party, makes available an electronic file including a bibliographic header file for each item of documentary material it possesses, it also must have provided an authentication

statement that indicates where an authenticated image copy of the document can be obtained. Further in this regard, the term “image” is defined in section 2.1001 as “a visual likeness of a document, presented on a paper copy, microform, or a bitmap on optical or magnetic data.” The objective of the authentication provision in section 2.1003(a)(1) is to not only alert other LSN participants as to how they can obtain an actual image of a document but also to provide a potential check on whether the electronic copy of a document is actually what it purports to be, i.e., it accurately captures the content of the original document. Authenticity of a document will also be one factor in determining the admissibility of a document in any later adjudicatory proceeding on a repository license application. In that context, the term “authenticated image copy” denotes an image copy of an item of documentary material that, if presented in an evidentiary hearing in an unadulterated form, will not be subject to a sustainable objection based on a lack of “authenticity.” Put another way, the “authenticated image copy” should be one that, if the generating entity itself sought to introduce it into a judicial proceeding, would not be subject to exclusion as not being authentic.

For purposes of the application of section 2.1003(a)(1) to Federal government agencies, as well as other governmental units (states, counties, tribes), as long as the agency or governmental unit follows the applicable laws and regulations (for example, the National Archives and Records Administration regulations on the federal level) for maintaining official record copies of their documents, whether for paper copy or electronic copy, that official record should be considered the authentic image of the document. Where authentication becomes an issue, either before the Pre-License Application Presiding Officer or before the Presiding Officer at the adjudicatory hearing on the DOE license application, the sponsoring agency would need to authenticate the document. Satisfying the authentication requirement in that context will be based primarily on the document’s reliability which will be dependent on demonstrating that the requirements for maintaining the integrity of the document have been satisfied. Generally, under the Federal Rules of Evidence, public records are authenticated by proof of custody. Advisory Committee notes, Fed. R. Evid. 901 (1998). Furthermore, the Advisory Committee notes to Rule 901 of the Federal Rules of Evidence, indicate that the principle of authentication for public records has been extended to include data stored in computers.

It should be emphasized that the authenticated image may be a paper copy of a document. However, it could also be an electronic copy. As Federal agencies and other governmental units move toward replacing paper copies of documents with electronic copies for purposes of the “official agency record,” the following United States Department of Justice guidance regarding electronic records provides useful guidance:

As with evidence in any form, electronic records must meet the legal requirements for “admissibility” before the government can use them in court. Generally, the party seeking to introduce records must show that the evidence is “authentic” (that is, provides proof that it is what it purports to be), and that it is the “best evidence” (that is, that it is the original or an acceptable duplicate). In addition, in order to be able to use electronic or any other agency records as evidence, the government in most cases

must establish that the records were generated and maintained by a "trustworthy" process. If an electronic process does not reliably show who transmitted a piece of information to the agency, when it was transmitted, and that it has not been altered either intentionally or inadvertently, then, depending on the circumstances, the electronic record might not even be admissible, which means that the record could not even be considered by the judge or jury in deciding the merits of the government's claims or defenses.¹

United States Department of Justice, Legal Considerations in Designing and Implementing Electronic Processes: A Guide for Federal Agencies § II.C.4 (Nov. 2000) (<http://www.cybercrime.gov/eprocess.htm#IIC4>).

- **Please indicate “how” an LSN user would know whether a particular image is “authenticated” or not.**

Response: In accordance with Guideline 13, if a participant chooses to make a text document's corresponding images available on the web, those images must:

- meet the standards for an official record version if they are offered in response to meeting requirements for “notifying parties where an authenticated image may be acquired,” or
- identify individually for each image whether it is or is not an authenticated image.

Therefore, if an image is provided on the LSN by a participant, and the image is not identified as a non-authenticated image, the image would be considered authenticated as it has been downloaded from (and is verifiable against) the authenticated image available on the LSN.

38. It was stated that the LSN would support “2 million users.” Please provide the assumptions and basis supporting the statement.

Response: Two million is the theoretical limit of the number of users the three LSN web servers can handle; however the LSN system has been developed and tested to accommodate 150 concurrent users.

- **Is this a change in the performance specifications in the requirements document?**

Response: 150 concurrent users was in the original performance specifications.

¹ For example, Rule 803(6) of the Federal Rules of Evidence (FRE), generally allows records of regularly conducted activity (commonly called "business records") to be admitted into evidence. The government frequently relies upon that rule to introduce agency records into evidence. But the rule also provides that such record will not be admissible if "the source of information or the method or circumstances of preparation indicate lack of trustworthiness." Similar provisions are contained in, for example, FRE 803(8) (pertaining to admissibility of public records and reports) and 804(b)(3) (statements against interest).

- **What would you view to be the impact that the participant sites need to plan for?**

Response: Participants should plan for the volumes of documents, users, and file delivery as outlined in the baselined design requirements document.

39. **The NRC stated that the LSN site would be fully 508 compliant. It is understood that the NRC portal site can be developed to be fully 508 compliant, however:**

- **DOE has a very large document collection that currently is NOT 508 compliant. The vast majority of documents that DOE will post to the OCRWM/LSN will be posted using existing technology. DOE is currently assessing the impact of upgrading the existing Records Management System running on BasisPLUS to a new EDMS that would meet the 508 compliance requirement as part of the 508 Compliance Plan. However, full implementation and cutover to a new 508 compliant EDMS is well past the current planning horizon. DOE questions whether the technical distinction between the “NRC portal site” and the “LSN total system” is clear in the minds of the potential user community who hears such a statement. Care should be taken that potential users are not given to misunderstand that documents retrieved in response to any search and retrieval activity will be 508 compliant as well.**

Response: Federal participants are advised to consult with their Office of the General Counsel (OGC) regarding compliance with Americans with Disabilities Act section 508. We believe the participants have been adequately briefed and notified on the distinction between the LSN portal site and the total system, and we emphasized 508 compliance being applicable only to the Federal participants. In discussions with their individual OGCs, we suggest that the Federal participants emphasize the following points: The participant collection servers are not directly searchable and retrievable by a public user, e.g., contain no search and retrieval software per se. The LSN central server, on the other hand, does provide search and retrieval access to the public and is Section 508 compliant, although it provides no document storage. Federal participants should follow their own OGC guidance for the documents they make available on their document collection file servers.

OBSERVATIONS ARISING FROM LSN TWG MEETING (Tuesday):

40. **It is asserted that the daily information transfer of directory information from the OCRWM LSN site (assuming 500K documents) would be 3.5 hours. Please provide the assumptions and calculations that substantiate that assertion.**

Response: This calculation was based on known machine speeds, communication speeds, and assumptions made about document sizes and hypothetical file directory setups at that time. NRC continues to work directly with DOE to fine-tune the processes to be used in transferring data.

QUESTIONS FROM NEI IN TWG MEETING

- 41. If a document is classified or a rock [Ed. Note: the reference to a “rock” is to any material specimen, such as a core sample, water sample, etc.], is there guidance as to where, in the header for example, that they should put this information.**

Response: Classified materials should not be made accessible through the LSN since the site and the internet communications channel are not systems that provide an appropriate level of security for such documents. They are excluded from the LSN by 10 C.F.R. 2.1005.

LSN Guideline 5, page 5-2, section 5.3, 2nd paragraph states, "If an electronic image is not associated with the bibliographic header, the Comments field in the bibliographic header should be used to document where an image version of a document may be acquired." Similarly, a comment to the effect that the header is intended to denote a “physical specimen” of a particular type should suffice.

- 42. What happens if someone posts their LSN login id and password on a bulletin board and masses begin using it. Do we have some way to protect against or remedy this problem?**

Response: The LSNA has the ability to see how many times a user is logged in. Therefore, if a user is concurrently logged in an excessive number of times (e.g., 1,000), the LSNA can disable the account and contact the participant to discuss this anomaly. The LSNA also can raise such issues with the Presiding Officer, who may choose to discuss it directly with counsel to the party involved.

- 43. FOIA - How does the LSN relate to FOIA?**

Response: 10 C.F.R. 2.1007(b) states: “Public availability of paper and electronic copies of the records of NRC and DOE, as well as duplication fees, and fee waiver for those records, is governed by the regulations of the respective agencies.”

- 44. If someone wants a document posted (the 5 day rule) how do they initiate that (e.g., ask the LSNA, Pre-presiding officer, or ...)?**

Response: Assuming that the party that possesses the document has declined a request to post the document on the LSN, per the response to Question #5, ask the Pre-License Application Presiding Officer.

10 C.F.R. 2.1010(a)(1) states, “The Commission may designate one or more members of the Commission, or an atomic safety and licensing board, or a named officer who has been delegated final authority on the matter to serve as the Pre-License Application Presiding Officer to rule on disputes over the electronic availability of documents during the pre-license application phase ...”

10 C.F.R. 2.1010(b) states, "The Pre-License Application Presiding Officer shall rule on any claim of document withholding ..."

10 C.F.R. 2.1004 states, "Any document that has not been provided to other parties in electronic form must be identified in an electronic notice and made available for inspection and copying by the potential party, interested governmental participant, or party responsible for the submission of the document within five days after it has been requested unless some other time is approved by the Pre-License Application Presiding Officer or the Presiding Officer designated for the high-level waste proceeding. The time allowed under this paragraph will be stayed pending [Presiding] Officer action on a motion to extend the time."