LICENSING SUPPORT NETWORK - ADVISORY REVIEW PANEL
TECHNICAL WORKING GROUP
PARTICIPANT TECHNICAL DISCUSSIONS
AGENDA - January 2001

8:00 – 8:15 Introductions - Dan Graser/Matt Schmit
- NRC/GRCI Team
- Participants

8:15 – 09:15 Proposed LSN Solution
- Docket, Electronic Motions, & Electronic Courtroom - Dan Graser
- Search, Retrieve, and Audit Web Portal - Derek Weaver (GRCI/AT&T)
- Milestones and LSN Updates - Dan Graser

09:15 – 09:30 Break

09:30 – 10:15 Autonomy Discussion - Jeff McCormick (Autonomy)
- How Documents and Content Are Indexed
- Searching and Personalization
- What to Expect When Crawler Visits Your Site

10:15 – 10:30 Break

10:30 – 12:00 Participants Information Gathering Roundtable (Part 1) - Systems
(Participants bring completed questionnaire)
- Working POCs (Name, Phone, E-mail, etc.)
- Participant LSN Servers (O/S Version, Web Server Version, Dedicated, Availability)
- Participant LSN Users (O/S Version, Browser Version)
- Configuration of Participant Website (IP, URL, Directory Tree)
- Tracking a Site's Usage (Log File Format)
- Security / Firewall Access / Any Scheduled Maintenance Windows
- NRC's Compliance with Section 508 and Cookies

12:00 – 1:00 Lunch

1:00 – 4:15 Participants Information Gathering Roundtable (Part 2) - Integration Discussion
- Header Information (Logical View, Physically Stored, Format (ODBC, XML, etc.))
- Document Types (text, HTML, Word, WP, Non-electronic, etc.)
- Image Types (Tiff, PNG, PDF, etc.) and Image Format (page per file or multi-page)
- Participant Method for Adding Documents
- Audit; Integrity of Documents (Log files)
- Testing (Provide test data, Participant systems available for test)
- Participant Backup and Recovery
- Delivery of Document Images (on-line or off-line)
- Help (E.g., Phone Number) for Participant Server/document Problems

2:30 and 4:15 Break

4:30-5:00 User Session
- Priority Access and Priority Users
- User Interface
- Wrap-up

LICENSING SUPPORT NETWORK (LSN)
Technical Discussion Questionnaire

Participant: _____

Technical POC: _____ Office Phone & E-mail: _____

Business POC: _____ Office Phone & E-mail: _____

Other POC: _____ Office Phone & E-mail: _____

What is your LSN website address (IP and URL)?	
Estimate the number of files you may post.	
What is your directory tree structure for LSN Files: Header, text and image?	
Is your server going to be solely used for the LSN?	
What are the formats of the files you will be posting? (text, html, word, wp, etc.)	
What is the format of your headers? (xml, text, odbc, etc.). If ODBC, what is the schema for the header database and what make and version is the database? (Access, Oracle, etc.)	
What is the format of the images you store? (tiff, png, etc.)	
What operating system does your server use? (NT, UNIX, etc.) What version of the operating system?	
What web server software do you use? (Apache, IIS, etc.) What version is you web server software?	
What is the bandwidth from your site to the internet?	
Is your site behind a firewall? If yes, what is the ip / proxy?	
What is the maintenance schedule for your server?	
What is the format of your log files? (common log format, etc.)	
What is your user's operating systems? (Windows 95, 98, NT, etc.)	
What browsers and version is installed on your LSN user's machinet? (IE, Netscape, etc.)	
What is your document unique ID format?	

Licensing Support Network Advisory Review Panel

Technical Working Group
Participant Technical Discussions
January 2001

Introductions

- **NRC**
- **GRCI/AT&T**
- **Participants**

Proposed LSN Solution

- Document discovery via LSN portal
- Docket
- Electronic motions

Electronic courtroom (parallel project)

Development Schedule

February – design

March – development

April – open URL (static information)

April/May – revised rule in effect

LSN Resources

- Search, retrieve, and audit web portal
  - Participants provide file servers
  - Uses NRC infrastructure for docket and motions practice

LSN Search & Retrieval System Goals

- To allow participants and the public to easily search and retrieve relevant information pertaining to the high-level waste hearing
- To organize and uniquely identify documents for use in the hearing courtrooms
- To ensure document integrity

User Friendly Design

- Quick and intuitive navigation
- Clean web page design and layout
- Consistent organization and style
- Browser neutral
- Compliant with Section 508 of the Rehabilitation Act and government "cookies" policy

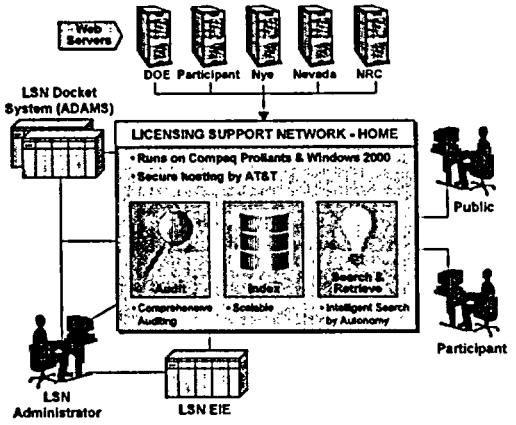
LSN System Overview

LSN portal:

- Windows, Compaq
- Autonomy

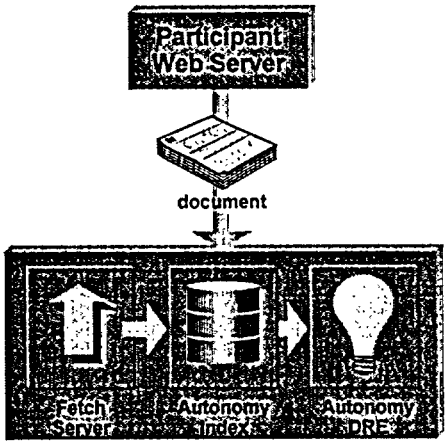
LSN audit:

- Web trends, WhatsUpGold



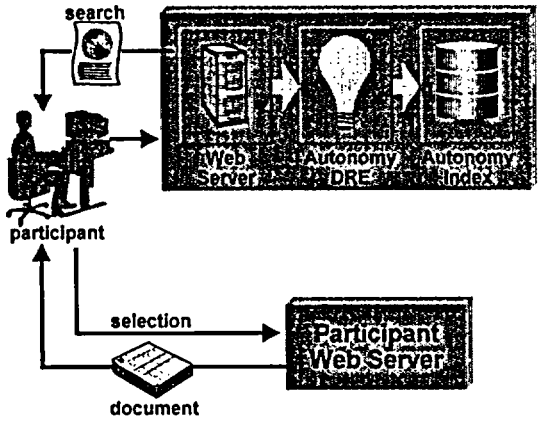
Indexing Participant Sites

- Crawl daily
- Distributed, parallel process
- Capture participant server performance
- Build intelligent concept index to support relevancy searching



Search & Retrieval

- State-of-the-art searching
- High availability
- Content retrieved from participant site and delivered to user



Autonomy Discussion

- How documents and content are indexed
- Searching and personalization
  - Boolean, phrase, proximity, results ranking, wildcard, etc.
- What to expect when the crawler visits your site
- How the portal and the participant file servers interact

LSN Site (AT&T Hosting)

- Security
  - Firewall, intrusion detection
- 99.4% availability
- Disaster recovery plan
- Load balancing
- Windows/Compaq platform

Integration Sessions

LSNARP TWG Meeting

Meeting Date: January 18, 2001

Location: Hillshire Building (DOE) Las Vegas, NV.

Those attending:

NAME	REPRESENTING
Daniel J. Graser	Nuclear Regulatory Commission
Matt Schmit	Nuclear Regulatory Commission
	Nuclear Regulatory Commission
Derek Weaver	GRC International (NRC Support)
Angela Tayfun	GRC International (NRC Support)
Jeff McCormick	Autonomy (NRC Software Vendor)
Dee W. Jensen	DOE OCRWM/OIM
Alesia Boone	DOE OCRWM/OIM
David Warriner	DOE OCRWM/OIM
Bob Brown	RSIS (DOE Support)
Lew Robertson	Booze-Allen, Hamilton (DOE Support)
Barbara McKinnon	Booze-Allen, Hamilton (DOE Support)
Tom Nartker	UNLV/ISRI
Jerry Weiser	M&O/SAIC
Roberta Stambaugh	M&O/MK
Harry Leake	M&O/TRW
Brad Stillman	M&O/TRW
Chris Ideker	M&O/SAIC
Marty Bryan	M&O/SEA
Robert Cray	M&O/TRW
Bill Schrock	M&O/TRW
Robert Stoner	Bechtel-SAIC
Paula Thompson	Bechtel-SAIC
Jeff Halliday	Bechtel-SAIC
Steve Cereghino	Bechtel-SAIC

The action items agreed upon during the meeting may be summarized as follows:

Action No.	Responsible Party	Action
1	Graser	Instead of text and image pages being fed from the OCRWM's LSN site directly to the requestor, DOE has requested the file(s) be provided to the requestor through the NRC server, i.e., the NRC's server becomes a "proxy server". The NRC agreed to the DOE proposed architecture.

2	Graser	The NRC will clarify the process to change pages posted to the OCRWM/LSN server with cleaned-up text should an error be discovered.
3	Graser	For documents where images are inserted, replaced, or deleted, new participant accession numbers should be submitted with the new document. NRC will review the associated requirement and provide clarification.
4	Jensen	What is the relationship between author and author organization in the DOE Records Management System? DOE is to review and provide definition.
5	Shepherd	Margie Shepherd (NRC) will look at the reasoning for keeping multi-participant accession number.
6	Jensen	DOE needs to look at the use of multi-value of version number for multiple drafts while retaining one accession number.
7	Graser	Change the value of Document type from '3' to '10'.
8	Graser	NRC to look at the number of values associated with "Access Control" information.
9	Jensen	The DOE will update the OCRWM/LSN on a weekly basis, or as agreed between NRC and DOE. The update process will not affect the required certification and re-certification processes.
10	Graser	NRC to check on bibliographic header field "number of images" where 'N' is defined for participant required.
11	Graser	The "physical location reference information" field is a candidate to be eliminated. NRC to review.
12	Jensen	The NRC is going to set up an LSN test bed of approximately 1,000 documents, to go on-line in April 2001. The NRC requested DOE participate in the design and implementation of the LSN test. OCRWM has agreed to participate with the NRC by bringing up a file server that will contain a sample set of documents.
13	Jensen	How are e-mail and attachments going to be represented in the LSN? DOE to research and present a proposed resolution.
14	Graser	A question to be posted to the LSNARP Chairman is: "Can we classify forms as image only?"
15	Jensen	DOE will provide metrics supporting the use of forms as "image only" to the NRC. Additional metrics will be provided relevant to the accuracy rate (percentage) of the image to text conversion of LSN documentary material.
16	Graser	NRC will work with the DOE on the file directory structure that will be used for the participant documents on the participant server.
17	Graser	The NRC will evaluate the compression algorithm that is currently being used by the DOE for color images, i.e., TIFF encoded JPEG images.

18	Jensen	DOE will develop and publish a formal OCRWM/LSN Backup and Disaster Recovery Plan that aligns with the NRC's LSN requirements.
19	Jensen	DOE will work with the NRC technical staff on designing the layout of the html page that will represent a document to the requestor(s). The design of the html page(s) used will include the distinction as to whether the header data will be submitted as a comma delimited file or as xml tags in the document html page itself.

The meeting minutes are recorded in the following four sections.

1.0 Introductions and Current LSN Status

The agenda for this meeting is attached. (Attachment 1). Dee Jensen (DOE OIM LSN Lead) opened the meeting and after all attendees introduced themselves, (Attendance sheet is Attachment 2), turned over the meeting to Dan Graser (NRC LSNA).

Matt Schmit, LSN Project Manager for the NRC, discussed the make up of the LSN development team under the newly signed NRC contract. The LSN development team consists of the following:

- GRCI – System integrator
- Autonomy – Portal software selected
- Commerce One – Web page development
- Management Systems Development (MSD) – Systems analysis and database development

The LSN is to be hosted at one of two AT&T sites in either Mesa, Arizona or in New Jersey.

Dan Graser then provided a report on the current status of the LSN. The meeting minutes from all three meetings (first one in northern Nevada, second one with DOE, and the third with other southern Nevada parties) will be provided to the LSNARP Chairman for distribution to all LSNARP members.

The NRC has budgeted for the development and implementation of the LSN along with a five-year plan to operate and maintain it. There are three parts identified in the Rule that make up the licensing process. They are the LSN, the Docket, and the EIE. The purpose of the LSN is to make available all documentary material for discovery purposes for all participants prior to and during the hearing process.

The Docket will be used during the licensing process to hold all exhibits, motions, and decisions. The Docket is to be initiated with the electronic filing of the LA to the NRC. The NRC is hoping to use the ADAMS system to support the Docket if it is operationally acceptable. Protective

order information, which consists of attorney/client privilege information, company private information, etc, will be locked in a safe and made available to the judges when the information is warranted. The documents entered into the Docket are required to be available to the LSN prior to their introduction to the Docket. One scenario was defined that would allow for information to be introduced into the court by having attorneys down load their evidence and exhibits from the LSN and then submit them to court from their laptops. Once introduced, the information would be moved into the Docket.

The Electronic Information Exchange will be used to link the motions and orders between the judges and attorneys.

Judge Ann Young is looking at the design and requirements of an electronic courtroom, including a server room. The analysis will require developing a Feasibility Study and performing a Cost/Benefits Analysis. The final solution will be determined over the next two years. Regardless if the electronic courtroom is used or not, it is expected that the DOE will be required to re-run computer models used to support the LA as a part of the Docket.

The LSN is in the design phase through February with a go/no go decision made before going onto the next step. In March, the development phase will see the software and hardware integrated and installed. Pilot testing will be conducted in April. The current plan calls for making the approximately 18,000 NRC documents available through the LSN sometime in July. Dee Jensen committed OCRWM to participate in the test phase.

Six sets of comments were received regarding the last draft Rule that was sent out for review by the NRC in August 2000. The NRC responses are being developed and reviewed by the Office of General Council (NRC) and expect to be released along with the Final Rule sometime in April 2001.

Matt Schmit described the LSN search and retrieval system goals. The LSN will allow participants and the public to search and retrieve relevant information pertaining to the high-level waste hearing. The participants will be able to use this information to prepare for the hearing process. The web interface will be clean and will allow for intuitive navigation.

Derek Weaver (GRCI LSN Development manager) provided a system overview of the LSN. Autonomy will be the portal software that crawls each participant set of documentary material. The index of the information crawled will be stored at the LSN home site which will be stored at an AT&T site (as previously mentioned). Intelligent Search by Autonomy will be used for the search and retrieval mechanism. To ensure data integrity, the LSN Administrator will perform audits of each participant's documentary material.

2.0 AUTONOMY PRESENTATION AND DISCUSSION

Jeff McCormick (Autonomy) made a presentation on Autonomy's capabilities and how it might be used in the LSN environment. Autonomy's strength lies in its high-performance pattern matching algorithms. These algorithms are informed by Claude Shannon's principles of information theory, Bayesian probabilities, and the latest research in neural networks. This

technique enables Autonomy's software to identify patterns in text and look for similar patterns in other sources, quickly and automatically.

Using the technology within Autonomy, the software can analyze a piece of text and identify the key concepts within the document because it understands how the frequency and relationships of terms correlate with meaning. Autonomy employs advanced pattern matching technology (non-linear adaptive digital signal processing) to extract a document's digital essence and determines the characteristics that give the text meaning. Once Autonomy's technology has identified and encoded the unique "signature" of the key concepts, Concept Agents are created to seek out similar ideas in other documents.

Autonomy Portal-in-a-Box is an out-of-the-box solution that enables companies to easily create and automatically maintain an Enterprise Information Portal which aggregates content from internal and external sources, eliminating the need for any manual labor in the process. It provides advanced personalized features that can allow users to access customized information available from any source of relevant, accurate and timely content needed to make informed business decisions. For Intranet developers, this means that they can set up an easy-to-navigate corporate portal that packages the best of the web and the company's Intranet. Autonomy can also profile users for "pushing information" to them.

There are several features that Autonomy technology allows for that will not be employed within the LSN. They include:

- The "non-push" technology that allows users to create and store personal agents for subsequent use,
- The interface customization that allows users to customize their interface and save it for future use, and
- Any other Autonomy features that requires a user to have an account and log in with a password.

It was also discussed that maybe the features that require account creation will be made available to the users that request it, but not required for general public internet access.

Autonomy will be used to spider the information stored at each of the participants web site location. The index of that information will be stored at the central LSN site. Instead of the text and image pages being fed from the participant site directly to the requestor upon performing a search and retrieval operation, DOE has requested that the requested information be fed to the requestor through an NRC proxy. This will limit the potential for security violations at the DOE site since only the NRC server would have access to the participant server.

Marty Bryan (M&O) asked a question concerning the relevancy and the order of documents returned from a user's query. It was made clear that Autonomy will present the results set in order of what it believes is important based on the uniqueness of the documents and the query submitted. However, the return order is not to be construed as relevant in regulatory sense. The NRC is contemplating using the Topical Guidelines as an organizer and may not show the

relevant percentage Autonomy assigns to users.

3.0 TABLE DISCUSSION

The following are topics that were discussed in a roundtable type discussion.

- The protocol for exchanging information between the DOE and the NRC was discussed. The DOE OIM point of contact is Dee Jensen. The DOE OLRC point of contact is April Gil. Harry Leake (M&O) and Jerry Weiser (M&O) were identified as the technical points of contacts. All information or conversation between the technical contacts and the NRC are to be provided to both April Gil and Dee Jensen.
- HTML Page - A question was raised to the NRC that the HTML pages that present text and images should look similar in appearance. Later in the meeting, Harry Leake (M&O) provided a demonstration of how the DOE's RISWeb produces the HTML dynamically. The DOE will work with the NRC technical staff on designing the layout of the html page that will represent a document. The design of the html page(s) used to represent a document will include whether the header data will be submitted as a comma delimited file, or as xml tags in the document html page.
- The Functional Requirements may be updated in the next couple of weeks based on the comments received and the information and discussion from the three TWG meetings.
- There was some discussion concerning the current DOE Internet collection, which is heavily hyper-linked. The version of documents that have been placed on the DOE Internet have had minor modifications and are not considered a copy of the record copy and should not be considered for inclusion in the LSN unless they are verified. For use in the Docket, the court will probably require each participant to provide the true agency copy certified by the agency records manager. In the case of DOE, the true copies are the TIFF images. Copies of these documents from the RMS will be used in the DOE tests.
- NRC will work with the DOE on the directory structure that will host the participant documents on the participant server. Since the NRC will placing their collection into the LSN first, the DOE needs to have the ability to follow suit. Consequently, by allowing DOE to participant in the directory design, DOE can identify and propose alternatives for any design issue that might negatively impact the ability to copy the NRC directory structure on the DOE server.
- Although it is not required to provide all of the images on the OCRWM LSN, DOE will provide them in order to satisfy all potential requests as stated in the Rule.
- The NRC will look at the compression algorithm that is currently being used by the DOE for color images - TIFF encoded JPEG images.
- When there are changes made to the bibliographic header, no new accession number is required from the participant.
- The NRC will review and determine if it is ok to change textual pages with cleaned-up text.
- For documents where images are inserted, replaced, or deleted, new participant accession

numbers should be submitted with the new document. The NRC will review the associated requirement and provide clarification.

- Concerning the removal of files, any files that are candidates for removal such as attorney/client, copyrighted, safeguards related records will be presented to the hearing Judge. Once the Judge agrees that the file should be removed, the file should be removed from the OCRWM LSN server and the LSNA will post a history of the LSN accession number and remove the header and text from the central LSN site as appropriate.
- What is the relationship between author and author org in the DOE RIS? DOE is to review this problem.
- Enter supercede and superceded by in bibliographic header for newer versions of a document.
- Margie Shepherd (NRC) will look at the reasoning for keeping multi-participant accession number.
- DOE needs to look at the use of multi-value of version number for multiple drafts while retaining one accession number.
- NRC to change the value of Document Type from '3' to '10'.
- NRC to look at the number of values associated with "Access Control" information.
- NRC to check on bibliographic header field "number of images" where 'N' is defined for participant required.
- The "physical location reference information" field is a candidate to be eliminated. NRC to review.
- NRC will need an account on the UNIX machine for Autonomy spidering purposes.
- The DOE will update the OCRWM LSN on a weekly basis or some workable period of time. This will not affect the required certification and re-certification processes.
- The NRC is going to set up a test bed with 1,000 or so documents. The DOE is requested to participate in the design and implementation of the LSN design. Dee Jensen has committed OCRWM to participate in the tests.
- The LSN is to be available 7 days a week from 6:00 AM EST to 9:00 PM WST.
- How are e-mail and attachments going to be represented in the LSN? DOE to research this topic.

4.0 Discussion concerning Text Accuracy

Dr. Tom Nartker, University of Nevada, Information Sciences Research Institute, explained that, based on the results of their research, a large percentage of the DOE documents in the OCRWM Records Management System (RMS) are forms containing handwriting. Dr. Nartker and his staff have been performing document analysis for the DOE relevant to use in the LSN and the RMS for the last 10 years. Dr. Nartker's research facility is regarded as one of the few document capture, conversion, search, and retrieval experts. In addition, there are a large percentage of draft documents containing marginalia in OCRWM's RMS. Since most of the searchable

information associated with forms is contained in the header, it is expected that little would be gained by processing forms through an Optical Character Recognition process wherein the images are converted to text. A question to be posted to the LSNARP Chairman is: Can we classify forms as image only? DOE will provide metrics supporting this issue to the NRC as well as metrics relevant to the accuracy of OCR conversion of OCRWM documents.