Collaborative Electronic Notebooks as Electronic Records: Design Issues for the Secure Electronic Laboratory Notebook (ELN)

James D. Myers
Pacific Northwest National Laboratory
PO Box 999
Richland, WA 99352 USA
+1 610-355-0994
JIM.MYERS@PNL.GOV

Key Words: Electronic notebook, electronic records, web applications, collaborative systems and applications

Abstract[1]

Current electronic notebooks (EN) can be grouped roughly into two general classes—personal/group productivity tools and enterprise records/knowledge management systems. Personal/group productivity-oriented ENs extend the notebook metaphor in terms of supporting multimedia annotations, automating workflow and data processing, supporting simultaneous use by distributed researchers, providing displays on personal digital assistants (PDA), etc. Enterprise records/ knowledge management-oriented ENs tend to limit content to static, paper-like documents (e.g., saved in portable document format [pdf] using Adobe Acrobat) and to emphasize searching, digital signatures, and enterprise-level management functionality. A variety of complex technical and business issues make it difficult to bridge the gap between these types of ENs. The latest version of Pacific Northwest National Laboratory's (PNNL) Electronic Laboratory Notebook (ELN) incorporates a variety of features for both personal and group productivity as well as records management, thereby providing an example of how such capabilities might be combined to meet the needs of individual researchers and their organizations. This paper outlines the issues involved in using ENs as records, discusses the design choices made in the ELN to address these issues while maintaining a focus on collaboration and researcher productivity, and highlights additional areas where conflicts between productivity and records functionality remain and would have to be resolved, via technology and/or policy, to successfully use the ELN as an official record.

[1] This manuscript was written by an employee of Battelle Memorial Institute, which operates Pacific Northwest National Laboratory for the U.S. Department of Energy under Contract DE-AC06-76RL0 1830. The U.S. Government retains and the publisher, by accepting this article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

1. INTRODUCTION

Traditional paper research notebooks, and the policies for managing them, are highly constrained by the requirement that they serve as an archival record that is both legally and scientifically defensible. While variations still exist, the general features of research notebooks and policies for their use have become part of the research culture. Notebooks are bound and have numbered pages. Entries must be made using indelible ink. Blank areas must be of limited size, revisions must preserve the original entries. Notebooks have a defined lifecycle that includes checkout, inspections, check-in, and archiving. Researchers must sign entries and have them witnessed on a prescribed schedule. While researchers complain about these constraints and few would claim that paper notebooks are an ideal solution, paper notebooks are universally accepted because of a combination of good design, familiarity, and a lack of alternatives.

Electronic notebooks promise an alternative, one with clear advantages to the researcher in terms of multimedia inputs, automation of entries, location independence, group access, etc. Trends in scientific research—towards distributed project teams, discovery- and informatics-based research, large-scale coordination of efforts through Collaboratories and Grid-based virtual organizations—emphasize the collaborative advantages of ENs and provide a strong driver away from paper records [1,2,3]. However, to succeed in displacing paper notebooks, ENs must deliver these advantages and, at the same time, fulfill the paper notebook's traditional role as a legal record. Unfortunately, the technologies and processes developed to meet records requirements with paper notebooks do not map directly to digital media, and/or they conflict with the type of productivity enhancements noted above.

Signatures are an obvious case. Written signatures on paper are relatively hard to forge and, once created, are hard to delete or copy. Direct digital equivalents—recordings of the image or pen strokes attached to a note—can be duplicated or deleted easily. One must look to public-key cryptography, a very different technology, to find a functional equivalent of hand-written signatures, and must also amend policies to make such public-key based 'digital signatures' the legal equivalent of

handwritten signatures. Analogous issues exist with the solutions used in paper notebooks to assure that all content is recorded and it is never altered or deleted—the notebook binding, notebook and page numbers, using indelible ink, crossing out blank areas on a page. An EN system could emulate the write-once semantics of these features, but this approach would not stop someone with access to the underlying database or operating system from altering the information in ways undetectable to the EN application. Again, cryptography provides a means to implement a functional equivalent.

This type of deconstruction and redesign is necessary across the range of EN capabilities. Are two-dimensional, static representations of notes a requirement or just a paper-based solution to a more fundamental need? If so, are there better electronic solutions to the underlying need? Do such solutions have implications for the procedures and policies related to notebook use?

The BACKGROUND and DESIGN sections of this paper provide a brief review of EN research and an overview of the basic architecture and capabilities of PNNL's ELN. The section entitled RECORDS FUNCTIONALITY IN THE ELN follows with a discussion of the mapping of records functionality from paper to electronic notebooks and the design choices made to extend the ELN for use as a record. The last section—CONCLUSION—provides a summary of where additional technical capabilities or policies and procedures would be needed to make the ELN usable as a record and a commentary on the prospects for future EN systems that would combine best-of-breed productivity and records features and, thus, provide a general replacement for paper notebooks.

2. BACKGROUND

Researchers have explored a variety of directions in the development of ENs over the past two decades. Initially, ENs focused on text- and hypertext-based annotation to support individual researchers [4,5,6]. Inspired by the Internet and later the World Wide Web, ENs added support for multimedia annotations and collaborative use [7,8,9,10,11]. Pen-and-voice input, as well as support for lightweight input devices (e.g., wireless laptop computers, PDAs, and tablet computers), has been investigated, as has the concept of the EN as an active component in the data analysis workflow [12,13].

More recently, digital signature technologies, aided by the increasingly recognized advantages of knowledge management, have prompted investigation of ENs at the enterprise level as legally defensible records. Efforts in industry and government have documented the need to recognize ENs as primary, legally defensible records at the enterprise scale [14,15]. Custom solutions have been built and deployed in industry, and commercial notebooks that target structured environments, such as analytical chemistry and pharmaceutical laboratories, have begun to emerge [16,17,18]. However, these types of notebooks tend to restrict themselves to a fixed set of two-dimensional annotation types that can be stored in printable documents.

The ELN originated in EN research begun at PNNL in 1994 and has been developed primarily as part of a three-way exploration of EN concepts involving researchers at PNNL, Lawrence Berkeley National Laboratory (LBNL), and Oak Ridge National Laboratory (ORNL) within the U.S. Department of Energy's DOE2000 Collaboratory program [19,20]. The DOE2000 notebooks have made significant contributions in the areas of personal and distributed group productivity noted above as well as in defining standard data formats and providing extensibility through programming interfaces [21,22,23].

3. DESIGN

The core of the ELN consists of an interactive browser-based client and a notebook server implemented using common gateway interface (CGI) scripts that run on a Web server (see Figure 1)[24]. The client interface is rendered by a locally installed Java application that handles user interaction—login, selecting page views, searching, entering notes, etc.—and the browser, which serves as the initial launch point and displays the contents of individual pages. A small, signed Java applet within the browser launches the application and maintains communication with it through a socket. Interaction with the server is solely through HTTP requests. The server, a CGI script written in Perl, responds to HTTP requests from the application to implement the basic functionality of logging in, discovering the notebook contents, submitting new entries, etc. Requests to view page contents are forwarded from the client to the browser via the applet. The server then responds to HTTP requests from the browser by generating the HTML and JavaScript

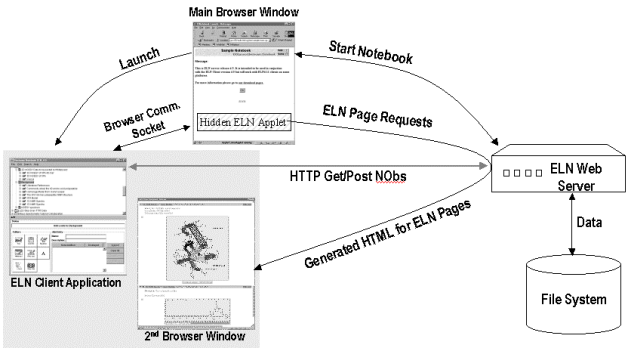

Figure 1. (Reproduced from [24]) The ELN architecture showing interactions between the ELN Client application, the user's web browser, and the web server hosting the ELN server CGI scripts. Gray box: The ELN interface consisting of the Table of Contents and palette of editors in one window (left), and a selected page of the notebook showing two entries (right).

necessary for the notebook page views. This architecture has allowed us to make use of browser-based HTML rendering while freeing the bulk of the client logic from being limited to the lowest common denominator of Java support across browsers.

The central data management concept in the ELN is the notebook object (NOb), which consists of an opaque, typed piece of data and key/value pairs of metadata—data about the data. The NOb data are the text, images, or other data composing an entry, or it can be the list of other NObs, thus allowing representation of a potentially infinite hierarchy of chapters, pages, notes, and sub-notes. The metadata include a core set of key/value pairs {authorname, datetime, label, description, objectid, datatype, and dataref} whose meaning is understood by the notebook, as well as arbitrary key/value pairs with meaning for the applications that produce or process NObs.

New information is added to the ELN though editors that are invoked via a button palette within the main client window. The standard set of eight editors allows the creation of text (plain, HTML), equations (TeX), and sketches, uploading files, and capturing images from the computer display. A simple programming interface allows new editors to be dynamically added to support new data types. On the server side, rendering scripts that generate HTML and/or invoke applets and plugins appropriate for the content type controls the display of individual entries. Again, new rendering scripts can be dynamically added. The ELN also supports a lightweight forms-based input mechanism via a forms editor/renderer combination that allows new forms to be defined using an HTML-centric syntax. Finally, a specialized ELN Wizard client exists that can be incorporated into data acquisition or analysis software to partially or completely automate the creation of new notes while eliminating the need for a web browser.

4. RECORDS FUNCTIONALITY IN THE ELN

While many of the design decisions for early versions of the ELN anticipated its use as a record, it is only with the most recent versions (Version 4.5+) that significant records-oriented capabilities have been added. The development of these capabilities has been partially guided by regulations such as the FDA's CFR 21 Part 11 standard for electronic signatures and Good Laboratory Practice (GLP) standards, work within the Collaborative Electronic Notebook Systems Association (CENSA) to define requirements for ENs as records systems, academic research on the mapping between paper and electronic records, and ongoing discussions with records managers and intellectual property specialists within the DOE national laboratories [25,26,27,28]. As discussed below, making the ELN strictly compliant with existing standards has been secondary to the goal of understanding how technology and policy might evolve to allow the creation of advanced ENs that capture the productivity enhancements demonstrated in current ENs while meeting the underlying requirements to serve as records.

In simple terms, the functionality required of a record is the documentation of the familiar 'who, what, when, where, why, and how' of events plus the preservation of this information for long periods of time. Preservation includes maintaining the ability to access the information and providing evidence that the information is authentic and has not been changed.

4.1 Expanded Content

With paper notebooks, the original documentation is all produced directly by the researcher and is limited to what can be recorded by pen or, in some cases, taped into the notebook. Information that cannot be easily reduced to paper may be referenced in the notebook and stored as a separate record (e.g., computer files on CDROMs or voice recordings on audiotape). The accuracy and completeness of the information are limited by the errors and costs involved in manual transcription and coordinating multiple independent records.

ENs can provide capabilities directly analogous to paper–manual entries made by keyboard, mouse, or pen (e.g., using a graphics tablet). However, if the underlying requirement is producing complete and accurate documentation rather than migrating manual processes directly, the 'productivity' features of an EN, such as support for multimedia notes and automation of entries, can be seen as providing an enhanced initial record that unifies notes and data and captures more detail with fewer errors and omissions. In a collaborative research scenario, where a primary purpose of the EN record is to allow distributed teams to evaluate, reproduce, and/or extend the documented experiments, the inclusion of 'raw' data is key—the ability to reanalyze data, potentially using different procedures, is an essential technique for learning.

However, such entries raise concerns with respect to exposing additional information to later scrutiny and being able to reproduce the exact view seen by the researcher. The concern regarding the unification of notes and data is simply that by making raw data part of an official record, one opens an additional step in the scientific process—the researchers' interpretation of the raw data—to external scrutiny. While this can certainly be seen as a source of risk in, for instance, future litigation concerning the efficacy and safety of a drug, it can also provide additional validation of a claim as well as providing opportunities for future data mining (re-analysis of data for other purposes). For intellectual property scenarios, in contrast to uses where liability is the primary concern, the benefit of additional evidence is

seen as outweighing the new risks. However, with increasing calls for publication of data in many scientific communities and growing expectations by government agencies for access to original data, the trend is clearly towards inclusiveness.

The concern regarding the capture of more detailed information relates to the fact that scientific data is not studied directly but via 'views' created with software, for example a rotatable three-dimensional image of a molecule's structure generated from a file containing the list of atoms and their relative X, Y, and Z coordinates [29]. Aside from the issue of preserving the view software (discussed below), the concern is that aspects of the view (e.g., rotation angles, zoom level, color assignments for atoms, etc.) are not part of the data *per se*, so simply recording the data does not capture what the researcher actually viewed. The importance of the view in the researcher's process and the many-to-many relationships between data sets and views argue for treating them as first-class entries on their own. The current state of the art for capturing views is probably reduction to a standard pixel-based image file such as GIF, JPG, or PNG. However, one might anticipate software that could generate more complete descriptions of the data processing performed, both the mathematical transforms applied to the data and the visualization methods used to generate a view, that would supplement or replace an image and be required for future best-practice. In either case, treating a view as a first class entry allows for ENs in which raw data has a shorter retention period than views, making it possible to tune an EN based on a perceived difference in the timescale over which benefits and risks accrue in a given situation.

The ELN, as originally designed, supports arbitrary types of entries, which can include raw data along with data views and manually entered text and drawings. The editors available for a given notebook installation can be configured by an administrator to limit input to specific types as desired. Further, whether and how entries are rendered can also be controlled. Thus, the ELN could be configured to only accept text and image entries, to also accept data files but provide no rendering of their content (simply providing a link to download a copy of the file), or as is the current default, to allow arbitrary content and provide a dynamic rendering of the content (e.g., a rotatable molecule, a zoom-capable graph, a playable movie). In the latter case, use as a record would probably include policies and procedures requiring capture of relevant static views via automated or manual means. The ELN's ImageCapture utility provides a simple mechanism for complying with this requirement as it can capture the current screen image from external programs or from the dynamic rendering of data within the notebook. Beyond the general Wizard interface for programmatic input, the ELN currently does not have a mechanism for automatically capturing static views from dynamic renderers; however, we are currently investigating means to support structured 'pedigree' information (e.g., documentation of the link between data and views including the specific data inputs used, the version of the program producing the view, the parameters describing the view, etc.).

4.2 Automation of Entries

As noted above, the ELN supports both manual and automated entries—the 'what' of the event. It completely automates capture of 'who' makes entries and 'when' they are created. Entries are automatically tagged as being authored by the person logged in to the ELN client, and the current time is attached to each entry upon submission. Describing 'how,' 'why,' and 'where' experiments are done remains primarily a manual process, though the ability to add new editors to the ELN and to create forms for input can provide some scaffolding to prompt researchers. (If such information is available in other programs [e.g., a procedure has been programmed into a robotic sampling system], the Wizard interface could be used to allow programmatic entry of the information into the ELN.)

4.3 Evidence of Authenticity

The combination of physical properties and operational procedures described above in the Introduction combine to provide this evidence in paper notebook systems. Physical access control to the notebook plays a key role. Binding and page numbering, using indelible ink, and writing dates and signatures (author and witness) provide the primary evidence that content is complete and unaltered and that the information comes from trusted sources, but external documentation of who had access to a notebook at what times adds important corroboration.

For ENs, the direct analogs of these forms of evidence are not credible. As noted in the Introduction, digitized written signatures are easily copied, and it becomes possible, without additional controls, to alter page numbers, dates, etc. Physical access controls to a laboratory and/or the location of the physical notebook server are not sufficient to limit access to the EN. Even the concept of applying such controls to one 'official copy' of the record (e.g., the original paper notebook and, at a later time, a microfilm copy) loses meaning when all copies are absolutely identical and the system itself may involve replicated servers and/or integrated backup capabilities. The use of ENs in collaborative scenarios does not fundamentally alter the records requirements. However, it does provide an incentive to use standardized technical methods to enforce EN policies and procedures due to the increased costs and risks associated with coordinating manual mechanisms across enterprises.

In the ELN, entries can be completely characterized by their content, associated metadata, and position in the hierarchy of chapters, pages, and nested notes, all of

which is recorded in the server file system. Due to the HTTP-oriented design, there is no server state to consider as part of the record. Thus, evidence within the ELN concerning the authenticity of the record must address four concerns: 1) that the entries have been made by known and trusted sources, 2) that they have not been altered during transfer from the source to the server, 3) hat they have been properly stored by the server, and 4) that they have not subsequently been altered on the server system.

Fortunately, there are functional analogs for ENs, primarily involving various aspects of cryptography. Some background is necessary to understand the high-level functionality needed by ENs and the assumptions underlying cryptography-based evidence. Two key concepts—hashing and encryption—underlie the higher-level capabilities of interest. Hashing involves producing a relatively short number from the information of interest with a 'one-way' mathematical function that is easy to calculate but hard to invert. Thus, given an EN entry, it is easy to calculate its hash, but given the hash, it is impossible to reproduce the entry and 'hard' to find a modified entry that has the same hash. (By choosing the hash algorithm and hash size appropriately, the cost of buying sufficient computing power to find a modified entry with the same hash can be made much larger than the value of the information in the entry.) Encryption involves using a secret key to encode information in a way that makes it unintelligible unless the key is available to decode it to its original form. Public-key encryption involves two keys; information encoded with one key can only be decoded with the other. Thus, if only the researcher knows the first key and the other is made 'public,' one can assure that anything that can be decoded with the second key was produced and encoded by the researcher. In theory, one could encode the entire entry using the first key. In practice, one usually encodes only the hash of the entry. This makes a readable copy of the entry available while still assuring that any changes can be detected; recalculating the hash and comparing it with that obtained by decoding the one created by the researcher will allow detection of any change. Encoding the hash also separates the 'digital signature' from the content, allowing it to be managed as a separate set of bits without disturbing the one-to-one cryptographic tie between them.

A few additional concepts are needed for a complete solution. To provide evidence that a given researcher authored specific content, one must know, in addition to being able to verify the cryptographic signature, that the public key being used actually belongs to the person in question (i.e., corresponds to their private key). In theory, one could publish a statement to the effect that "the following number is the public key of Dr. Jane Q. Researcher, employee of Science Corp." in the public record (e.g., in the *New York Times*), however this approach is cumbersome given the number of researchers and the cost of checking the public record each time a signature must be verified. In practice, 'certificates' are used. Certificates are essentially statements of the form above, digitally signed using a key belonging to an organization (the researcher's employer or a provider such as Verisign), which can be sent along with the signature. One must still ultimately verify the organization-level signature, but this can be done using fewer (relative to the number of employees) organization-level certificates with relatively long lifetimes that can be distributed outside the system—for example, as part of software installation (as occurs with commercial browsers today). The scalability of such a certificate system, and the ability it provides for local verifications of the authenticity of EN content (based on local cryptographic calculations versus trust in a remotely managed EN system), are particularly valuable in collaborative scenarios.

Evidence that a given researcher has digitally signed some information then rests on four assumptions: 1) that the organization issuing a certificate to the researcher is trustworthy, 2) that their procedures for assuring the identity of the researcher (e.g., by requiring the researcher to present a photo ID during the issuance procedure) are effective, 3) that the algorithms and software used in creating the certificates and signatures have not been compromised, and 4) that the researcher has kept their private key secure and that it has not been shared or stolen. The first two assumptions are essentially the same as required with paper notebook systems. The third assumption is a new concern that is best managed by choosing algorithms and key lengths that are well studied and have been accepted in the larger community and by using third-party cryptographic software (commercial or open source). The last assumption is a concern because a private key is just a series of bits that, unlike a written signature, can be shared and is too long to be memorized. Thus, policies must be implemented to prohibit voluntary sharing, and some mechanism of storing the private key in a secure manner is required. A standard means of satisfying these requirements is to store the private key, encrypted with a memorizable password/PIN, on the researcher's desktop computer. More secure protections can include using a biometric identifier to access the private key and/or storing the key in a removable smartcard rather than on the computer disk.

Digitally signing notebook entries not only provides evidence that the researcher created them, but also provides evidence that the entry has not been changed since being signed. Without access to the researcher's private key to create a new 'forged' signature, any change to the entry will be detectable. Signing metadata along with the content can 'bind' the information and offer proof that neither has been changed. Thus, a digital signature that includes content and metadata concerning the time the entry was made becomes an assertion on the part of the signer concerning the time of entry, an action

that is analogous to writing the date and time with a written signature. Similarly, signing content plus the statement "read and understood by" or "reviewed by" becomes the assertion expected of a witness. Signing the 'structure' of an EN (i.e., signing the list of notes on a page, pages in a chapter, and/or chapters in the notebook) can provide evidence that it has not changed, which is analogous in functionality to the page numbering and binding in paper notebooks.

Digital signatures provide evidence concerning the person who created them and assurance that the content, metadata, and/or structure have not been modified after signing. However, they do not prevent attempts to alter data. Changes made after signing would be detectable, but there are important cases where signing would be delayed—for example, a page or chapter cannot be signed until the information is complete (adding a new page to a chapter is a 'detectable modification' of the chapter), or an author or witness may sign batches of notes or pages on a schedule. This concern can be reasonably addressed by implementing access control, encrypting network traffic within the system, logging entries, and protecting the server storage electronically and physically, all precautions that are desirable for other reasons (e.g., protecting intellectual property from theft, and preventing malicious destruction of the record). Additional strategies, such as periodically signing the entry log or versioning all content and verifying the signature of the earlier version before allowing any modifications, could add additional security at the cost of additional complexity and computational overhead.

The secure version of the ELN uses software from Entrust, Inc., for its client-side cryptographic capabilities and to provide overall management of users and certificates. This software was chosen primarily for two reasons: 1) Entrust was one of the first vendors to release a Java development kit, and 2) it was possible for the ELN project to make use of an Entrust infrastructure being tested and deployed for use at PNNL within a lab-wide digital ID system. The Entrust system provides a variety of tools for managing certificates (issue, revoke, periodically replace, etc.). The Java toolkit provides high-level interfaces to access user keys and certificates, create and verify signatures, invoke management functions, etc. The capabilities provided by Entrust are encapsulated well in the ELN code, and it would be relatively easy to modify the ELN to use an alternative certificate management infrastructure.

The ELN server, which is written in Perl, relies on cryptographic functionality available in the open source OpenSSL toolkit, both directly and via the mod_ssl extension to the Apache Web server.

In the secure version of the ELN, the user and the web server hosting the ELN server are both issued certificates. The web server is configured to accept only HTTPS connections using secure sockets layer (SSL) encryption with mutual client-server certificate authentication. All ELN client server communications are thus encrypted, client access can be limited to authorized users based on an access control list maintained on the server, and the user can be assured they are connecting to an approved ELN server. During login, users must supply a username/password pair that is used within the Entrust components to retrieve their private key from an encrypted local store or an Entrust directory on the network ("roaming" credentials).

ELN entries are made in the same manner as previous versions, although the submit operation is now carried out over HTTPS. A design decision was made to make signing a separate operation, thus allowing unsigned entries to be submitted to the server (e.g., as drafts) and allowing operations such as witnessing and approvals to be handled via the same mechanism as author signatures. If required by organizational policy, the requirement for a signature during submission could be supported with a relatively minor change to the software.

The ELN signing operation can be applied to any entry or hierarchy node (e.g., a chapter or page). A signature window (Figure 2) shows all current signatures on the selected entry. Selecting 'Sign' from the menu pops up a dialog requesting user credentials and the purpose of the signature. Purpose statements are limited to those configured on the server. The defaults statements are 'Authored By,' 'Read and Understood By,' and 'Approved By.' These statements may be replaced or supplemented according to organization policy. Two-part credentials that include the Entrust profile name for the user and the password to decode the profile are required for each signature, which is consistent with the FDA's CFR 21 Part 11 guidelines for electronic signatures. Internal to the ELN client, the identity of the signer and their credentials are treated completely independently of those of the user currently logged-in, making it technically possible, for instance, for an author and a witness to add their signatures during the same notebook session.

The signing operation retrieves the content of the specified entry from the ELN server via an HTTPS connection. The metadata for the entry, which includes any prior signatures, is retrieved also. As a prerequisite to

Figure 2. Example of the ELN user interface for viewing digital signature information. The menu includes items for creating new signatures, verifying existing signatures, and configuring whether verification includes verification of the signer's certificate.

signing, all prior signatures are verified. By default, this includes direct verification of the signature itself but not the verification of the signer's certificate chain. An option in the signature window will enable certificate chain verification as well. The signing algorithms (currently hard coded to SHA1 and CAST for hashing and encryption respectively) are then applied to a concatenation of the selected purpose string, the current time, and the content. For signatures of unsigned entries, the content is the byte representation of the entry's content. For signed entries, the bytes of the most recent previous signature are considered to be the content. (Since the previous signatures have been verified, this approach still provides a direct cryptographic link to the actual entry, and it adds a mechanism to prove the ordering of signatures.)

Metadata properties are not signed as part of the entry in the current ELN. This decision was made for several reasons. Most practically, it allowed storage of signature information as additional properties. From a more strategic perspective, it allows the dynamic creation of additional properties (i.e., for categorizing the entry) without breaking signatures of the content. The most general solution, developing a mechanism to specify whether specific properties are or are not included in a signature, was beyond the scope of our development effort.

The signing process occurs on the client machine and the signer's private key is only used locally. When the 'Sign' button is pressed in the client, the signature is created and the bytes of the signature are transmitted securely to the server. As a convenience, in addition to storing the signature as a property, the signers name (common name from their certificate), their certificate, the time of signing, and the purpose are stored as additional properties. (All of this information can be obtained and verified from other sources.)

These properties are submitted to the server using the same method used to add content and non-signature metadata. To prevent attempts to use the method to modify content during signing (i.e., with a modified client), the ELN server prohibits any changes to content when signature properties are involved. Similarly, rather than trusting the time stamp provided by the client, the server verifies that it is within a configurable number of minutes of that on the server machine. Thus, the server clock is the ultimate time source for the current ELN system (which may then be synched with GPS or network time sources through other means/procedures).

The client signature window also allows the verification process to be run manually, with or without verification of the signers' certificate chains. As with signing, verification causes the entry contents to be retrieved from the server. Although verification is left as a manual process for efficiency, the fact that signatures exist on an entry is marked as part of the main ELN page display via a signature icon. Further, the signature window is synchronized with the ELN table of contents; selecting an item in the table of contents will update the signature window display to show the signatures for that particular entry.

In addition to recording digital signatures, the ELN server logs all ELN activity. Since all communication with the server occurs via the web server, all calls to the server are already recorded in the web server logs, but the ELN server maintains its own separate log that provides a more readable description of notebook activity. In addition, an upload log provides a summary of all additions to the ELN. Writing/copying the upload log to write-once media would make it possible to detect changes to the ELN such as the deletion of a signed page within an as-yet unsigned chapter.

4.4 Long-term Preservation

The combination of signatures and logs in the ELN provides evidence of the activities documented that can be verified using the ELN client. However, maintaining the ELN content and the ability to display and verify it over the long term—durations of 50–100 years or more—requires additional work. Over such timescales, one can expect computing hardware, operating systems, programming languages, and recording media to become obsolete. Assuming continuing advances in computing power and mathematics, one should also expect that the cryptographic operations involved in generating signatures will become 'breakable' (i.e., that the costs to forge them will decrease below the value of the signed data).

Because of all these factors, preserving digital data over the long term is a difficult problem. However, the difficulty is more in finding the best low-cost means of preservation rather than in finding a means. In theory, data can be migrated to newer media as required, data formats can be updated and revalidated, software can be maintained on legacy hardware or revalidated on newer hardware, etc. Revalidation involves not only testing that the new data format or software accurately reflects the old, but also the documentation of this process and the creation of a signed statement by a trustworthy entity that it was performed. The latter might, for example, take the form of a digital signature with the purpose 'Read, Updated, and Revalidated By.' One could also argue that for multimedia data (i.e., the types of information that are not recorded in current paper notebooks), simply maintaining the original information and ignoring issues of upgrading formats and software to maintain readability is a valid solution. As long as the original data, with metadata about its format, exists and is authentic, updating the data and validating new software to view it could be done on-demand when the record is questioned. Such a model, which separates the task of preserving the

data and evidence of its authenticity from the ability to view it, may be especially appropriate to collaborative ENs where users of the data may reside in different organizations with different computing infrastructures and different validation requirements.

The preservation of digital signatures can be handled in a similar manner. Signatures will lose their cryptographic link with the data if the data format is changed (which changes the bytes in the data) or if breaking them becomes computationally feasible. However, a process of inspecting data and signatures, and producing a new signature that states that the earlier signatures were valid as of the date of the new signature, can maintain a chain of cryptographic evidence. To maintain evidence during a format update, the new signature would be a statement that the signer tested the validity of the existing signatures on the original data and verified that the new data in the updated format does correspond to the original. To maintain evidence as the cryptographic strength of earlier signatures becomes questionable, the new signature, created using a more secure mechanism (new cryptographic algorithm and/or longer key lengths), would simply state that the signatures were tested and found valid as of the date of the new signature. The chain of evidence then rests on the trustworthiness of the new signer who could be a notebook archivist within an organization, an external notarization agency, etc.

In practice, predictions concerning which media, hardware, and software will require the least frequent upgrades and will have the lowest re-validation costs are not easy. The ELN was designed with preservation cost issues in mind. The ELN's stateless nature reduces the problem of migrating data to that of migrating files. The use of Java, Perl, Javascript, and HTML, and the release of the source code under an open source license, makes maintenance of the ELN software feasible. Work is underway to migrate to Java on the server and to incorporate XML, which is arguably the current best practice for creating maintainable software. The ability to configure the ELN to use external software for editing and viewing data allows organizations to make individual decisions about which data types are worth the cost of maintaining and/or whether to simply present multimedia entries as links to the data, thereby pushing the issue of reading and viewing these formats out of the scope of the EN system. The design choice to allow signature purpose statements to be configured and to implement signatures as a chain with each signature being applied to the previous signature (rather than all signatures being applied to the entry itself) provides infrastructure required to implement data and signature migration strategies as discussed above.

5. CONCLUSION

The discussion above and the design of the ELN v4.6 argue that ENs with strong personal productivity and group collaboration features also can be made to function as official records. As EN research continues and additional types of functionality are developed (e.g., support for laboratory workflow, federation with other data management systems, management, analysis, and mining agents), there do not appear to be theoretical barriers to maintaining records capabilities. While such a statement cannot be taken as an absolute given the lack of prescriptive rules for what types of evidence such systems must provide (a consequence of legal systems based on guidelines and precedents), there is nothing inherent in the mapping between productivity and collaborative functionality and records functionality, or any limitations in current technologies for implementing electronic records, that suggests cases which could not be handled. Cryptography-based digital signatures and good system design that includes protected information flows and comprehensive logging, can provide stronger evidence of authenticity than is possible with paper [30]. Further, signatures and timestamps based on public key certificate systems provide a scalable means of verifying the integrity and authenticity of information retrieved from remote EN repositories, thus reducing and formalizing the trust relationship required to operate a collaborative EN across organizations. Given these capabilities, it is hard to imagine barriers to the long-term acceptance of multimedia, collaborative ENs as records systems. However, as EN capabilities go beyond those of paper, the lack of clear requirements for legal defensibility will certainly slow adoption of EN systems. Thus, community efforts to define 'best practices' and educate EN stakeholders, such as those within CENSA, will be an important driver towards EN acceptance.

The current ELN provides an interesting test bed for discussion of EN best practices. It incorporates a wide range of productivity features and, combined with an Entrust infrastructure, state-of-the-art digital signature capabilities. Combined with appropriate organizational policies and procedures and configured to limit the types of data input, the ELN can meet current guidelines for digital records while providing extensibility and infrastructure to demonstrate additional possibilities for future practice and to support evolution and maintenance to new practices over time. As noted previously, the ELN records capabilities, while functional, were created as a proof-of-concept; for production use, the development of additional capabilities for administrating multiple notebooks, monitoring use, enforcing signing and

witnessing policies, etc., would be advisable to reduce overall systems costs. The ELN source code has been released under an open source license and is available for adaptation to specific uses or as a platform for continuing EN research. Within PNNL and ORNL, the ELN, and the design concepts for collaborative EN systems developed within the DOE2000 project will be leveraged in a new research project considering a broader view of collaborative scientific annotation that will continue to examine the interplay of productivity, collaboration, and records requirements in the design of EN systems.

6. ACKNOWLEDGEMENT

The author would like to acknowledge helpful discussions with Dave Long and Steve May of PNNL's Intellectual Property Department, the efforts of Elena Mendoza and Michael Peterson of PNNL's Computational Science and Mathematics Division in testing and debugging the records-related functionality of version 4.6 of the ELN and in packaging it for release, and the work of John McCoy and PNNL's Entrust administration team for help in debugging several PKI-related issues. This work was supported by the U.S. Department of Energy through the DOE2000 program, which is sponsored by the Mathematical, Information and Computational Sciences Division of the DOE Office of Science. Battelle operates PNNL for the U.S. Department of Energy.

7. REFERENCES

[1] Kouzes, R.T.; J.D. Myers; W.A. Wulf. 1996. "Collaboratories: Doing Science on the Internet." *IEEE Computer*.

[2] Foster, I. 2002. "The Grid: A New Infrastructure for 21st Century Science." *Physics Today*, No. 55: 42-50.

[3] Atkins, D.E.; *et al*. 2002. "Revolutionizing Science and Engineering through CyberInfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on CyberInfratructure." Draft 1.0, http://www.cise.nsf.gov/b_ribbon/.

[4] Kaplan, H. 1984. "An Electronic Notebook." *Popular Computing*, No. 3: 174-&.

[5] Costello, T.D. 1986. "RS/1, An Electronic Notebook." *Journal of Coatings Technology*, No. 58: 75-80.

[6] Figueras, J. 1987. "An Electronic Notebook for Chemists." *Acs. Symposium Series*, No. 341: 37-47.

[7] Shipman, F.M.; R.J. Chaney; G.A. Gorry. 1989. "Distributed Hypertext for Collaborative Research: The Virtual Notebook System." *Proceedings of the Hypertext '89 Conference,* ACM Press, pp. 129-135.

[8] Fowler, J.; D.G. Baker; V. Kouramajian; H. Gilson; R. Dargahi; K.B. Long; C. Petermann; G.A. Gorry. 1994. "Experience with the Virtual Notebook System: Abstraction in Hypertext." *In Proceedings ACM CSCW'94 Conference,* 133-143.

[9] Edelson, D.C.; R.D. Pea; L.M. Gomez. 1996. "The Collaboratory Notebook." *Communications of the ACM*, No. 39: 32-33.

[10] Hong, J.; G. Toye; L.J. Leifer. 1996. "Engineering Design Notebook for Sharing and Reuse." *Computers in Industry*, No. 29: 27-35.

[11] Strauss, D.; K. Meserve; T. Hess; R. Meyers; M. Kersey; J. Reilly; G. Thornton; G. Atwater; C. Jones; P. Graf; J. Jenkins. 1996. "Expressive Electronic Notebook Environment." *Abstracts of Papers of the American Chemical Society*, No. 212: 5-CINF.

[12] Malony, A.D.; J.L. Skidmore; M.J. Sottile. 1999. "Computational Experiments Using Distributed Tools in a Web-Based Electronic Notebook Environment, *High-Performance Computing and Networking, Proceedings*, No. 1593: 381-390.

[13] Malony, A.D., J.E. Cuny; J.L. Skidmore; M.J. Sottile. 2000. "Computational Experiments Using Distributed Tools in a Web-Based Electronic Notebook Environment. *Future Generation Computer Systems*, No. 16: 453-464.

[14] Lysakowski, R. 1993. "Electronic Notebook System Technologies and Architecture Base-Line Requirements." *Abstracts of Papers of the American Chemical Society*, No. 206: 103-COMP.

[15] Recordkeeping Requirements for Electronic Research and Development Notebooks, 2000, http://www.ornl.gov/records/notebooks.pdf

[16] Labbook, http://www.labbook.com/.

[17] IceBreaker, http://www.icebreaker.com/.

[18] R.M. Stember LABTrack. 2001. "Introducing a Legal Electronic Lab Notebook." *221st ACS National Meeting* San Diego, April 1-5, 2001, http://www.labtrack.com/acsindex.htm.

[19] DOE2000 Notebook Project Homepage, http://www.csm.ornl.gov/enote/.

[20] DOE2000 Project Homepage, http://www.mcs.anl.gov/DOE2000/.

[21] Myers, J.D.; C. Fox-Dobbs; J. Laird; D. Le; D. Reich; T. Curtz. 1996. "Electronic Laboratory Notebooks for Collaborative Research." In *Proceedings of the IEEE Fifth Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, Stanford University, California, June 1996.

[22] Mendoza, E.S.; W.T. Valdez; W.M. Harris; P. Auman; E. Gage; J.D. Myers. 1998. "EMSL's Electronic Laboratory Notebook." In *Proceedings of the WebNet '98 World Conference*, Orlando, Florida, November 7-12, 1998.

[23] Sachs, S.R.; C.D.S. Freitas; V. Markowitz; A. Talis; I-Min Chen; E. Szeto; H.A. Kuno. 1996. *The Spectro-Microscopy Electronic Notebook*, LBNL Technical Report #LBNL-39886, Lawrence Berkley National Laboratory, Berkley, California.

[24] Myers, J.; E. Mendoza; B. Hoopes. 2001. "A Collaborative Electronic Notebook." In *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications,* (IMSA 2001) August 13-16, 2001 Honolulu, Hawaii.

[25] FDA Office of Regulatory Affairs, Compliance References: Title 21 CFR Part 11—Electronic Records/Signatures, http://www.fda.gov/ora/compliance_ref/part11/.

[26] Collaborative Electronic Notebook Systems Association (CENSA) Homepage, http://www.censa.org/.

[27] Duranti, L.; H. MacNeil. 1996. "The Protection of the Integrity of Electronic Records: An Overview of the UCB-MAS Research Project." *Archivaria,* No. 42: 46-67.

[28] The InterPARES Project, http://www.interpares.org/
[29] Walther, D. 1997. "WebMol - A Java-Based PDB Viewer." *Trends Biochem Sci*., No. 22: 274-5.
[30] Dolak, L.A. 1999. "Patents Without Paper: Providing A Date Of Invention With Electronic Evidence." *36 Hous. L. Rev.,* No. 471.
[31] Myers, J.D. and A. Geist. Scientific Annotation Middleware (SAM) Project Website, http://www.scidac.org/SAM.