# *DOE-OSS File Room Project*
# *July 14, 1995 Update*

**James A. Rome**
**Johnny S. Tolliver**
**Patricia W. Payne**

Lockheed Martin Energy Systems

## *Goals of the Project*

***To convert the DOE-OSS File Room into a searchable, retrievable, electronic format.***

- The system must be useful, expandable, and upgradeable for at least a decade.
- The data-entry process should minimize required labor and training.
- The data should retain the format of the original document.
- The users of the database should require little or no training to use the software and to make complicated queries.
- Retrievals should be fast and accurate.
- Data should be in a form suitable for "cutting and pasting."
- Sensitive data and eventually Secret data should be protected.

## *How to Attack the Problem*

The data format and storage medium drive the other parts of the problem.

They

- determine how the data are entered and indexed.
- restrict the retrieval search engine and client software.
- Are the key to enforcing the security of the system.

## *Data Storage Format*

- Maintenance of original format
  - Scanned bitmap image + OCR searchable text.
    - > Bitmaps are hard to read or resize.
    - > Two retrievals are required to use the data — the bitmap and the OCR text.
  - Adobe portable document format (PDF).
    - > PDF is a compressed PostScript format that is usually smaller than a bitmap.
    - > The text is fully searchable and indexable.
    - > The page looks like the original, even using the same type of fonts, character attributes, etc.
    - > It can handle equations, pictures, and unknown characters as embedded bitmaps.
    - > Adobe Acrobat readers are free.
    - > The ($129) Acrobat Exchange has "groupware" features — comments, cross-references, searching.

## *Data Storage Media*

To handle multilevel-secure data imposes requirements on the system:

- All data must be labeled.
- Classification and declassification must be possible and easy to perform.

➢ In order to handle labeled multilevel data, we propose hosting the database on a powerful CMW workstation.

The data must be stored on a medium that satisfies several requirements:

- It can be written, erased, and rewritten.
  - Otherwise a single label must be applied to the entire medium, and individual items cannot be reclassified.
- The medium must be mounted using a labeled Unix file system.

➢ Today, only magneto-optical (MO) disks satisfy this requirement.

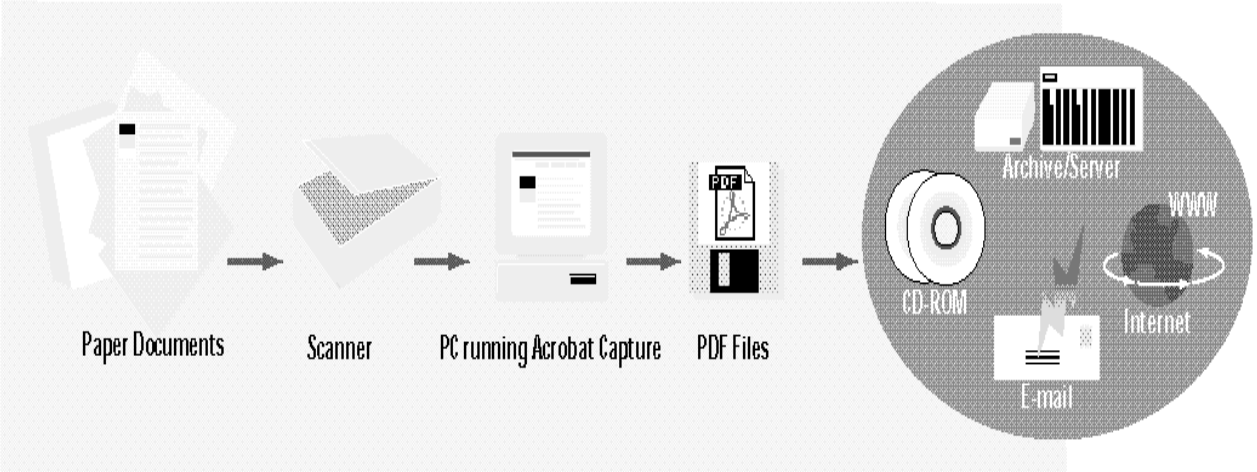
## *Magneto-Optical disks*

- Magneto-optical disks hold 1.3 Gb each (2 sides).
- They have unlimited lifetimes and probably do not require backups.
- They "look" just like a normal hard disk to the CMW computer. Files on these disks are individually labeled, so the security features of the CMW system are automatically available.
- Unlike CDs, individual data files can be added as they are generated.
- They can be stored in large jukeboxes.
- Their cost is about the same as WORM CD-ROM.

At the end of this year, rewriteable CDs are expected to become available, so this decision could change.

# *The Data Input Process*

The Adobe Capture software will be used to convert scanned document pages to PDF files.



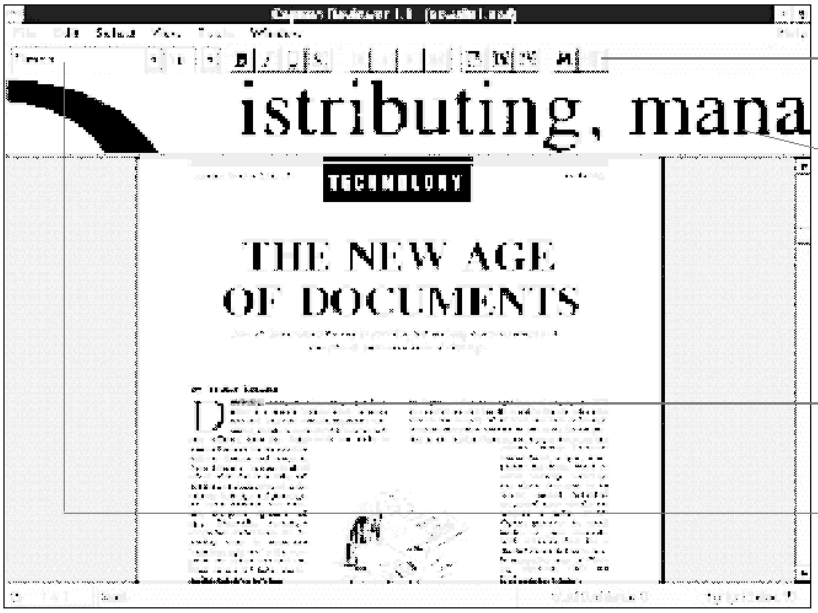

Acrobat Capture lets you scan paper documents and convert them to PDF files, producing an electronic version that replicates the original printed page and can be viewed or searched by users regardless of platform—Macintosh, Windows, DOS or UNIX.

Acrobat Capture lets you scan paper documents and convert them to PDF files, producing an electronic version that replicates the original printed page and can be viewed or searched by users regardless of platform—Macintosh, Windows, DOS or UNIX.

Adobe Capture and the PDF format have been chosen as the conversion/storage techniques for the LANL *Library Without Walls* project.

# *Adobe Capture Features*

**WYSIWYG Reviewer**



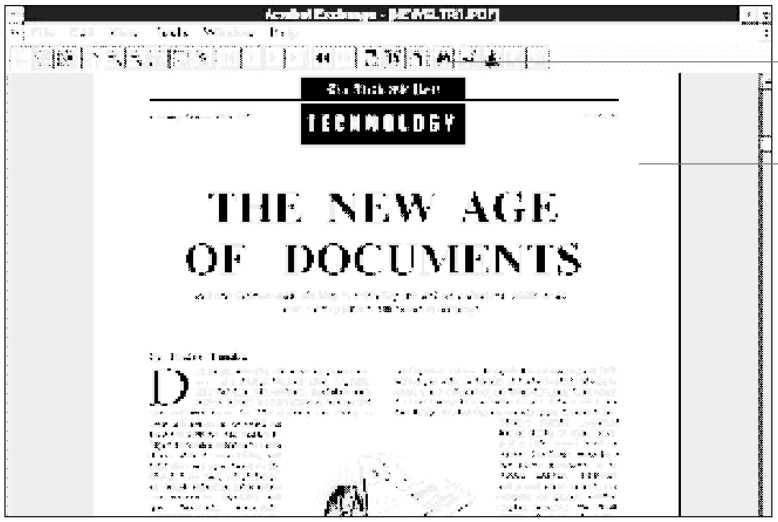

Built in dictionary checks spelling automatically and is fully customizable.

The Reviewer displays the complete page and a scrolling bitmapped image of the original document in full WYSIWYG mode.

Suspects are highlighted in yellow, making it easy to find and correct words, if desired.

After processing your file, you can select any word, line or text area and change the font style or size. Use any Type 1 font installed on your system.

**Your PDF file in Acrobat**




Full-text search lets you find information in seconds.

In Acrobat software, your PDF file looks exactly like your original document and works on any Windows, Macintosh, DOS or UNIX computer.

## *Other Document Conversion Issues*

- The document scanning, conversion, and indexing are all performed on Windows-based PCs.

- Separate indexes will be maintained for each classification level of document.

- Lexicographical indexes will be used.
  - ~ 30% the size of the data.
  - Stored on hard disks on CMW host.
  - Must be backed up.

- More utility can be added to the data search process by using *data thesauruses.*
  - Bill Clinton <—> President of U.S.

- The data must be written to the MO disks on the CMW host (or via a single-level remote session) so that the data can be properly labeled.

- A host-based reclassification procedure will have to be devised.

## *The Data Access and Retrieval Process*

We would like the database users to interact with the text retrieval engine by using a Web Client (e.g., Netscape or Mosaic) on their PCs.

The Web server on the host would enforce security (using the CMW host) and act as a front end for the text retrieval engine.

**Assumptions:**

- We can trust the user to declare the security level of his PC correctly.
- The network is secure.
  - We hope to support the Tessera Card when it is available.
  - We may turn on the Web-based security and authentication procedures (SSL and SHTTP) *in addition* to the CMW host procedures.

## *The Text-Retrieval Engine*

We have selected the Verity text retrieval engine called *Topic*.

- It is a quality product with many satisfied users.
- Adobe creates Verity indexes and uses their search engine in Acrobat Exchange.
- Verity has a version of Topic that is already integrated with a Web server.

*So, if all of this COTS software is available, why isn't the project already completed?...*

## *Remaining Challenges*

- Getting COTS software to run properly at multilevels on a CMW platform is nontrivial.
  - The Mosaic Web Server is running on our MLS HP. It works.
    But anyone on a listed host can access it.
  - We haven't tried the Verity search engine yet . . .

- Modify the Web software and the search engine to get them to utilize the security features of the CMW platform.

- Allow a single-level client (a Windows-based PC) to log in at a *selected* level on the CMW Web Server.
  - We have written a trusted program that checks the user's authorization and changes the logged-in security level.

- Getting HP to sell us the CMW server has been an exercise in frustration. But HP finally has sent us a copy of the CMW operating system.

## *Remaining Challenges (continued)*

- Devise a scheme to transfer PDF files and indexes to the secure server.
- Need a reclassification procedure.
- Determine the usefulness of setting up a thesaurus.