Transition from Paper to Electronic Documents

Jack Van Oosterhout*

Abstract

In a traditional environment documents are represented by paper. In the emerging electronic document world, all the normal functions associated with documents have electronic equivalents. Electronic publishing and the Internet represent a vast source of color and graphics, which are added to other documents.

Electronic document storage is enabled by new document interaction software. New user-friendly ad hoc document storage is needed for departmental use. Document file formats have to adapt to application independent format and display complex color documents. Printing and network scanning must be easy to use, directly interact with electronic documents and offer fast single copies.

Introduction

The world of documents as it exists for office equipment and solutions providing companies such as Sharp is changing rapidly. This article will review and predict the changes in office equipment and software that will be reinforce and enable those changes within the departmental office or workgroup. There are 3 separate trends that are driving these changes:

(1) Conversion to electronically based documents and work processes in offices.

(2) Ready availability of colorful and complex electronic documents via the Internet.

(3) Transition in the publishing industry to distribute and print and print on demand .

For a number of years writers have envisioned a world where all informational documents are processed electronically. This environment has often been called the "paperless office". In fact however, paper is far too convenient a means of displaying documents and making data available for rapid random access by people to be readily replaced.

The transition to a paperless office is actually a change in dealing with documents electronically rather than in the traditional form. We can examine the challenges and opportunities in the transition to electronic documents by understanding the functions of a traditional document system and extending these requirements to electronic formats. We can also see the need for ready conversion between electronic and traditional paper document types as well as new challenges brought by changes outside the office and home.

Our model for understanding the transition to electronic documents is the departmental or workgroup office environment. In a suitable organization all members of the workgroup are equipped with networked PCs and centralized or workgroup file servers. The department has its own filing system and library, probably in paper format. The model department's work process may be electronic or manual.

* Sharp Laboratories of America, Inc.

Many of the work processes in this kind of office are still more conveniently executed in paper format. There are several enablers needed for conveniently working with electronic documents in this kind of environment. These enablers are:

1. Software that allows convenient interaction with all documents equally.
2. Convenient reliable ad hoc departmental level electronic document storage that replaces library and filing systems.
3. File formats that can efficiently represent all documents without compromising image quality.
4. Simple to use and readily accessible ways to convert any document from electronic or paper format to its complement.

We can examine these enablers in detail.

1. Document Interaction

In traditional software a separate application is used to interact with each document, for example a separate word processing, spreadsheet, or e-mail applications be invoked, depending on the type of the original document. In practical office work process documents from several creation sources are handled simultaneously. Workers avoid using multiple software applications by dealing with documents in paper format, the common denominator. For such office work a single application that can deal with all or nearly all kinds of originals equally is critical. Such an application is frequently called desktop software.

At minimum desktop software should allow the user to:
▶ View all electronic documents without regard to document type.
▶ Add notations
▶ Remove or reorder parts of the document by page
▶ Store the documents conveniently
▶ Combine all or parts of individual documents to a new compound document
▶ Transfer the document by electronic methods (workflow, e-mail, GroupWare)
▶ Print the document
Optionally the desktop software may:
▶ Search documents based on full text and metadata
▶ Display document thumbnails for easy identification within folders
▶ Connect to document scanning systems
▶ Convert between document formats, including OCR for scanned text to word processing conversion

An example of this class of software application developed by Sharp is Sharpdesk®.

2. Document Storage

Enterprise level document management systems are quite complex and highly structured. There are several suppliers of tools and solutions for structured document management. In contrast departmental paper based filing and library systems are frequently ad hoc or designed without benefit of extensive systems analysis. The challenge is to provide the informal design and ease of use characterizing paper filing systems to a departmental electronic filing system.

At the present time there aren't any systems that really meet the needs of a departmental filing system. We can guess at some requirements, but can't validate by examining successful systems. With that proviso we could list requirements as:
▶ Accept documents from all sources directly.
▶ Connect seamlessly to enterprise document management, workflow, and backup systems.
▶ Support user driven file and filing hierarchy, including ready reorganization.
▶ Support user driven access, permission and privacy.
▶ Provide reliable backup and address user concerns about accidental loss.
▶ Support easy user search including by text, metadata, by reference, and visually.
▶ Support document filing and retrieval without intervening application software.

Some suppliers have shown preliminary systems that address this need. This kind of product will become important in the next few years.

3. Electronic Document Formats

Several electronic document formats are already in common use. Unfortunately, no single format meets all the requirements of electronic documents. In general the existing formats fall into two categories

Those formats intended for electronic originals and those formats intended to represent or mimic physical originals.

For the purposes of this discussion we may define document as **organized static information for human use.** By this definition document can refer to books, magazines, newspapers, web pages, e-mail, recorded

sound, video, dictionaries, forms and the like.

A significant limitation is the term static. Documents modified from time to time are still static in each instance and meet the common understanding of a document. Truly dynamic interaction such as with spreadsheets, active web pages, or in fact live television are beyond our understanding of a document.

3•1 Existing Formats

Existing file formats can be used for document storage and exchange in the office environment, but all the common formats have significant limitations. We will examine common classes of file format.

Application Files. These are probably the most common document formats are an important class that will not be supplanted. The disadvantages of application files are:

1. Requires complex custom display software-the application.
2. Document appearance will vary widely depending on application settings.
3. Files tend to be large compared to the human information, and files are frequently platform specific.
4. Information interchange among application using native file formats is usually difficult.
5. Indexing and searching is only possible in some application file formats
6. The number of incompatible file types is very large.

Publishing File Formats. These file types are intended to make a document display and print predictably in any target. Acrobat® and TIFF G-4 are most common. The emerging e-book formats belong to this class. The disadvantages of these formats are:

1. Acrobat® The reader application is complex and requires substantial computing power.

2. TIFF G-4 is only suitable for monochrome documents with simple graphics. The format doesn't inherently support indexing.

3. All As a universal file format these types have one major problem, they formats are not editable. Once a document has been "published" to the file format it can't be modified without first performing a complex conversion process

Image File Formats such as JPEG, are only suited for limited classes of original data. These formats aren't truly suitable for document applications.

HTML As the most common format used for displaying web pages HTML is almost universal. HTML, by supporting new tags, can deal with almost any data format. The shortcoming of HTML is that the appearance of the document can vary widely depending on local display settings. HTML extensions such as CSS (Cascading Style Sheets) haven't completely resolved this problem and are complex to implement.

3•2 Document Format Challenges

There are 2 important needs in electronic document formats that must be addressed by industry:

Universal format An appropriate universal document format is needed that allows predictable display, interchange among applications, editing, searching. This universal format needs to be Internet transmission compatible and should support high quality document display and printing with small computing resource. The format should be an industry standard and should be extensible. Compatibility testing should be available to assure compliance with the format.

This is a lot to ask for a single data format, but this is just some of the goals for XML (Extensible Markup Language) Initial versions of the XML standard and applications that support the standard already exist. Many industry standard groups are working to extend the standard. Organizations concerned with electronic documents should contribute to standardizing features in XML that will enhance the format for universal electronic document use.

Complex Color Physical Document Format Existing formats for representing physical original documents in electronic format aren't adequate for complex color documents. Bitonal standards such as TIFF-G4 and JBIG are excellent for text and linework but can't represent natural color images. Image centered formats such as JPEG are afflicted with objectionable artifacts for text and lines, especially at high compression ratios.

Developing a document format to meet the needs of complex color images is a current industry effort. The ideal complex color image document format would provide variable image quality sufficient for display and printing with no important degradation.

The most promising direction for complex color image coding is mixed raster format. In mixed raster content compression the original image is segmented into foreground, background and mask layers. Each image layer is separately compressed. Typically JPEG or wavelet compression is used for the background, lossless JPEG for the foreground, and JBIG for the mask layer. Examples of

this format include Lucent's DjVu. Standards for this kind of compression may be published by ITU-T within recommendation T.44 or as a part of JPEG-2000

The mixed raster content format offers a significant opportunity for proprietary implementations to deliver faster results with improved image quality that still meets the standard. The critical steps in processing the original image are segmenting the image into foreground, background and mask and in image processing to reduce bandwidth and artifacts. Competitors in the mixed raster content environment are working hard on these issues.

4. Conversion between Paper and Electronic Documents

Finally we look at the conversion from electronic to paper format and paper to electronic format. At a casual inspection printing is not affected by the electronic document processing revolution, but as we shall see the changes in printing requirements are substantial. In the other direction network or workgroup scanning is just reaching the stage where it is possible to understand how it will be used in departmental document conversion.

4•1 Printing

Electronic printing is the most familiar conversion system between electronic and paper documents. Even this familiar technology must undergo substantial changes as a result of electronic document technology and external forces.

One change in the printing paradigm is that paper documents are becoming an accessory display system rather than the primary distribution process. The impact of this change is that single copies of complete electronic documents will be printed as the preferred display mechanism. The user will expect instant and easy access, without regard to the document complexity. Printers and print software should be geared to printing single copies quickly.

We should expect that users could request the printing resource to select the document to be printed directly from departmental document storage. The present method of going to a 3rd place (the users PC) to coordinate printing a document stored on a departmental document server to a printer is not user friendly and doesn't seem logical when described.

To support this kind of operation the printer will have to offer a mechanism for traversing the documents stored and selecting the document or documents to be printed. In addition the printer will have to support direct printing of the most important electronic document formats without intermediate conversion on a PC.

4•2 Print on Demand

As mentioned above we expect that departmental document storage system will act as the departmental library. Despite the rise in electronic means for distribution, paper is still the preferred format for viewing long documents. The combination of these conditions forces the departmental electronic printing resource to meet print on demand requirements, which are:

(1) Direct printing of an electronic publishing document format. Currently Adobe Acrobat® and TIFF G-4 compressed monochrome documents at 200 to 400 dpi are the important standards. It is possible that XML will become dominant in the future.

(2) Printers capable of producing complete copies of original documents - books and the like, in a completely finished form, including covers, bindings tabs etc.

Existing Print on Demand systems are quite complex to operate. Typical interaction is for the user to prepare a request form for the document, which is delivered some time later. In user operated systems the user has to find the appropriate electronic file, launch an application and make the appropriate print settings. Actually printing the document uses significant workstation and print processing resource. In order for print on demand to become an integral part of the office environment, the operation will have to become much easier and more intuitive. This is the real software challenge for print on demand applications

4•3 Color Document Printing

In the past the use of color in the office and home was limited. The primary reason had been that there was no or little color content available. From a printing viewpoint the Internet is a vast source of complex and colorful documents, graphics and images. This source of color data is rapidly driving the color content of personal and office printing.

The rise in color content has an obvious effect on printing hardware. Low-end printers are already all color. We should expect that within a few years all printers will be color.

The effect on software of the increasing use of color is less clear.

4•4 Scanning

The paper to electronic document conversion paradigm for the office is less well developed. Until recently most such conversion took place in application programs, scanners acting as dumb devices directly connected to workstations. Office document scanning has been primarily a centralized activity, with paper documents moved to a scanning facility complete with operators and a complex work process that is analogous to the old fashioned centralized computer printing operations.

There is emerging into the modern office a new class of devices we call the network scanner. These devices provide easy workgroup access, fast and convenient operation, and flexible delivery that are analogous and complementary to network printing. The best network scanning solutions are based on robust and flexible paper handling of digital copiers. As such network scanning tends to be a digital copier option.

To meet the needs of the modern office for conversion to electronic documents need to fit easily into the environment. Goals include:

1. Easy and intuitive use.
2. Works with any kind of paper original, as does a copier, preferably including color documents.
3. Converts to all electronic document formats important for the office.
4. Deliver the document for its intended application, such as e-mail, fax, filing, or to the desktop.
5. Works seamlessly with the office network environment.

Conclusion

There are substantial challenges and opportunities in the transition from paper based to electronic documents as the primary working document in the departmental workplace. Several of the key technologies needed to complete this transition are now in development. Successfully addressing these challenges will offer significant growth opportunities and challenges for office solutions.