

University of Leeds – Institute of Communications Studies
COMM 5010: Communications and Technology

David J. Aumueller¹
MSc in Communications Studies
May 15, 2003

Second Segment Assessment, 13 pages

The technology for the paperless office is only virtually real

Table of Contents

1. Technology Exemplar Selections	2
2. Technology Assessment Report.....	2
2.1. The technology for the paperless office is there.....	2
2.1.1. Scanner with optical character recognition	3
2.1.2. Document server.....	5
2.1.3. Internet Technology.....	6
2.1.4. Handheld PC.....	8
2.2. Integration of technologies to form the paperless office.....	9
3. Technology Opportunity Article	11
4. References.....	12

¹ ics2dja@leeds.ac.uk

Technology Exemplar Selections

Four technology exemplars as key foundations for the paperless office:

- **Scanner/fax server with optical character recognition**
Paper documents need to be scanned and digitized to be stored either as image file or text file after OCR (optical character recognition) process.
- **Storage device with content management**
Document management for indexing and retrieving virtual documents (physically stored on magnetic hard-disks in the server) supporting search algorithms.
- **Network-infrastructure using Internet technology**
Documents residing on the server need to be transported digitally over a network. Internet technology provides standard-protocols for this, e.g. TCP/IP (transfer control protocol/internet protocol).
- **Portable computer with integrated modem (and phone)**
A mobile device may be small as a mobile phone or personal digital assistant (PDA), or big like a tablet PC with a large screen acting also as touchpad for a stylus for input, or may just be a normal laptop – or anything in between.

Technology Assessment Report

1.1. The technology for the paperless office is there

In order to change a traditional paper-based office into a future paperless one, a range of technologies needs to be installed which strongly need to work together to benefit from all its offered possibilities. “Concurrently, both hardware and software technologies are evolving in ways that make it possible to maintain large amounts of information on-line and to have access to this information in conjunction with the communication networks from almost any location with the available distributed databases” (Grauer, 2002: 7476).

Firstly, all available information on paper should be digitized. Although it can be streamlined this will take a reasonable amount of time and probably should be done alongside daily work. More important is to digitize all *incoming* paper-based documents, make them available in the document database and get rid² of the paper. Ideally, all textual information is available as structured text and not only as image data. That way, the texts can be indexed and made available for easy retrieval within

² In practice incoming documents will get archived for a specific time, I assume.

the document system. Also, text files are much smaller in byte-size than even compressed image files. The document server being networked with all workstations and connected to the Internet it is possible to retrieve any document onto the workstation PCs as well as onto other computers connected to the Internet or even onto mobile devices with access to the Internet.

1.1.1. Scanner with optical character recognition

To move from a paper-based to a paperless office documents have to be scanned, digitized and stored as computer files. For documents exclusively containing text it is more than useful to translate the scanned image of the document into plain text due to a variety of reasons. Firstly, it saves a lot of storage space; secondly texts can be indexed allowing better access to specific documents. Since some documents contain signatures or other illustrating images that cannot be translated to text it is important to think of an appropriate file format for image storage. This non-textual information may be “extracted through a segmentation process, stored, and compressed separately” (Baeza-Yates and Ribeiro, 1999: 158).

Concerning incoming faxes (facsimiles) they do not need to be printed out in the fax machine. Instead, the incoming data is already in discrete form and the digital data can be run through the OCR-process and translated into email messages sent straight to the identified recipients.

Analogue to digital conversion of images

Analogue data, such as pictures and text on paper is continuous in three dimensions, namely the two dimensions of the flat space (e.g. x- and y-axis) and the intensity of the printed ink. If there is only black and white as colour 2 bits for this information would suffice, although 8 bits might produce a nicer reproduction due to anti-aliasing. If there are more colours they get divided into their red, green and blue fraction/value.

Figure 1 A prism splits the light spectrum into red, green and blue fractions

To convert analogue signals into digital ones two operations are necessary. The sampling in space divides the whole area of the paper into small discrete areas, so called pixels (picture elements) or dots. Common values for sampling are values between 300dpi (dots per inch) and 2000dpi. Ideally, the sample rate has to be determined using Shannon’s law based on the resolution needed for later processing

of the image (the input frequency of an image is hard to determine). Quantisation of the colour intensity for each pixel is the second operation, usually done with 8 bits per pixel, i.e. 256 different values for intensity.

The intensity is captured using CCD-sensors (charge-coupled device) which are integrated circuits containing an array of linked, or coupled, capacitors. Projecting an image on the capacitor array causes each capacitor to accumulate an electric charge proportional to the light intensity at that location. This value then is quantised. The overall result is a sampled data-stream of digital data that can be serialised into a file. (Lewis, 1997: 144)

Image files, compression

A common file format for high resolution bitmap images is TIFF (Tagged Image File Format) which usually codes the image data uncompressed. A more interesting image file format is JPEG (Joint Photographic Experts Group). JPEG is very commonly employed in many products and used to interchange images on the World Wide Web, for instance.

Most image data contains some redundant data which allows high compression by reducing redundancy. JPEG compression is a set of processes, centred upon a spatial frequency representation of the image data. The image is converted from RGB (red, green, blue) into a different colour space called YUV. The Y component represents brightness of a pixel, and the U and V components together represent the hue and saturation. Each component (Y, U, V) of the image is tiled into sections of 8 by 8 pixels each, then each tile is converted to frequency space using “discrete cosine transform coding” (Lewis, 1997: 146).

Human perception has a much greater sensitivity to intensity than colour. This is exploited to reduce the amount of data. Since our visual sense interpolates and manages without very fine detail some detail can be discarded. Typically an overall data reduction of 20 to 1 or better is achieved. (Hoyle, 2002)

JPEG, like many other compression methods are ‘symmetrical’, i.e. it takes the same time to compress as to decompress.

Optical Character Recognition (OCR)

OCR means the processing/translation of images of typewritten text into machine-editable text. Text is encoded using a standard encoding scheme, usually ASCII (American Standard Code for Information Interchange). Most fonts/typefaces are recognized with a high degree of accuracy. Some systems are even capable of

correctly identifying columns and non-textual images and producing output that places the text and scanned images equivalently in a common word-processing document format. Recognition of handwriting in general is still very difficult although some systems cope with 'trained handwriting'. (Hoyle, 2002)
As indicated above, textual representation rather than image representation has advantages that will be described next.

1.1.2. Document server

All digital documents – be it images, texts or even sound and video data – need to be physically stored on storage devices. To organise the data and be able to retrieve documents again more or less sophisticated file or data structures are available. The most common way to store huge amounts of data are relational databases. Although object-oriented databases once seemed very promising it is the relational paradigm that still is prevalent. Compared to the normal way of storing files on non-volatile memories such as hard and floppy disks, CD-ROMs and DVDs, which is done using a hierarchical structure using directories/folders using standards such as FAT (file attribution table) or NTFS (new technology file system) storing information in databases has certain advantages. Data stored in tables/objects can easily be cross-referenced and it is the database management system (DBMS) that takes care of data consistency and data durability. Most interesting here though, databases can maintain indices of all stored textual information creating means to easily access and retrieve searched-for documents. (Henry, 2003)

Database with document management

Document or knowledge management is necessary to store, index, and retrieve documents using search algorithms. Thus, a 'knowledge base' of electronic documents is built. Another concept to not get lost in huge amounts of information is called information retrieval which denotes the indexing, searching, and recalling of data, particularly of text or other unstructured information. It is the process of determining the relevant documents from a collection of documents, based on a query issued by the user. Especially, methods such as word-stemming and finding similar terms using thesauri help to improve the search result. (Baeza-Yates and Ribeiro, 1999: 1-9)

Non-volatile storage on hard disks

All data needs to reside physically on storage devices, the actual database is a special piece of software running on the operating system (OS) of the server – more and more the free Unix-like system Linux is used as OS. Hard disks are the industry standard for retaining fast access for on-line use by computers. They provide a good compromise between storage volume, access time, transfer rate and cost/bit³. They are cheaper but slower than electronic storage which is volatile memory losing the status of its bits without electronic current. (Hoyle, 2002)

Hard disks are usually single or multiple ‘platter’ disks divided into tracks and sectors. The disks rotate at constant angular speed, spinning continuously. The magnetic read/write head flies aerodynamically without touching the surface of the disk. High density of information can be stored with local magnetic domains being very small. As rough number of storage space around 100 GB (giga-byte = 10^9 bytes) is common for a PC whereas for a document server a size of 1 TB (terra-byte = 10^{12} bytes) might be anticipated – depending on the size of the company and type of documents to be stored – achieved by building a RAID (redundant array of independent disks) of individual hard disks shaping this amount of available space. RAID can be configured to automatically shadow disks in case of a single unit failure. Decisions for the right hard disk have to be based on its access time (in ms), transfer rate (or information flow rate in bits per second), and cost. (Hoyle, 2002)

Figure 2 Platters and read/write heads of a hard disk

1.1.3. Internet Technology

A network-infrastructure using Internet technology for transmission of data connects computers in order to exchange documents residing on the server. Common Internet technology provides with its standard the protocols for this packet-switched network which regulate the transmission. The TCP/IP (transfer control protocol/internet protocol) provides an interface between the physical network layer below and the application layer above – similar to the ISO OSI (open system interconnection) reference model. (Irvine and Harle, 2002: 5)

TCP/IP slices every stream of data into small data-packages. These packages contain information about their destination, e.g. the IP-address of the requesting computer.

³ Prices for 1 GB are down to 1 Euro, e.g. for 120 GB disks.

The protocol guarantees that every package sent from one machine arrives at the other without loss. TCP uses a number of mechanisms to achieve this robustness/reliability and high performance. These include using sequence numbers for ordering received TCP segments and detecting duplicate data, checksums for error detection and acknowledgements and timers for detecting and adjusting to loss or delay, which can easily occur since each package may take a different route over various computers called Routers. At their final destination the packages are put together to files again. (Irvine and Harle, 2002: 81-123; Brownlee and Claffy, 2002)

Internet applications

Common applications act on top of the TCP/IP layer, e.g. HTTP (HyperText Transfer Protocol) for the Web and FTP (File Transfer Protocol) for transferring larger files, or telnet and SSH (Secure Shell) to login to a remote computer/terminal and SMTP (Simple Mail Transfer Protocol), POP3 (Post Office Protocol version 3) or IMAP (Internet Mail Access Protocol) for email.

All these applications can be used to access documents from either the document server (e.g. via FTP, HTTP) or directly from other users on the network/Internet via email.

However, TCP is not appropriate for applications that need especially continuous delivery like real-time applications. These often do not suffer from some loss, errors. For example real-time streaming multimedia is better controlled by the user datagram protocol (UDP) (Irvine and Harle, 2002: 101). In the context of the paperless office video-conferencing would be a technology exemplar, since some issues can be discussed with distant partners on-line without the need to write down elaborate reports.

Local and wide area networks

Connecting a group of PCs locally using this Internet technology, e.g. within an office, one speaks of a local area network (LAN). Workstations and servers are physically connected using Ethernet cables, but also wireless technology becomes more and more common for in-house networks. This might bring some security problems with it, though. If the network is not configured properly with a firewall people from outside the company may succeed in logging into the firm network using their laptop or handheld PC and use the 'free' Internet access, or worse, spy for confidential data.

A wide area network (WAN) is a computer network covering multiple buildings, often across the world. The Internet is the best example of a WAN. WANs connect local area networks together, so that users and computers in one location can communicate with users and computers in other locations.

LANs are typically faster than WANs to enable quick transmission of larger documents inside the company. (Irvine and Harle, 2002: 65)

1.1.4. Handheld PC

Modern business people often have to travel widely and exchange electronic messages and other documents on the move (Hoyle, 2002). The mobile equipment they use can quite vary – it could be a normal laptop connected to the Internet or the network of the office via a mobile phone, or a future third generation (3G) phone with built-in multimedia support. Thus, it is possible to read, edit, and save documents from and to the company's database server.

Connectivity

Mobile phone technology is divided into generations, with current technology in use being mainly second generation (2G) phones using GSM (Global System for Mobile communications) which is the world-wide standard for digital wireless mobile phones. The so called 2.5G indicates technology between the second and third generation (3G). 2.5G uses GPRS (General Packet Radio Service) in the TDMA-based (Time Division Multiple Access) GSM system to provide high-speed packet switched data transmission. Timeslots in GSM is normally allocated to create a circuit-switched connection, whereas the GPRS timeslots are allocated to the packet-connection on a need-basis, freeing up frequency bands for other users. Transfer rates in GSM are 9600 bps (bits per second).

The third generation mobile phone technology uses the Universal Mobile Telephone System (UMTS) technology, which supports up to 2 Mbps (Megabits per second) data transfer rates. (Irvine and Harle, 2002: 223-227)

With such high transfer rates it is in effect possible to send any type of media stream to the mobile device, enabling for example mobile videoconferencing. Either, the mobile phone itself provides the ability to decode the incoming data or the phone is linked up with another handheld device such as a personal digital assistant (PDA), handheld PC, or laptop. For this short distance link between phone and computer a

cable or the Bluetooth⁴ technology could be used, the latter being a communication standard primarily designed for a distance of up to 10 meters and low power consumption. Battery life still is a major problem in mobile devices (Savage, 2003: 46).

Input interface

Depending on the size of the mobile device/computer, various means as input are possible. A tablet PC will have an on-screen touchpad on which the user can write with a stylus pen or 'press' the keys of a 'virtual' displayed keyboard and click on icons. Handheld PCs usually have a smaller version of a normal laptop keyboard, whereas at mobile phones only an array of 3 by 4 keys is common. Incorporating more or less intelligent dictionaries it is possible to type in words with pressing buttons used for multiple (3 to 4) letters only once. This method is named T9 which stands for "Text on 9 keys". (Savage, 2003: 45)

Figure 3 Examples of modern handheld devices integrating phone and PC

1.2. Integration of technologies to form the paperless office

The integration of the described technology exemplars to form the paperless office of the future is seamless as indicated already explaining each technology. Paper-based documents get scanned and maybe translated into a text format that can be indexed in the database. Either as text or image data, the document is stored on the database which allows retrieving the information from connected computers. These workstations or remote computers – also mobile devices – connect to the server via the local area network or via the Internet from outside the company using wide area networks. Thus, even being on the move documents can be reviewed, changed and sent back to the office or to other business partners around the world as mentioned in the introduction.

⁴ Bluetooth is named after the Danish king Harald Blatand (Blue Tooth) II of Denmark, who united the Scandinavian countries.

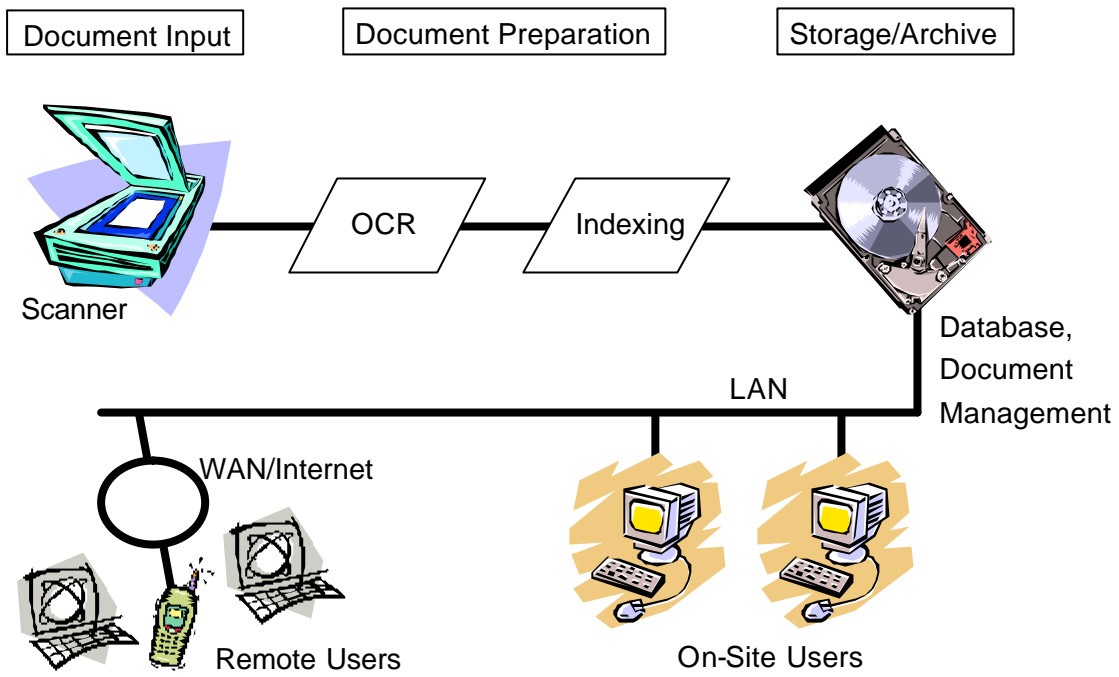


Figure 4 Integrated technology exemplars to form the paperless office

Technology Opportunity Article

When Tim Berners-Lee invented the World Wide Web in 1990 as collaborative virtual work-environment he made the first step to remove the big file-cabinets from our offices. Using already common Internet technology he established a now very popular way of organising electronic documents – maybe even too popular. The Web became home of more documents a human or computer can ever overlook and much information provided proves is useless...

Gladly, there are now more sophisticated document management systems running on databases that can host all the old ‘paperwork’ in new digital format. These reliable storage systems allow easy retrieval of once lost documents in huge filing cabinets. After scanning the old paper-based documents and turning them into textual representation using streamlined optical character recognition (OCR), all documents available in one company are stored and indexed in the database. A simple keyword search delivers the one document the user is looking for. Hopefully! If OCR is not possible due to the sensible content of the document it still can be stored as compressed image JPEG-file. But any contained text is not available for the search algorithm then.

With the Internet technology connections from anywhere in the world can be established to your home office. Wherever you go – you do not need to remember which articles and reports you have to bring with you, they are all available from the document server in your office. With new mobile devices work really becomes portable – one stroke with the pen on the screen and you are connected to the company’s knowledge base. Any document opens in its appropriate application to edit and save it back to the server. New 2.5th and third generation phones make it even possible to discuss in videoconferences with business partners around the globe to clarify the last open issues... Can I never close my eyes on the train and have a nap? The excuse of no network coverage will not work anymore – is this the surveillance society we are waiting for since 1984? Not yet...

Although all these technologies are basically available nobody really cares about using them to that extent. The biggest step towards a paperless office is to learn how to break old ‘analogue habits’. Paper is not system dependent and does not require special equipment or software to access its information. Some bits represented physically as magnetic fields cannot replace the touch of paper which naturally proofs the existence of the information *at hand*, making it seem more valid. Human computer interfaces lack usability; taking a note still is faster on a plain sheet of

paper instead of having to open the right application and key in the information. The idea might be lost already before it is noted down. Technology based forecasts of the paperless office “ignored the habits and values of readers and writers and the countervailing technical advantages of paper, such as the relative difficulty of reading an electronic screen” (Dutton, 2002: 2483).

One should not be too pessimistic though – information technology is at least ready for a ‘less paper office’.

References

- Baeza-Yates, R. and Ribeiro, B. d. A. j. N. (1999) *Modern information retrieval*, Addison-Wesley Longman, Harlow.
- Berners-Lee, T. (1990) *Design Issues for the World Wide Web*, <<http://www.w3.org/DesignIssues>>, Accessed 08/05/2003.
- Brownlee, N. and Claffy, K. (2002) Understanding Internet Traffic Streams: Dragonflies and Tortoises, *IEEE Communications Magazine*, 10, 110-117.
- Cherry, S. M. (2002) Weaving a Web of Ideas, *IEEE Spectrum*, (9), 65-69.
- Crowley, D. and Heyer, P. (2003) *Communication in history : technology, culture, society*, 4th ed., Allyn and Bacon, Boston, MA.
- Dusch-Feja, D. D. (2002) Communication: Electronic Networks and Publications. In *International Encyclopedia of the Social & Behavioral Sciences*.
- Dutton, W. H. (2002) Computers and Society. In *International Encyclopedia of the Social & Behavioral Sciences*.
- Grauer, M. (2002) Information Technology. In *International Encyclopedia of the Social & Behavioral Sciences*.
- Henry, W. (2003) *CoOL : Conservation OnLine : Resources for Conservation Professionals*, <<http://palimpsest.stanford.edu>>, Accessed 08/05/2003.
- Hoyle, B. S. (2002) Information Representation, Storage, Retrieval and Processing. In *Lecture Notes School of Electronic and Electrical Engineering*, University of Leeds.

- HP (1997) *Digital Modulation in Communications Systems : An Introduction*, Hewlett Packard Company.
- Irvine, J. and Harle, D. (2002) *Data communications and networks : an engineering approach*, J. Wiley, Chichester.
- Langley, G. and Ronayne, J. (1993) *Telecommunications primer*, 4th / ed., Pitman, London.
- Lax, S. (1997) *Beyond the horizon : communications technologies : past, present and future*, University of Luton Press, Luton.
- Lewis, G. (1997) *Communications technology handbook*, 2nd ed., Focal Press/Butterworth-Heinemann, Boston, Mass.
- McFarland, W. J. (1998) Wireless Communications: A Spectrum of Opportunities, *The Hewlett-Packard Journal*.
- Naughton, J. (2001) *Contested Space: the Internet and Society*, <<http://molly.open.ac.uk/21c>>, Accessed 08/05/2003.
- Savage, P. (2003) The Perfect Handheld : Dream On, *IEEE Spectrum*, 1, 44-46.
- Zimmer, D. E. (1999) Das grosse Datensterben. Von wegen Infozeitalter: Je neuer die Medien, desto kuerzer ist ihre Lebenserwartung?, *Die Zeit*, 18/11/1999, Hamburg.