1 The Difference between Electronic and Paper Documents

In principle, electronic discovery is no different than paper discovery. All sorts of documents are subject to discovery electronic or otherwise. But here is where the commonality ends. There are substantial differences between the discoveries of the two media.

The following is a list of discovery-related differences between electronic documents and paper ones. We assume that a paper document is a document that was created, maintained, and used manually as a paper documents; it is simply a hard copy of an electronic document.

1.1 The magnitude of electronic data is way larger than paper documents

This point is obvious to the majority of observers. Today's typical disks are at several dozens gigabytes and these sizes grow constantly. A typical medium-size company will have PC's on the desks of most white-collar workers, company-related data, accounting and order information, personnel information, a potential for several databases and company servers, an email server, backup tapes, etc.

Such a company will easily have several terabytes of information. Accordingly[1], such a company has over 2 million documents. Just one personal hard drive can contain 1.5 million pages of data, and one corporate backup tape can contain 4 million pages of data. Thus the magnitude of electronic data that needs to be handled in discovery is staggering. In most corporate civil lawsuits, several backup tapes, hard drives, and removable media are involved.[2]

1.2 Variety of electronic documents is larger than paper documents

Paper documents are ledgers, personnel files, notes, memos, letters, articles, papers, pictures, etc. This variety exists also in electronic form. But then spreadsheets are way more complex than ledger, for example. They contain formulas, may contain charts, they can serve as databases, etc.

In addition to all added information, e.g. charts, ability to view the actual computations involved, e.g. formulae, the electronic spreadsheet supports experimentation with *what-if* version the discoverer may want to investigate.

Personal digital assistants, pocket PCs, palm devices or BlackBerry devices, are subject to electronic discovery. Many of these devices can be

---

[1] High-Risk Insurance Company Reduces Risk Of Losing Documents, Business Solutions, March 1998, http://www.businesssolutionsmag.com/Articles/1998_03/980324.htm

[2] Linda G. Sharp, The complexity of electronic discovery requires practitioners to master new litigation skills, Los Angeles Lawyer, October 2005, Vol. 28, No. 8.

used to send and retrieve e-mails. Since an e-mail deleted from a network may still exist on an individual employee's PDA, parties may demand discovery of the contents of PDAs.

1.3 Electronic documents contains attributes lacking in paper documents

Computers maintain information about your documents, referred to as "metadata," such as: author's name, document creation date, date of it last access, etc. A hard copy of the document does not reveal metadata, although certain metadata items may be printed. Depending on what you do with the document after opening it on your computer screen, the actions taken may change the metadata collected about that document. Paper documents were never this complex.

1.4 Electronic documents are more efficient than paper documents

Paper documents are delivered by mail and stored locally in filing cabinets[3]. For multiple users to access documents simultaneously one needs a set of documents per each accessing person. File cabinets are bulky and use up valuable office space. Paper documents are difficult to search, carry, copy, and modify. Paper documents are easily damaged, misfiled or misplaced.

Electronic documents are delivered by networks, disks, flash memory and CD/DVD and are stored on a file system. Multiple users can read and review electronic document simultaneously. Computer file systems are getting smaller and contain more data every year. Personal file systems are physically smaller than a small cell phone; only very large companies need massive file systems that occupy a lot of real estate. It is almost too easy to search, carry, copy, and modify electronic documents. Electronic documents, in a well run operation, have copies and damage to a single copy causes extra work but no loss.

1.5 The structure of electronic documents may reach complexity absent from paper documents

A description of the structure of an object (i.e. document) identifies its component parts and the nature of the relationships between those parts[4]. Describing documents (i.e. objects) this way points to the complexity of electronic documents.

The following list shows some aspects of the complexity of electronic documents.

- An electronic document may consist of subdocuments that do not even have to reside on the same computer.

---

[3] Content Management, Ryerson University's Open College unit, Xerox Process Study, 5/30/2001.

[4] Pete Johnston, Document Structure, in Effective Records Management Project, UNIVERSITY OF GLASGOW, May 1998. http://www.gla.ac.uk/infostrat/ERM/Docs/docstr.htm#Heading4

- An electronic document may be written in HTML and displayed by a Web browser. Such paper documents don't even exist.
- An electronic document may have hyperlinks[5] to other documents.
- An electronic document may have dynamic parts. This holds in an obvious way to Web pages that can include (and invisible to the reader) programs in languages such as JavaScript and ASP, which run in the browser and the reader views only the results of these runs. It may also hold to text document or spreadsheets as a result of embedded executable fields, functions and macros.
- Electronic documents have much wider spectra than paper ones as they can include not only the classic, word-processed text but tables, databases or a part thereof as well as image, voice and video.

1.6 Electronic documents are more persistent and more difficult to destroy than paper documents

Paper documents are easy to destroy. Throwing away or shredding makes paper documents disappear. Deleting an electronic document eliminates only the ubiquitous accessible copy. The document, i.e. its data, still exists and in systems such as Windows and Mac OS, an accessible reference to deleted documents may be in the trash bin. Restoring a document in the trash bin, i.e. a deleted document, revives the document to its original glory.

Even removing the document from the thrash bin does not erase the documents data off the disk. Once removed from the thrash bin, documents data areas on the disk go into a "fee list" that makes those areas available for future data creation needs. The free list contains all areas not currently allocated to active documents as well as to deleted documents still in the trash bin. How long will an area stay on the free list (thereby still containing the deleted documents data)? That is difficult to predict due the huge variability of factors such as: future demand for disk space, size of current and future files, the current availability of disk space, etc.

Even the complete deletion of a document, its trash bin instance and the allocation of the document's data area on the disk does not typically extinguishes the document altogether. Several practices create copies of documents and are only marginally affected by document deletion:

- Backups – most organizations and individuals create back up copies of documents as regular practice as precautionary actions. The backups are maintained independently of the document itself.

[5] James H. Pence, How to Do Everything with HTML, McGraw-Hill Osborne Media; (May 22, 2001)

- Documents may be exchanged by email, access through web pages and manually handed electronic copies. Thus created copies continue to exist after the deletion of the original document.
- Even work on a simple text document is quite frequently preceded by creating a copy of the document being edited. Once again, such copies persist beyond the deleted document unless specifically deleted.

1.7 Electronic documents change faster, more frequently and easier than paper documents

Changes to an electronic document are fast and easy. The reason is obvious; all you need to do is make the change and save it. Changes to paper documents, however, require retyping the whole document.

There are many other reasons to the difference in speed and frequency. We already said that documents may be dynamic. Web pages are made dynamic in order to ease change.

For discovery, faster and frequent changes imply a need for a more meticulous and length monitoring of document discovery.

1.8 Electronic documents last longer than paper documents

Paper deteriorates with time; paper documents can be destroyed by flood and fire. Although these factors have their parallels in electronic documents, e.g. a flooded computer loses its data; typical backups of the documents practices maintain copies away from the “office.” Paper documents may enjoy the same treatment, but the frequency, extent and usage of such backups is substantially smaller.

Electronic document suffer from upgrades in technology. If one used a peculiar word processor, e.g. WordStar, to write a document ten years ago, today it will be difficult to find a tool to read that document. Same holds for spreadsheets, databases, etc. Again, most companies have practices that avoided such problem by evolving documents with time.

1.9 The redundancy in electronic documents is higher than in paper documents

There are several levels of redundancy to electronic documents.

- Due to the type of recording used for electronic data, minor errors in a document can be corrected by computer tools. The tools rely on the redundancy of checksums and other devices
- Due to frequent changes in documents, individuals learn to save previous versions of the documents. Doing that generates redundancy of document versions.
- Emails, flash memories, CDs all proliferate documents and result in high redundancy
- Most companies and many individuals make backup of documents

- Tool that control versioning of files create built-in redundancy

1.10 An electronic data is more likely to be created by several individuals than a paper document

MS Word supports "Document Collaboration.[6]" Where this term implies: "new objects, properties, and methods of the Word 10.0 Object Library shown in this article allow you to change the display of revisions and comments, accept and reject revisions, and start and end a collaborative review cycle."

Another tool, Workshare 3[7], is an add-on to Microsoft Word that manages collaboration on Word documents and integrates this activity with email and the organization's document repository tool.

Collaborations on databases (e.g. people using a bank's ATMs update the bank's database), spreadsheets (e.g. BadBlue[8]), and Web sites are commonly practiced.

This dwarfs collaborations on paper documents possible.

For discovery it implies that the author of a Word document may not be the only person involved in writing the document. One has to determine all the parties that collaborated on the document.

1.11 Electronic documents may be created by electronic means while paper documents are created by humans

Paper documents are always written by human beings. That is not necessarily the case with documents. We start with a simple, and rather common, example. The Quicken financial program will generate financial reports from a database of financial transactions.

---

[6] Lisa Wollin, Creating Custom Solutions for Document Collaboration, Microsoft Corporation, April 2001, Applies to: Microsoft® Word 2002.

[7] Martin Langham, Closing the Collaboration Gap, IT-Director.com, September 2003. http://www.it-director.com/article.php?articleid=11205

[8] BadBlue Excel Web Sharing FAQ, http://www.badblue.com/helpxls.htm

Category Description	7/1/2004 Actual	- Budget	3/31/2005 Difference
INCOME			
Advertising Website	0.00	0.00	0.00
Educational Activities	20,400.00	20,000.00	400.00
Interest Inc	282.41	180.00	102.41
Other Programmatic Activities	0.00	0.00	0.00
Other Revenue	0.00	0.00	0.00
Rebates	693.00	650.00	43.00
TOTAL INCOME	**21,375.41**	**20,830.00**	**545.41**
EXPENSES			
Admin and Oper Exp	2,782.62	2,547.00	-235.62
Advertising Expenses	0.00	0.00	0.00
Educ Act Expenses	13,625.00	13,280.00	-345.00
Other Exepnses	0.00	0.00	0.00
Other Programmatic Expenses	0.00	0.00	0.00
Scholarships	1,500.00	1,500.00	0.00
Website Related Exepnses	0.00	0.00	0.00
TOTAL EXPENSES	**17,907.62**	**17,327.00**	**-580.62**
OVERALL TOTAL	**3,467.79**	**3,503.00**	**-35.21**

This is an application created document.

Using MS Word and its Autosummarize tool on a large document we got:

Patient Monitoring Techniques in Telemedicine

Through the leverage of these devices we can formulate distributed algorithms and create effective data structures to properly monitor patients. Every patient will have very specific needs and we need a real time system to properly monitor the status of every single patient.

Each individual patient will be uniquely identified with a combination of building, floor, room, and patient id. ***Senior Citizen Patients Monitoring Tree***

Lastly, each room contains one patient.

The objects could be customized to contain all pertinent monitoring information of each respective patient. Our goal is to formulate a Medical Object Query Language (MOQL)
The medical devices can interface with each object api to continuously update each patient object (MP). ***Research Goals***

The tool created the document within the box. In this case, discovery has to find the person that wrote the original document. That is not needed with paper document.

1.12 Electronic discovery requires support of an infrastructure that paper discovery has never needed

The large volumes of data, its complexity, its variety of electronic documents have brought about many types of computer tools to help overcome the obvious difficulties.

Socha Consulting[9] provides the following entries in its Tools section (we drop the commercial part and use just the generic description):

- Electronic discovery software; allows users to evaluate and manage electronic documents
- Automated litigation support software; allows users to organize, search, and retrieve e-mail with attachments
- Open, view, print and convert various files types
- Review, acquire and analyze digital information on individual machines or across a wide-area-network
- View and access contents of various file types
- Automated litigation support software; allows users to process electronic files

In the chapter dedicated to ED tools, we will discuss tools in a generic way and demonstrate their functionality.

2 Document Discovery in the Electronic Age

2.1 Introduction

Electronic Discovery (ED) describes the process of identifying, locating, securing and producing electronic data for the purposes of obtaining evidence for civil or criminal litigation.

The discovery of electronic data has "become commonplace in litigation.[10]" That manual states that a discovering party "must be able to learn the underlying computer theories and preparation procedures in order to understand the meaning of computer information, particularly if it will be used at trial."

More than 90% percent of the world's data is in electronic format. Most of these electronic documents are never printed[11]. Reviewing electronic data, therefore, is essential to conducting comprehensive discovery—otherwise most of the potential evidence is left unexamined.

2.2 Discoverable Data Sources and Types

Electronic evidence has been broadly defined as "any electronically-stored information subject to pretrial discovery." .

---

[9] http://www.sochaconsulting.com/tools.htm

[10] David F. Herr, Annotated Manual for Complex Litigation, 4th, 2005 ed. Thomson-West.

[11] Andre Guilbeau, Getting a Grasp on Electronic Discovery, Houston Business Journal, Vol. 34, No. 49, 2004.

2.2.1 Data Context of a Generic Company

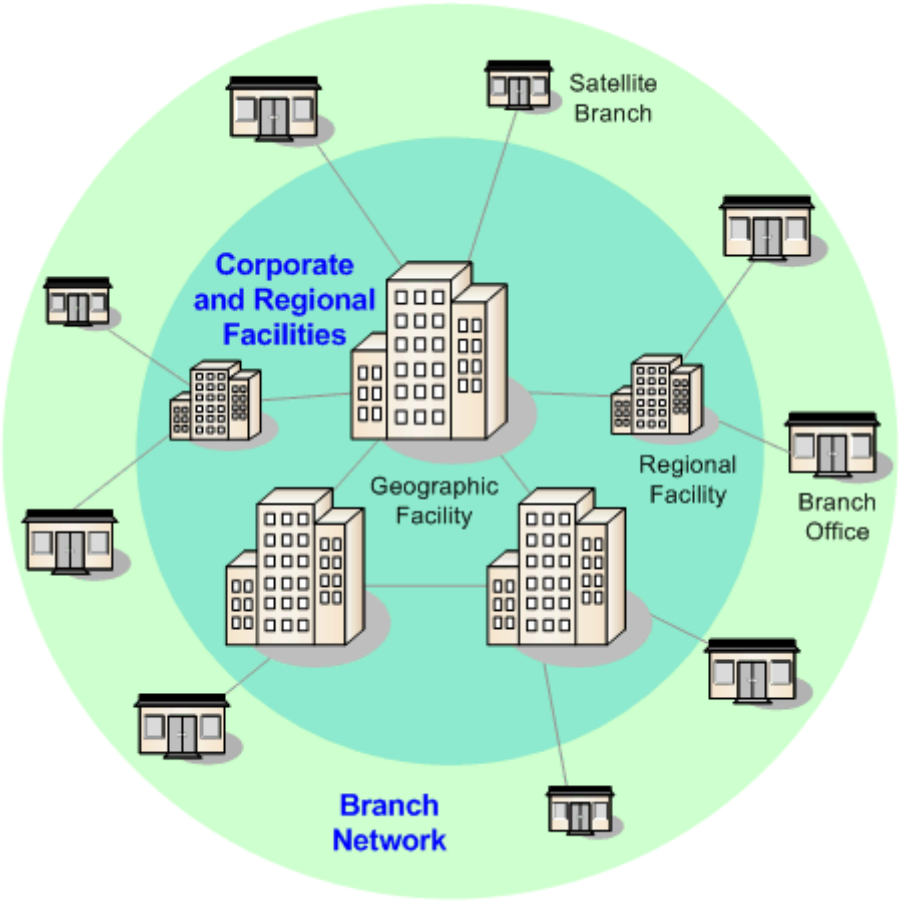


Figure 1 – Company's Organization Structure

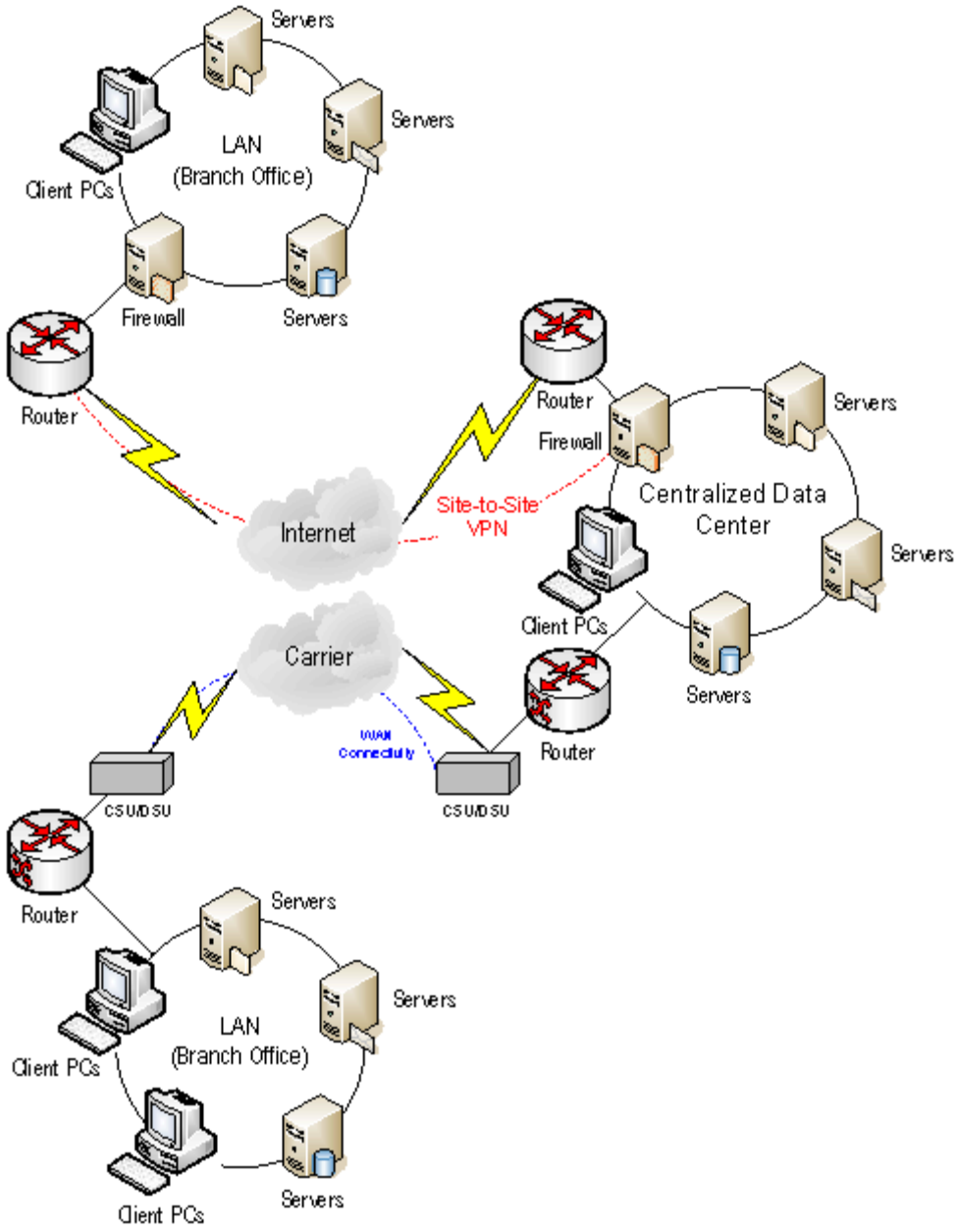


Figure 2 – Company Network

The typical company office will have several servers, a wired local area network (LAN), several PCs, printers, and scanners. Many will now also have a wireless network with several hot points.

2.2.1.1 Servers

According to Wikipedia[12] a servers is (references are ours):

[12] Nora Miller, Wikipedia and the disappearing "author".(piece of writing) : January 1, 2005, International Society for General Semantics, Volume: 62 Issue: 1 Page: 37(4), and http://en.wikipedia.org/wiki/Main_Page

> "A computer software application that carries out some task (i.e. provides a service) on behalf of yet another piece of software called a client. In the case of the Web: An example of a server is the Apache Web Server[13], and an example of a client is the Mozilla Web Browser[14]. Other server (and client) software exists for other services such as e-mail, printing, remote login, and even displaying graphical output. This is usually divided into file serving, allowing users to store and access files on a common computer; and application serving, where the software runs a computer program to carry out some task for the users. This is the original meaning of the term. Web, mail, and database servers are what most people access when using the Internet."

The typical office now supports an Internet server that connects the office to the outside world and may offer some externally accessible office pages. Most offices have an e-mail server (e.g. Microsoft's Exchange Server[15]). Other servers than may be found are: a database server (e.g. Oracle[16]), an application server (e.g. IBM's WebSphere[17]), and a file server (e.g. Linux[18]), etc.

The servers may share computers or may each have a whole computer for them. Each of these servers is a separate computer application with its own complexity, infrastructure, configuration, file and data.

2.2.1.2 Local Area Network (LAN)

The LAN consists of cables that are the wires through which the data is communicated, a router - a device that forwards data packets toward their destinations through a process known as routing – switches - connect different types of network segments together to form a heterogeneous network - hubs and the computer to which the LAN is connected.

Although LANs are used to transmit data from one computer to another, they still have some data content. There are two main data storage facilities with discovery implications. Routers and switches may have three data components:

- Software modules
- Configuration parameters
- Statistical data

Higher-end routers and switches can be loaded with software by the user. This software may be bought or home-made and may be of interest in discovery. Routers and switches have to be configures to the environment where operate. For that purpose, the local network administrator sets parameters that make the network operate as required. In well

---

[13] Katherine Wrightson, Kate Wrightson, Apache Server 2.0: A Beginner's Guide, Osborne/McGraw-Hill; 1st edition (September 5, 2001)
[14] Nigel McFarlane, Rapid Application Development with Mozilla, Prentice Hall PTR; 1st edition (November 7, 2003)
[15] Scott Schnoll, Microsoft Exchange Server 2003 Distilled, Addison Wesley Professional, 2004.
[16] Kevin Loney, Oracle Database 10g: The Complete Reference (Osborne ORACLE Press Series), McGraw-Hill Osborne Media; 1 edition (May 5, 2004)
[17] Tim Francis, Eric Herness, Rob High, Jr., Jim Knutson, Kim Rochat, Chris Vignola, Professional IBM WebSphere 5.0 Application Server, Wrox-Wiley December 2002.
[18] Brian Bilbrey, Linux Transfer for Windows Network Admins: A Roadmap for Building a Linux File Server, Hentzenwerke Publishing (November 1, 2003)

run companies, such parameters, their justification, previous settings and related information are kept as paper or electronic logs. Routers and switches continuously monitor the network and collect traffic statistics. These devices have predefined means to retrieve such statistics form a controlling computer and make them available. Statistical data may be significant for discovery. For instance, huge peaks of traffic on certain days may indicate activity counsel may be looking for.

Storage Area Network (SAN): is a high-speed special-purpose network that interconnects different kinds of data storage devices with associated data servers. It is predicted that SAN will replace file servers, at least, in large enterprises and while they are not referred to as 'file servers,' their function, though not performance, is that of file servers and may contains large amount of data.

2.2.1.3 Personal Computers and Workstations

Data found on personal computers and workstations that are dedicated to specialized tasks, e.g. graphics computer for better drawing, have all the typical types discovery is interested in. A detailed discussion of these types can found below.

2.2.2 Data Types

We divide discoverable data into two distinct groups. One group is the file system data. Basically, this group is the electronic version of the folders and files we had in the file cabinet and the drawers. True, the electronic version expanded the complexity of the documents, their volume and their variety. Yet, by and large a ledger is now a spreadsheet (now with built-in calculations, use of statistics and math, etc.). The same holds for memos, articles, pictures, etc.

The second group is the new guy on the block; it is data we did have when we used just paper. This data is organized differently than the old file cabinet. The prominent data in this group is Email. Email just resists the neat organization of a file system. It just piles up as time pass; frustrating not only lawyers, but also administrators that manage it. There are other types of data that do not tow the line.

We will call the first type of data structured data and the second type of data piled data.

2.2.2.1 Structured Data

2.2.2.1.1 Active Data

Active data are the files on the computer at time of discovery. Most such files can identified by listing the content of directories, themselves a type of a file, and may be associated with applications as input, output, parameters, configuration, measurements, etc. Active data may be examined by a "discoverer" using simple text viewer (e.g. word processors), or may need to be examined using an application (program). Common applications are e-mailer (Outlook is probably the most used of the lot), a spreadsheet application (e.g. excel), a database manager (e.g. Oracle), a Browser (e.g. FireFox), a Weblog aggregator (e.g. SharpReader), etc.

Active data may be identified and its associated application known, but the data itself may be protected and its content will not be visible with getting permission to get through the protection. Active data may be password protected or it may be encrypted; in both cases, and especially in civil discovery, further progress with the file examination requires the consent of the owner in the form of the encryption key, password or the production of files that are freely accessible.

Among the files on the computer at time of discovery there are many data item that are only indirectly accessible. These items are removed files that still reside within the recycle bin, history files, temporary files, buffered data, Browser caches, cookies, file caches, system registry files (if the computer has a registry), operating system logs, performance logs, etc.

Most of this type of data is not pertinent to the case on hand. For instance, in a patent infringement case the Browser cache on the secretary's PC is probably of little use for the opposing counsel, but in an harassment case, such a cache may contradict a statement such as "I was out of the office last week."

The variety of active files is unbounded. Different computers, e.g. Mac and Windows, will have their additional active files types, either through the operating systems or the different applications they run. Different companies will run different applications and will have different types of files. Even different versions of the same application, e.g. MS Word 98 and MS Word 2003, may have different file types. This list does not end.

One needs an expert to navigate the discovery of such variety and a single expert will not do. Different applications/cases may require different experts. It will be wrong to assume, for instance, that files can always be reduced to some visual representation either textual, graphical or even using sound and motion.

2.2.2.1.2 Archival Data

Archival data is information copied to an off-line medium. (Off-line is the opposite of online which is described as: something is said to be online if it is connected to some larger network or system). Most businesses have their computer networks backed up, i.e. archived, on a regular schedule. Archival data of a time frame provides a historical record of the active data on this time frame. A backup is nothing more than a snapshot of the data being backed up. It is a reflection of that data at a particular moment in time.

Archival data is backed up to: tapes, cartridges, CDs, network servers, network storage or even the Internet. Typically, backups of electronic records have been used for making copies of files available for disaster recovery when the original files are not available due to damage.

There are several types of backups[19] (and therefore, destinations for discovery):

- Full Backups - every single file is written to the backup media
- Incremental Backups - if a file's modification time is more recent than its last backup time. If it is not, then the file should be backed up
- Differential Backups - once a file has been modified it will continue to be included in all subsequent differential backups until the next full backup

Backup and restoration of active data has been around almost from the very beginning. Thus, this process is well understood, relatively efficient and widely practiced. Most computer facilities have a cyclic schedule of backup tasks whose main purpose is to create a complete copy of, mainly, users' active data.

The reasons archival data is created are varied. The more frequent reasons are listed below:

- Need to go back to an earlier file version because of a mistake made when modifying the file
- A user deletes a file by mistake
- A computer virus or worm destroys files
- Hard drive failures
- Disaster recovery

Due to the size of many enterprises and the large amount of active data, backup of mission critical data has become a big business. Companies offer off-site backup processing and facilities with guarantee of restoration in case of need for recovery[20].

Archival data can be staggeringly large, particularly in large enterprises. Currently, the tapes used for backup are almost a terabyte in capacity (LTO tapes under the name "Ultrium[21]"). Furthermore, until late 2003 close to 90 Million such tapes have been shipped[22]. Clearly, these tapes can accommodate an astonishing amount of data.

The archival copy of an active data file does not fully reflect all of the data that can be identified and extracted from the active file.

---

[19] Red Hat Linux 9: Red Hat Linux System Administration Primer, Chapter 8. Planning for Disaster, 2003 by Red Hat, Inc.
[20] Ira Gupta, Data-recovery Plans Can Avert Disaster, ITAudit, Vol. 7, November 1, 2004.
[21] Wikipedia, Linear Tape-Open entry.
[22] Krishna Kumar, Storage & Backup, State Of The Mart, PC Quest, September 17, 2003.

Individual applications sometimes need to be restored from backup. This backup can easily be botched and the application may not be restorable. On the Windows XP system there is an operating system file called the Registry which may have to be backuped as well, but backup and restore of the Registry is not straightforward[23].

2.2.2.1.3 Replicant Data

Replication of data occurs for several reasons and for several different purposes. An automatic backup, which most companies and agencies use, is executed on a regular basis. This process creates and stores archival data in the computer network which it is responsible for. Replicant data, when viewed from the archival data aspect, refers to side effects of the archival rather the whole volume of data. The reason archival data may also be Replicant data multi-faceted.

- By definition, a backed up file is a replica of the file. Such a replica has its own life and may reappear in many places
- Archival data may contain multiple copies of the same file. For instance, two successive full backups will maintain 2 copies, at least, of each file that was in the system at both backups
- With time, the metadata (see later in the chapter) of a backed up file changes with necessarily changing the content of the data resulting in the original and the replica not being identical anymore.
- Deleted active files may still be fully "alive" on the replica

Typical examples of places replicant data can be found are printer buffer memories, which store unsaved data that is sent to a printer, and backup tapes, which hold "information copied to removable media in order to provide users with access to data in the event of a system failure."

To improve the performance of computer systems the operating system may decide to replicate files whose size and amount traffic, i.e. reads and writes. Replication, in this case, is done to make files available "locally," located in the network near the user of the file, thereby reducing traffic and avoiding bottle neck. (Disney has a park in California and a Park in Florida, had it stayed with the original park in California all customer would have flown, stayed and visited a single park; Disney reduced the traffic in California and made visits more pleasant.) Users of implicitly replicated files are not aware of files existence and have no control over them. Replicant data copies are subject to discovery and may exist after the original file is long gone.

Other sources of Replicant data are:

- Email attachments – files attached may exist in email folders and in personal directories
- Personal copies made for ease of use

[23] Peter Hipson, Mastering Windows XP Registry, 2002 SYBEX Inc., Alameda, CA.

- Leftover files that were restarted work in different folders
- Multiple media (e.g. a copy on PC disk and a USB flash card)
- Some applications are designed specifically to maintain a complete history of documents. Such application goes under the generic name *Version Control*. (The name implies coexistence of several *versions*.) For instance, NextPage[24] provides version control for Excel, PowerPoint and Word documents. It also integrates with Windows Explorer and Outlook. There are other products[25] such as the latter.

Another important aspect of replicant data is transaction logging. Applications such as word processing programs (MS Word or Wordperfect), databases (Oracle, MySql), and firewalls[26] use transaction logging. Logging allows restoring previously applied transactions. In this respect, it is a Do/Undo mechanism although; it can be much more than just an potential Undo list. Because any Do/Undo mechanism creates a log of what was already done, it is a replica of other information.

As illustrated by this example, discovery of replicant and archival data (e.g., earlier, deleted, or modified backup drafts) can be especially problematic because early drafts may contain information that does not reflect the company's "true or final position, but are discoverable nonetheless."

2.2.2.1.4 Latent Data

Latent data (also called residual and ambient) is data that has been *deleted* but actually remains in the system. Deleting a file does not remove that data from the computer's disk. Instead the file is removed from its current location and moved into a *trash bin*. The location of a file is an association between the data and the system of files that is external to the data and is recorded only by the system of files management data. The file stays in the trash bin until it is finally removed altogether, can be done from the trash bin, or restored to its original location. Physically, deleting a file simply changes the location of the file and makes it inaccessible except for finding it in the trash bin. Thus, the physical recording of the data on the disk is not involved in any of the stages discussed here. In other words, even files removed altogether may still have their data on disk in full or in part. "Old data" is recycled. That is, space on the disk can be overwritten if the disk space is needed. Unallocated recycled data is accessible through special tools, i.e. tools read the physical data, and therefore discoverable.

The degree of recoverable old data is arbitrary. The number of variable involved in the life span of deleted data is large. Amount disk traffic, disk capacity, disk utilization, and size of the deleted files are only some of these variables. Discovery success cannot, therefore, be predicted. Furthermore, tool to erase all latent data are a dime a dozen and it seems likely that sooner or later companies and individuals will start to use them on mass. Since these tools leave no marks, their use will increase.

---

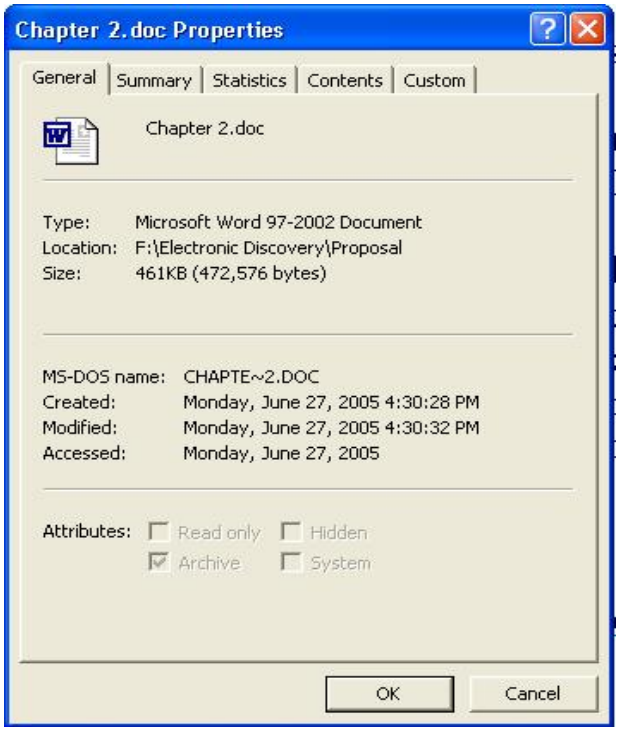
[24] nextPage® 1.5 product overview, http://www.nextpage.com/products/index.htm
[25] CS-RCS Basic 4.0.270, http://www.softpedia.com/get/System/File-Management/CS-RCS-Basic.shtml
[26] Configuring Transaction Logging, Chapter 9, Cisco Application and Content Networking Software Caching Configuration Guide, 2002.

2.2.2.1.5 Embedded Data

2.2.2.1.5.1 Simple Metadata

"Embedded data" is data that is automatically created and stored by computer programs that is not part of the content of the document. Embedded data is metadata, which is "definitional data that provides information about or documentation of other data managed within an application or environment[27]." For example, word processing programs store information about when data files are created and when, and who accesses them. The figure below is an example of some metadata of the document this text appears in. In shows the following items that are not part of the file content:

- Document name
- File location
- Size
- Creation, modification and last access times
- Attributes of different kinds



Microsoft Office documents, which play a central role in today's basic infrastructural documents, automatically store information, i.e. metadata, about the individual working

[27] Dictionary.com, http://dictionary.reference.com/search?q=metadata

on the documents, her firm, her network and each time your create or access a document file.

The following are most of metadata information found in a typical Word or Excel document:

- Author's name and initials
- Author's organization name
- Server name on which the document is stored
- File properties/summary information
- Non-visible portions of embedded OLE objects
- Previous author's names and initials
- Document revisions
- Document versions
- Template details
- Hidden text
- Hidden Cells
- Comments
- Smart Tags
- Network and World Wide Web links

Word's Object Linking and Embedding (OLE) capabilities allows placing a presentation slide or cells from an Excel worksheet into a document, and Word associates that object with the program used to create it[28].

Microsoft's Smart Tags[29] allow items in Word docs, spreadsheet cells and other Office applications to have properties attached to them. For example, a person's name in the document could have knowledge of its entry in a address book, or as the author of a book, or as a family member in some file. Tags may have multiple associations of this sort.

Working with an item with Smart Tags attached to it, the individual is presented with a number of options for actions to be taken in association with it. Conceivably, the actions could be automatically carried out. The action can apply to multiple targets the number of which is unlimited. Smart tags, therefore, have great potential. These tags are a great mechanism for automating connections between different files based on different applications. As metadata and for discovery such complex and rich source may be a God's send.

---

[28] Gregory Harris, Cut your Word docs down to size with these OLE techniques, TechRepublic, Networking and Communication, 8/26/2002.

[29] Darshan Singh, **Integrating Office XP Smart Tags with the .NET XML Web Services,** e-doc (Adobe Reader), Apress; ISBN: B0007MHF54; (May 20, 2002)

Smart Tags are written in XML[30] which is a widely used mark up language used widely by industry for which there are unlimited number of support tools as well as usage.

MS Word, like many other applications, makes use of embedded commands of different kinds. To demonstrate this type of metadata, we start with *Field Codes* in Word. An example is:

> The following field displays a Microsoft Graph object embedded (embed: To insert information created in one program, such as a chart or an equation, into another program. After the object is embedded, the information becomes part of the document. Any changes you make to the object are reflected in the document.) in a document:
>
> **{ EMBED** MSGraph.Chart.8 * MERGEFORMAT **}**[31]

Embedded commands are small programs that direct the application, in this case the application is Word, to execute a sequence of step to form a fragment of the document. Spreadsheet application may have mathematical function and even small programs embedded in them.

A paper version of the document does not even hint at the existence of such data, but even an electronic version does not show commands unless explicitly asked for.

2.2.2.1.5.2 Variety of Simple Metadata

Metadata is not restricted to documents. Cookies are small text files that are stored on a user's computer by a Web server explicitly permitted to do so by the user's browser software. A cookie itself is not typically read by the user. Rather, it is an identifier used by the Web site that originally placed it on your hard drive. Cookies can contain any arbitrary information the server chooses and are used to introduce state into otherwise stateless transactions.

Due to the centrality of Microsoft Office products in the office setting most discussion of metadata centers around MS Office metadata. Clearly, other products have their own metadata instances and typical use. Adobe documents, PDF, and WordPerfect also have metadata[32].

Typically cookies are used to authenticate or identify a registered user of a web site as part of their first login process or initial site registration without requiring them to sign in again every time they access that site. Other uses are maintaining a "shopping basket" of goods selected for purchase during a session at a site, site personalization (presenting different pages to different users), and tracking a particular user's access to a site.

Metadata types and variety is vast, growing, changing and requires a detailed understanding and the particular domain to which the metadata applies. Fortunately,

---

[30] Elliotte Rusty Harold and W. Scott Means, XML in a Nutshell, Third Edition, O'Reilly; (September, 2004)
[31] MicroSoft Word Help file.
[32] Donna Payne, Metadata – Are You Protected? Payne Consulting Group, 12/7/2004.

discovery will typically cover limited domains and a single domain expert may suffice. This book *Metadata in Practice*[33] provides a background on the world of metadata. It details a wide range of different metadata projects that involve an education digital library, statewide collaboration efforts, museums, university libraries, an image database, geographic data, aggregation, and sharing. It discusses the future of metadata development and practice by exploring its standards, harvesting, reuse, repurposing, and interoperability, among other topics.

2.2.2.1.5.3 Complex Metadata

Most publications dealing with documents focus mainly on textual documents such as: memos, articles, letters, notes, draft documents, manuals, emails, etc. All these formats are the creation of word processors. The latter may be MS Word, WordPerfect or a simple computer notepad. We have already mentioned metadata associated with these document formats, but there are potentially other metadata that are less common these days but are being used increasingly with time.

A new generation of documents based on markup language, data description and style formats is already used in many applications. Web pages are mostly written in HTML, which stands for Hyper Text Markup Language. A simple example clarifies both HTML and the basic notion of markup. To display **EXAMPLE** in HTML one writes <B>EXAMPLE</B>. Where < and > indicate a markup, i.e. formatting instruction, B indicates Bold. With enough markup variety one can use a markup language to do everything a typical word processor does.

When viewing Web pages one does not see the markup unless it is asked for explicitly. A markup is an executable form of metadata. This metadata has its own metadata. HTML uses Cascading Style Sheets (CSS) that are style sheet language used to describe the presentation of a document written in a markup language. CSS is used to define colors, fonts, layout, and other aspects of document presentation. CSS is the metadata of HTML while HTML is the metadata of the Web page document. Different style sheet will render different colors, fonts and basically change the visual depiction; the content is not affected.

Why would the requestor be interested in HTML markups? In other words, although technically HTML markup is metadata, it is seemingly of no interest for legal purposes. That is a wrong assumption. At least two markups may be sources of significant information. HTML has a comment markup, e.g. <!-- Hello -->. However, instead of Hello is may say: <!—copyright by ACD Inc. --> and ACD happens the requestors' client. In other words, this metadata is the smoking gun!

Digital Libraries

2.2.2.1.5.4 Application-related Metadata

2.2.2.1.5.4.1 Revision Control

[33] Diane I. Hillmann and Elaine L. Westbrooks (Editors), Metadata in Practice, American Library Association (June 1, 2004)

Many documents are singletons. Such documents may be related through topics, date, version or source, but each is saved separately and independently of the others. For instance, when counsel takes depositions of several witnesses, each deposition record is totally independent of the other records. The depositions are related by the case under hand, some witness may cover overlapping areas, etc.

Above we mentioned NextPage as a revision control application. Here is what Wikipedia says about such an application:

> Revision control is an aspect of documentation control wherein changes to documents are identified by incrementing an associated number or letter code, termed the "revision level", or simply "revision". It has been a standard practice in the maintenance of engineering drawings for as long as the generation of such drawings has been formalized. A simple form of revision control, for example, has the initial issue of a drawing assigned the revision level "A". When the first change is made, the revision level is changed to "B" and so on.

Revision control (RC) maintains detailed information about the documents it contains and on the revisions made to them. This is metadata. Since this metadata pertains to document that already have simple metadata, this metadata augments that original one. We list some common metadata records maintained by a revision control application:

- Time and date document entered RC first time, which is later than the document creation time
- Time and date when was document changed
- Commentary for change (e.g. reason, nature of change, supervisor of change)
- Change itself
- A complete history of changes to document
- Documents may be grouped together into “projects,” which have their own metadata

Some example of project metadata that is (or could be) persisted in the project directory are:

- Project nature and users
- Launch configurations – how to use the documents
- Document manipulating application (e.g. MS Word)
- Tasks/Bookmarks
- File encodings
- Dead documents

The metadata maintained by RC contains way more information that the individual document contains. Most striking is the fact that RC may provide not only creation, modification and last access dates for a document, but also any access to the data, all the changes and who made them. That is a substantial history of a document.

2.2.2.1.5.4.2 CMS – Content Management System

Wikipedia says:

> A content management system (CMS) is a system used to organize and facilitate collaborative creation of documents and other content. A CMS is frequently a web application used for managing websites and web content.
>
> CMS allow authors to provide new content in the form of articles. The articles are typically entered as plain text, perhaps with *markup* to indicate where other resources (such as pictures) should be placed. The system then uses *rules* to style the article, separating the display from the content, which has a number of advantages when trying to get many articles to conform to a consistent "look and feel". The system then adds the articles to a larger collection for publishing.

The nouns italicized are metadata. CMS also has workflow, approval of changes record, session logs, templates and a lot of the RC metadata.

Again, CMS augments each document's metadata substantially.

2.2.2.1.5.4.3 Digital Library

Digital libraries are heavily used by universities, states and the Library of Congress. A digital library is a collection of documents. The table of core metadata elements for library of congress digital repository has more than 50 elements. Some digital library tool are already available and sooner or later the commercial world will start using it.

2.2.2.1.5.4.4 Other Applications

RC and CMS are commonplace. The detailed mention of them is useful because one will encounter them frequently doing ED. But there are many other many other applications in use with the functionality to create "projects" – families of documents – and enhance the amount of metadata available.

Some generic applications that come to mind are:

- Integrated (software) development Environments – for software development
- (Software) Testing tool
- Simulation application
- Weblog tool – for running weblogs
- Photo album software

We have used the generic name of the applications. In reality one uses a product, commercial or public domain that deviates from the generic application. Deviations are typically minor additions, omissions and changes.

2.2.2.1.6 Associated Data

Association data is data associated with the document that may, generally speaking, be defined as metadata data but is *not* part of the document electronically or otherwise. Most ED-related publications use metadata to mean data physically associated with a

document. For that reason, we have chosen the term “implicit data.” We make it clear that the electronic copy of the files does not necessarily contain implicit data.

We describe the following types of association data:

2.2.2.1.6.1 Passwords

In today’s document world there is a non-negligible use of passwords to protect reading (and modifying or controlling) of documents. If the document is discoverable, then access to the electronic copy of the file may require the password, which therefore is also target for discovery. In the extreme case, the password may not be recorded anywhere! Someone remembers it. There might even be genuine cases of a discoverable document whose password has been lost. Such files are far from being a lost cause, there password breakers that will probably succeed in making the file readable. Counsel will have to get the opponent’s or the judge’s permission to that first.

2.2.2.1.6.2 Encryption

This is a case with some similarity to the password case. One needs a key to decrypt the file in order to read the document. The key is not part of the physical document, was likely, and is either stored somewhere else, remembered or even lost. So far password and encryption files seem identical relatively to discovery. Here is where the similarity breaks; there are no tools that will decrypt files without the key.

2.2.2.1.6.3 Split Documents

Most document come as an atomic unit (i.e. one does not see the parts that make the document up). There are, however, deviations from that nice and clear existence. A large document may be formed by have a master document broken into smaller components. Using a master document is a method of organizing a large document by breaking it down into several smaller, manageable files[34]. These smaller files are referred to as ***subdocuments*** and you can insert these subdocuments into the master file. Imagine having the subdocuments but not the master. How can you tell that the subdocuments have any relationship between them? Proper discovery must make sure that the master and the subs are provided together or a single document consisting of all subs is produced. Note that the latter product has removed from production the fact that there are subdocuments. To see the value of the subdocuments let us review an example. Master file M consists of subs A, B and C. A was written by Jill, B by Joan and C by Jillian. Your counsel was not aware that Jillian works for the “other guys.” After all, she used to work for “us.” In other words, the smoking gun is called Jillian.

Files may be split for other reasons such as size or organization. Many email systems have size limit on emails. There no general guidelines on that but 2, 5 and 10 megabytes are sometimes the size limitations. Given a file of size 50 megabytes, it may be justified to work on it as ten 5 megabytes files. Most publications on ED seem to deal with text document, where the content of the file may indicate its companion files. Many cases, however, deal with non-textual data. It may images, computer applications, numerical

[34] Nancy Fitzpatrick, Microsoft Word Large Documents, Integrated Technology Services, January 11, 2005.

files, etc. In such cases, one has to be given a list of the companion files in order to put this humpty dumpy together again.

2.2.2.2 Linking

MS Windows supports the following operation. Copying information from document A and pasting it into document B, the information that you paste becomes part of the document B. If you change the copied information in document A, those changes will not be reflected in document B.

For Windows-based programs that support drag-and-drop functions, you can use linking or embedding to transfer information from one document to another document. Linking the pasted information retains the connection to the information stored in the document A. Linking causes changes in A to be automatically reflected in B.

2.2.2.3 Linked Files

Before we present the next distinct ED data item, we look at a short description of the term “directory” that is quite well known and understood. (http://wikipedia.org/)

> A directory, catalog, or folder, is an entity in a file system which contains a group of files and other directories. A typical file system contains thousands of files, and directories help organize them by keeping related files together. A directory contained inside another directory is called a subdirectory of that directory. Together, the directories form a hierarchy, or tree structure.
>
> If you imagine the computer's file system as a file cabinet, high–level directories may be represented by the drawers, while lower–level subdirectories may be represented as file folders within the drawers.

We can now present the concept of hard link.

> A hard link is a reference, or pointer, to the physical data on a volume. On most file systems, all named files are hard links. The name associated with the file is simply a label that refers the operating system to the actual data. As such, more than one name can be associated with the same data. Though called by different names, any changes made will affect the actual data, regardless of how the file is called at a later time. Hard links can only refer to data that exists on the same file system.
>
> The process of unlinking disassociates a name from the data on the volume. The data is still accessible as long as at least one link that points to it still exists. When the last link is removed, the space is considered

Links, missing links and dead links may all complicate discovery. Multiplicity of files can cause confusion. Take two files A and B, both linked to the same physical data. Analyzing the produced files may lead to search for reasons for having two identical files, why two files in different directories are identical and what is the real motivation behind all that. In reality it may simply be a convenience ploy. No computer system is perfect; links go missing and the result is another complication.

A symbolic link is a special type of directory entry in modern Unix (or Unix-like) file systems that allows to almost transparently refer to another directory entry, typically a file or a directory.

In contrast with hard links, there are no restrictions on where a symbolic link can point, it can refer to a file on another file system, to itself or to a file which does not even exist (e.g. when the target of the symbolic link is removed). Such problems will only be detected when the link is accessed.

Symbolic links can be dangerous. Although at first glance links to files create the illusion of the file being present in several locations at once, just as hard links, these locations are not equivalent and deleting the wrong pathname destroys the file completely.

2.2.2.4 Context Data

Suppose the request is for 10 insurance policies. The target of the request has 7 policies in a folder called regular and 3 policies a folder named *Risky*. Upon request all 10 policies are produced. The folders themselves *Regular* and *Risky*, however, were not requested and are not produced. Clearly, the receiver of the files has lost potential insight into the fact that the opponent has deemed 3 of the policies as risky. Context data was lost.

There are other types of context information.

2.2.2.5 Structure Data

The very nature of a file system that holds documents is hierarchical.

2.2.2.6 New Hardware Data

Probably the most substantial revolution caused by ED is the tremendous increase in the number and variety of discovery sources. While in the “pure” paper days the sources were the different paper documents used by the discovery target, ED came about at an age where computers – themselves a huge expansion of targets – go hand in hand with cell phone, answering machines, ISPs (Internet Service Providers), PDAs (Personal digital assistants), pocket PCs, pen registers, USB flash memory, iPods (portable digital audio player), video cameras, digital cameras, programmable phones, tablets, car navigation systems, etc.

The list will change, items will be added and others will disappear. Most of the devices have components or similarities to the general discussion so far.

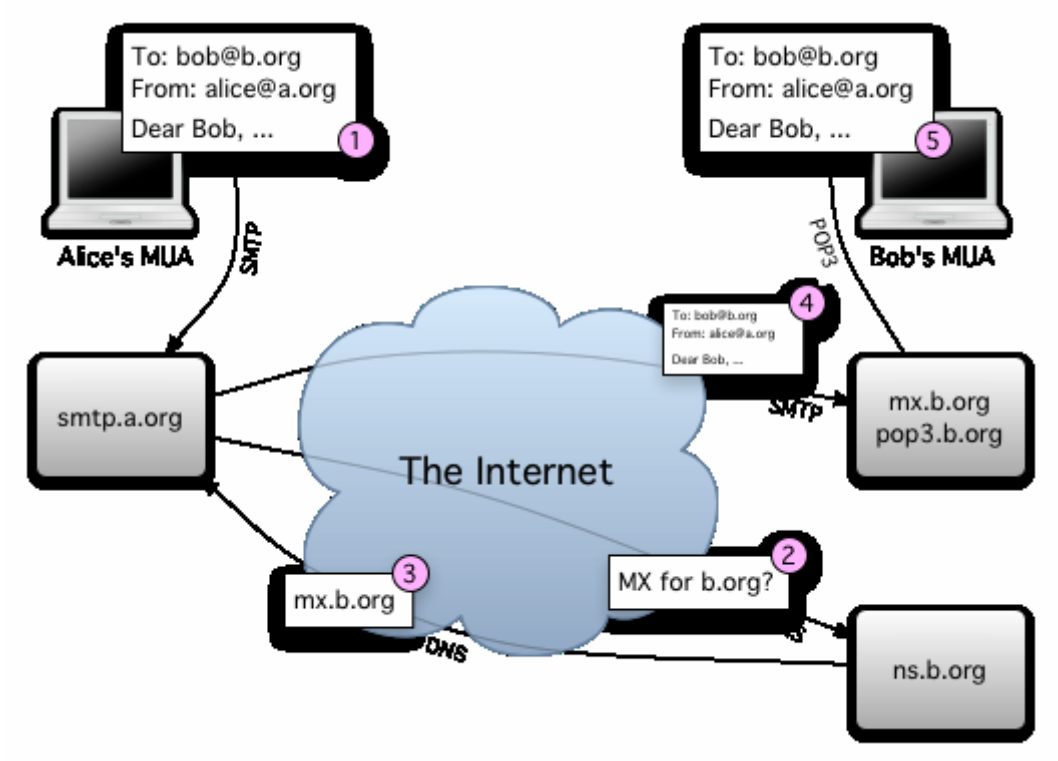
- PDAs, pocket PCs and tablets are just personal computers and these are dealt with anyway.
- USB flash memory and iPods are instance of file systems discussed already.
- Answering machines and car navigation system have a small number records, in different formats, on them. Retrieving this information is rather simple. Answering machine may have the same behavior as deleted files. The message may have been deleted but is still physically available and, therefore, accessible.

- Programmable phone and cell phones have three sorts of data:
    - Address book type data consisting of records with a name, its associated phone number and a set of attributes – type, ring, etc.
    - Small log of recent calls
    - Miscellaneous data – notes, alarm times, filtering information, etc.
- Video cameras and digital cameras have images and have programmable data similar to phones

The list of digital devices above is far from complete. Air conditioning systems may be programmable as others may be: security systems, home and car locks, alarm clocks, etc. It is important not to drop any item that has a discoverable data. Yet, discovery on such devices is not automatic.

2.2.2.7 Piled Data

Piled data is data that has no discernable data. It still has it active, archival and replicant counter parts. As for the rest of the attributes, they are are more dependant on the actual data.



2.2.2.7.1 E-mail

Electronic mail, abbreviated e-mail or email, is a method of composing, sending, and receiving messages over electronic communication systems. Most e-mail systems today use the Internet.

E-mail messages have a fixed format and consist of a relatively small number of components:

- Headers - Message summary, sender, receiver, and other information about the e-mail
- Body - The email text message itself, usually containing a signature block at the end.
- Attachments – Files that accompany the text and offer additional information, pictures, figures and documents.

Email headers typically have at least four fields:

- *From*: The e-mail address of the email message sender
- *To*: The e-mail address of the destination of the email message
- *Subject*: A short phrase or sentence that serves as a headline; typically a summary of the contents of the message
- *Date*: The local time and date when the message was sent

Other header fields include:

- *Cc*: Carbon copy or e-mails of additional destinations of the message
- *Bcc*: Blind carbon copy - the recipient of this copy will know who was in the To: field, but the recipients cannot see who is on the Bcc: list
- *Received*: List of mail servers that have handled the message

```
Received: from carbon.sag.gwu.edu (carbon
[192.168.61.231])
 by mail.sag.gwu.edu (iPlanet Messaging Server 5.2
HotFix 2.04 (built Feb  8
 2005)) with ESMTP id
<0IJ000GXCCYP8C@mail.sag.gwu.edu> for x@gwu.edu;
 Sat, 02 Jul 2005 12:11:19 -0400 (EDT)
```

- *Content-Type*: Information about how the message has to be displayed, usually a MIME type

The header has also automatically created fields and partial trace of the path the email has taken.

```
Received: from mail.gwu.edu (mail.gwu.edu [123.456.7.89])
by mailhost.some-isp.com (7.7.5/7.7.7) with ESMTP id
QBB43210 for <joe@some-isp.com>; Wed, 28 Mar 2007 14:39:24
-0800 (CST)
Received: from beta.gwu.edu (beta.gwu.edu [123.456.7.87])
by mail.gwu.edu (8.8.5) id 004Q43; Tue, Wed, 28 Mar 2007
14:36:17 -0800 (CST)
From: rth@bieberdorf.edu (R.T. Hood)
To: joe@some-isp.com
Date: Wed, 28 Mar 2007 14:36:14 PST
Message-Id: <tts031898234123-00000298@mail.gwu.edu>
X-Mailer: Morris v7.1
```

Subject: Dinner?

A mail server receives email from email clients or other mail servers. Such a server typically consists of storage, a set of rules and a database of user accounts. The storage stores mail for local users and messages in transit to another destination temporarily. The rules determine how the mail server determines the destination of email messages or reacts to the sender of the message.

An e-mail message passes through several e-mail servers each of which will store it temporarily, until the message reaches its destination. After the message is received at its destination, it is stored on personal computer or on the local e-mail server. Technically, once the user files or deletes the e-mail message, the local or remote servers can delete the message. Legal obligation mandates that the server's owners keep an archive of email messages sent and received by the server.

E-mail volumes overall stands in the trillions annually and the volume continues to grow. Growth in e-mail traffic will cause, in most likelihood, a commensurate increase in discovery requests seeking access to e-mails. Like other electronic data, e-mail is not easily deleted. Even if the sender and recipient delete the file, the e-mail message, if not the e-mail itself, may still exist.

Discovery of e-mail has two different facets. An e-mail can be viewed as a document, in which case it may be considered active, archived, replicant, latent data and embedded data.

When presented with a discovery request for a selected set of e-mail, companies failing to institute such policies are often unable to comply, claiming that to locate and provide the selected set would be too costly, difficult, or time-consuming. Courts can and have ordered such companies to then turn over all of their electronic mail. Once all the e-mail is turned over, the receiving party's search has the potential to uncover much more information than would have been possible if only the requested set of e-mail had been provided.

The major difficulty with e-mail discovery is the unstructured nature of this collection of documents. The magnitude of e-mails make discovery even more complicated. Email discovery therefore overwhelms the practitioners due to the sheer numbers and make the task an enormous hurdle to overcome to achieve quality discovery. According to IDC, in 2004, thirty billion messages crossed the Internet each day[35]. Assuming traffic between 20 and 50 e-mails per day and several thousand e-mail accounts per company, an e-mail system can easily consume two to three terabytes of storage capacity annually. Even small firms are looking at several hundred gigabytes of e-mail data per year[36].

[35] Dark Traffic Email Report, Q1, 2005, Tumbleweed.
[36] Nigel Williams, E-mail shots cost administrators dearly, ecominfo.net 2005.

Size by itself is not everything. Equifax, the giant consumer credit bureau "contains files on 210 million adult consumers[37]." Each record can be very large and covers 6 years and 9 months of personal financial records. Although concrete size figures are proprietary, an guesstimate of up to 100 terabytes is acceptable. The consumer credit report database is not impossible to navigate. Although most consumers appear several times, I doubt the 210 million consumers – at times Equifax, Experian and Trans Union had up to 700 million consumers in the database, production of the individual credit report is lightening fast.

What sets large email archives apart from huge databases is structure and rules. A database is structured, one searches for keys in combination with constraints, multiple hits are simply combined. In case of ambiguity, e.g. are John Doe and Johnn Doe the same person? Semantic rules aid in determining the right answer. Email archive lack the structure and the semantic rules to allow easy resolution of ambiguities.

We briefly outline email's special traits:

- Lack of structure – as stated before, emails are a pile. One can sort emails by destination, origin, time and date, but none of these attributes has anything to do with email content.
- The *subject* field is highly informal and of limited use as a discriminator. Peolple frequently drag the same subject for a long exchange of emails. Thus, a useful subject is more a result of good luck than smart analysis.
- Limited metadata –email metadata contains[38] is mainly:
    - Sender, Recipients
    - Subject and Body
    - Date when the email was sent
    - Attachments

  This set of attributes is of limited use.
- The associated data starts from the address book and continues with the same associated data other structured documents have.

A tool[39] can use email metadata by storing it in a searchable database (a simple enough tool that can be written by a good student in two weeks). This data can be searched as fielded information in a document management system. The messages can then be searched in more specific ways. When reviewing thousands, or millions, of email messages, attorneys have to be able to manipulate, manage, and search in any number of ways, and the electronic capture of metadata makes that easier. The problem, though, is that on huge amount of email, the searching itself is almost Google-like, i.e. the success rate is low. Surveys[40] have shown that: the success rate of these searches is less than 50

---

[37] John A. Ford, Chief Privacy Officer, Equifax Inc., Atlanta, Georgia, Before The House Committee On Financial Services/Subcommittee on Financial Institutions and Consumer Credit, Spencer Bachus, Chairman, June 4, 2003.

[38] Jason R. Baron, E-mail Metadata In A Post-Armstrong World, 1999, http://www.computer.org/proceedings/meta/1999/papers/83/jbaron.html

[39] Patricia Vinci, Electronic Discovery: How to Conquer the Basics, May 22,2002, IBIS Consulting.

[40] Laura Gordon-Murnane, The Invisible Web: What Search Engines Can't Find and Why, University of Maryland Libraries Digital Dateline Series, November 5, 2003.

percent and 13.3 percent of those in the survey found what they were looking for half of the time.

2.2.2.8 Access control lists

Most networks employ access control lists to "limit users' right to access, view, and edit various files otherwise available on a network." Access rights often vary depending on employee job titles or positions in the company, with some employees allowed read-and-write access and others read-only access. If the issue for which discovery of access lists is sought "centers on a particular file or group of files, identifying who had access rights to the files and the type of access each person was allowed can establish data ownership/authenticity of files."