OUTLINE OF ISSUES RELATING TO ELECTRONIC RECORDS

Brian Nelson Burford
NH State Archives

General Issues	*Sub-Issue*	*Specifics or Examples*
Definitions		
	What is a "record"?	Durable information showing that an event happened, or a decision was made, and created at or soon after the event or decision by a person who witnesses the event or decision.
	What is an electronic file?	Information stored as a combination of machines that manipulate physical characteristics of media in such a way as to store meaning for later use. Electronic Records require a combination of hardware and software to interpret this information. The media is organized in such a way as to inform the machine if the "switch" is on or off (also expressed as 0s and 1s, or binary data). Each "switch" is called a bit, and they are organized in groups of 8 called a byte. A file may contain instructions (platform or application) or data to be manipulated according to the application instructions.
	What is Digital Preservation?	The ability to maintain data (zeros and ones) in a structure over time in such a way to allow one to reinterpret those zeros and ones with the same meaning as when the data was created. The purpose is the preservation of the information preserved in the "zeros and ones" for later use by humans.
Sources of Data		
	Capture born digital from sources inside organization (government or corporate department, division, or other agency)	files created by users on corporate network
	Capture web-sites	documenting information available for the public to utilize over time
	On-line forms	information supplied by users of a form

		located on a web-site
	Import from outside sources	files created by users outside firewall, and imported
	conversion of magnetic to digital	e.g., VHS to digital
	Scan existing fixed documents (e.g., paper, microfilm, sound recordings, motion picture film) to raster image, sound or video files	

Hardware		
	CPU speeds	
	internal or external storage	
	devises	floppy drives, CD readers and writers; DVD readers and writers
	documentation	
Software		
	Programming Languages	Java C++ Perl Python (an *interpreted*, *interactive*, *object-oriented* programming language) HTML XML (universal language) [many, many others]
	Platform	Microsoft (MS) DOS[2] Unix[2] Linux[1] FreeBSD[1] NetBSD[1] Debian GNU/Linux[1]
	Desktop	Windows XP, NT, 2000[2] Gnome[1] KDE[1]
	Server	MS Server[2] Apache[1]
	Applications (examples)	MS Word (text)[2] MS Excel (spreadsheet)[2] MS Outlook (email, a form of text)[2] Mozilla (email program)[1] MS Access (database)[2]

		MSSQL (database)^{2} Oracle (database)^{2} MySQL (database)^{1} PostgreSQL (database)^{1} DSpace (database)^{1} Greenstone (digital library) ^{1} JHove (metadata extractor)^{1} Manifold (GIS)^{1} AutoCAD (CAD)^{2} Ascent Capture (Electronic Record Management [ERM])^{2} Laserfiche (ERM)^{2} Alchemy (Object database)^{2} Veritas Electronic Vault^{2} (electronic record archiving software) iLumin (email archiving software)^{2} Zantaz EAS^{2} (email archiving software) University of Michigan Digital Library eXtension Service (DLXS) (digital library)^{1} ePrints (Institutional Repository software)^{1} *In the list of applications above, the trailing superscript number indicates Open^{1} or Proprietary^{2}*
	Open^{1} vs Proprietary^{2}	Open Source: the source code is made publicly available and a general license to use software is granted for future use. Proprietary: the computer code is owned by a private entity, and future rights of use and changes are controlled by that private entity.
	Versions & backwards compatibility	
		new features which interpret data differently
		could effect “presentation” or appearance of document – and hence our interpretation of the document (e.g., formatting: columns, paragraph indentations, some characters, etc)

		documents containing inserted objects (photos, graphs)
	documentation	instructions on how software works, required hardware environment, and how it interprets data
Data		
	File Formats	*Notice that this list is very similar to the "application" list above, because specific applications usually create a data file in a specific open or proprietary structure or format. The preservation issue is, in fact, whether data files created by one application can be properly interpreted by another application program. A FORMAT is data arranged in a specific way following systematic rules, so the data interpreter will understand the rules and know how to retrieve the knowledge contained therein.* ASCII (text) txt (text) RTF (rich text format) DOC (MS Word text) PDF/A (image) PDF version 1.4 [ISO 19005-1.] security hole? PDF files after 1.4? TIFF (raster image) JPEG2000 (raster image) GIF (raster image) BMP (bitmap – raster image) PNG (personal network graphics – raster?) AI (Adobe Illustrator) (vector image) CDR (CorelDRAW) (vector image) CMX (Corel Exchange) (vector image) CGM Computer Graphics Metafile (vector) DXF (AutoCAD) (vector) WMF Windows Metafile (vector) MPEG (motion image) RealAudio (sound) WAVE (sound) MP3 (sound) ABC Amber Converter (Conversion)

		Xena (XML intermediate) raster is “resolution dependent”– number of pixels. The more pixels, the clearer the image. Vector defines shapes and combinations of shapes mathematically.
	Multi-formatted file structures	For example, TIFF files may have IPTC or XMP metadata formatted within it, meaning one file, but it needs two or three different software “translators” to make sense of it. Also, GIS ESRI 3 files have three separate bit-streams
	Optical Character Recognition (OCR)	ABBYY Finereader Prime Recognition
	Compression	Loss-less such as Group 4 *(i.e. TIFF4)* Lousy such as the LZW rules Mixed
	Aggregation	Zip files
	Authenticity & Certification	malicious alteration; unintended alteration *(e.g., during migration);* data corruption checksums *(sum of zeros and ones in the original data)* hash Values *(value derived from the relative positions of the zeros and ones in the original data)*
	File Naming Conventions	to ease recovering info; complexity begins when multiple users are creating files
	Dynamic Records	indexing changes to file to show earlier versions of data
	Organization & Indexing	Object database Folder hierarchy ERM (electronic record management program)
	Corruption	natural decay of magnetic medium; chemical reactions in CDs or DVDs; physical damage to medium
	“bit-level” preservation	keeping the digital structure of the original electronic document

Metadata		
	Standards	Dublin Core; modified Dublin Core Australian Govt.

	Capture method	automated;
		user determined;
		mixed

	Method to tie metadata to data file	XML wrapper
	Thesaurus	used to assist indexing or searching on metadata Protégé[1] (Stanford) TheW32[2] (Tim Craven) – assists user to create a Thesaurus.
Security		
	Firewalls	control of access
	Encryption	control of access
	Read-Only	limitation of access

	Malware	Viruses;
		Worms;
		Trojans;
		Spyware;
		sometimes embedded or inserted in applications or data files

	File Sharing	
	Electronic Signatures	
Preservation Storage Media		
	Magnetic	hard drive(s) CDs (not all are magnetic) *there are even differences between magnetic (writable) CDs—gold or silver CDs have longer "life" expectancy* DVDs (not all are magnetic) floppy disks *(very short expectancy – 2 years from manufacture?)* tape storage sticks
	Optical Disks	(also called WORM – "Write Once Read Many" – but this is not the same as Worms listed under Security). Optical storage works on the technology of bounced light off uneven reflective surface – more durable than magnetic

	Fixed	Computer Output to Microfilm *(COM)* print to paper
	Mixed	Information Life-cycle Management *(ILM) – tiered according to value; uses paper, film, electronic according to need*
Preservation Strategy		
	Retention Schedules	NH Municipal Records Board RSA
	Migration	open the file in a newer version, allowing that program to change the data-structure to the new requirements
	Encapsulation	storing both the data and the application files together, perhaps allowing the future use of the application program to open and interpret the data
	Emulation	writing an application program that translates former application instructions into current instruction for data interpretation (similar to “drivers” which translate instructions for hardware devises)
	“Computer Museum”	maintaining working examples of all hardware and software in order to run the programs and data files into the future – requires being able to obtain “antiquated” parts which wear out. And requires knowledge of how to use program (so save the documentation). Requires storage space.
	Hybrid Technology	Computer Output on Microfilm (COM) or scanning film images [computer (text, raster, vector, spreadsheet) to microfilm to computer (raster)]
	LOCKS	“Lots of copies keeps it safe” – many copies distributed to various institutions improves chances of survival (“playing the odds”)
	Periodic “inspection” for data loss	quality control. Whether microfilm, magnetic, or optical.

	Periodic medium refresh	
	Checksums; error-correcting codes	audits of the bit-streams to ensure the data is transferred to new medium intact.
Access		
	Right-To-Know (a.k.a., Freedom of Information)	RSA 91A
	Privacy	This is for several reasons including identity theft or to prevent abuse of personal information (e.g., HIPPA, Patriot Act.)
	Intellectual Property Rights	ownership of the original medium; copyrights to information/interpretation
	Discovery	court-ordered during legal suits; requirement to preserve and share ALL information described in writ
	Forensics	methods used to recover data thought to be deleted or lost Restorer2000 WordFIX (Cimaware) AccessFIX (Cimaware) ExcelFIX (Cimaware) OfficeFIX (Cimaware) OfficeRecovery (http://www.officerecovery.com/) OnTrack
Transfer of Data		
	Compression	Lossless vs. lousy
	Security	encryption checksums & hash values
	Methods	CDs or DVDs File Transfer Protocol (FTP)
Disaster Recovery		
	back-ups	kept at a different site
Disposal		
	Delete key	erases pointer, but not data until overwritten
	Reformat disc	Formatting a magnetic medium

		prepares the medium for use by the operating system. The surface of the disk is checked for physical and magnetic defects, and then an address structure is added (e.g., FAT, NTFS, etc). The addresses may be made up of tracks, sectors, clusters, cylinders, etc. Included is a root directory, which is a list of addresses.
	Software scrub	overwrite with random data[3]– usually multiple overwrites. One theorist (Peter Gutman in Australia) recommends up to 35 overwrites. SDelete
	Degaussing	erase all magnetic info on entire disc with a powerful magnetic
	Destruction of medium	physically destroy the hard drive, floppy disc, CD or DVD – most secure

created June 8, 2005
revised July 12, 2005; July 13, 2005
minor addition Feb 25, 2006; March 14, 2006, March 20, 2006
Revisions April 4, 2006: in medium, file transfer, applications, XML

Brian Nelson Burford
NH State Archives

[3] "random" sequences are not technically possible in computers, and are really only QUASI-random. Hence, because patterns may arise, "randomization" is not an absolutely secure form of data destruction.