Permanent Digital Records and the PDF Format

Permanent Digital Records and the PDF Format
10:30 AM to 11:45 AM, Tuesday, October 24, 2000

Defining a Permanent TransFormat Records Management System, A Hierarchy of Record Storage Formats, Five PDF Formats, and Document Copying/Migration

Non-Technical Abstract

There are two parts to the problem of preserving electronic documents for long periods of time. The first is getting the images of the documents to display and print properly after we move only a few decades into the future and find that the versions of the software used to create the documents are no longer available. The second is preserving the bits, the ones and zeros, that make up the computer files that hold the records.

A digital copier or a facsimile machine creates what is usually considered an identical copy of a document. For this reason, the simple raster image format used by facsimile machines is a good choice for long term reservation. (For preservation, a resolution that is greater that the facsimile resolution is used to improve document quality.) More complex formats, such as word processor file formats, are subject to distortion by future versions of the word processor software or future versions of document viewers. Nonetheless, these more complex formats are an important component of a long term preservation strategy.

Standing between the differing levels of complexity of the facsimile file format and the formats of word processor files, are the Adobe.com PDF (Portable Document Format) files. PDF files include raster (fax-like) images, and an outline font format that falls between raster images and word processor files in both complexity and longevity.

Preserving the bits, the ones and zeros, requires high quality digital media. In addition, error correcting codes, such as the code used in RAID (Redundant Array of Inexpensive Disks, see bibliography) allows for the correction of the small number of errors that are inevitable on even the best of media.

Preserving the bits matters little if the chain-of-custody for the records is broken. The system used to maintain the chain of custody must be secure from tampering and secure from internal errors. It is necessary to know where backups are: for both backup restoration and for the deletion of records that have reached the end of their retention period. Ultimately, it is necessary to be certain that the right document have been preserved and to know how to locate them in the system. An example of a system to permanently preserve digital records, in the City of Los Angeles, Department of Public Works, is presented.

Technical Abstract

A TransFormat records management system (TF system) stores multiple formats of each record (document) to avoid the problems that are inherent in storing each format alone. A TF system also provides a means of bridging the gap between the document formats of today's document creation applications and the formats that will be used by those applications decades into the future.

The comprehensive records management requirements, that define a TF system, describe a system that is different than the many types of systems that have evolved previously to support one or more aspects of records management.

A TF system stores electronic records that include scanned images, computer generated documents such as word processor documents, and metadata about both the electronic records and paper and microform hard copy records, including their retention schedule. The records stored by a TF system, including the metadata records, are designed to last forever (a permanent retention period). While a given TF system may become obsolete over time, future TF systems can reconstruct the metadata from the stored TF system records, thus recovering the obsolete TF system.

TF records are stored in fixed length units called fascicles (which are modeled on the fixed length packets of the Internet and the fixed size digital imaging picture elements called pixels) to facilitate the electronic sealing of records and to simplify handling during backup and off-site storage.

A TF system includes the functions of a traditional records management system, but is different than a traditional records management system, which only stores metadata about the location of paper and microform records and the retention schedule for those records.

TransFormat (TF) records management is also different than document management, archives management, email and system directory management, knowledge management, and library management.

A TF system manages records stored in all known formats: paper, microform, and electronic. The TF system also stores the electronic records. The formats of electronic records can be arranged in a hierarchy of complexity, with raster images being the lowest (and most desirable) common denominator. Raster images can be used to validate the software viewers and emulators that reconstruct the more complex document formats, when the native applications, such as word processors, that created documents in those more complex formats, are no longer available.

The Adobe PDF (Portable Document Format) format, a leading contender for use in records management, is described as a single format, but in fact there are at least five PDF formats that must be managed separately. One of the PDF formats, the Normal format, is unacceptable for permanent records storage because it modifies the contents of stored electronic records.

While there is complexity at every turn in systems design, the only possible way to operate a successful TF system is to keep it simple. The definition of what is to be done must be carefully drawn. Conversion of record formats is a task that must be done by users, before records are submitted to records managers for storage. While being kept simple, the TF system must manage the permanent and long term (referred to collectively in this paper as permanent) storage of all paper, microform, and electronic records.

Creeping capabilities, such as the management of employee schedules and calendars, must be kept out of the TF system. Other, more complex systems, such as document management systems, that use the TF system to access electronic records in permanent storage, should be the recommended recipient of creeping capabilities.

An installation of a TF system in the City of Los Angeles, Department of Public Works, Bureau of Engineering, is described.

Non-Technical Introduction

Preserving document for long periods of time requires the ability to interpret the format in which the documents are stored. Preservation also requires the ability to recover the bits, the ones and zeros, that have been written on long lasting media. In addition, the problem of a continuing cycle of the obsolescence of successive records management software packages, each of which is followed by the need for document migration, must be solved with an easily understood certainty.

Formats

Documents come in several formats, each of which has advantages. All of these formats (when

available) should be preserved, for every document, to gain all of the benefits that are provided by the different formats.

The most common format for computer-generated documents is the native application format. An example of a native application format is a Microsoft Word document file. An increasingly common format is the raster scanned format, stored as a TIFF (Tagged Image Format File) or a PDF (Adobe Portable Document Format) file.

The native format has the advantage that all of the native application's features are available for updating a document. The raster format has the advantage that it is very simple and will therefore be interpretable for a very long time. The native format has the disadvantage that only the exact version of the application that created the file can produce the exact same image of the document when printed. Even worse, after only a few decades, documents stored in a native format may not be interpretable at all. The raster scanned format has the disadvantage that it carries very little of the formatting information that was created by the native application when the document's author originally laid out the document. In addition, when a raster format document is OCRed (Optical Character Recognition), the resulting OCRed text and formatting is ambiguous (wrong) and must be proofed. The proofing is the equivalent of re-authoring the document (or re-engineering an engineering drawing). However, unproofed (uncorrected), but hidden (in the PDF hidden text format), OCR text output is very useful for full text searching.

For many documents, the electronic copy of the native format is unavailable, often because of incompatibilities in the way the native format was stored. These documents must be scanned, and therefore, only the raster format will be available.

Raster images are created in laser printers when native format documents are printed, because laser printers can only print raster images. These same raster images can be created using Adobe Acrobat and other printing software. These computer created raster images have all the desirable properties of the simple scanned raster images. In particular, computer-generated raster images will be interpretable for a long period of time. If the computer-generated raster format of a document is saved, along with the native application format, the benefits of both formats will be available.

In addition, there are synergistic advantages of storing the multiple formats. For example, decades in the future, when the then current version of the native application, or third party document viewers, are used to render a new raster document image from the native format, the new and old (preserved) raster formats can be computer-compared. The computer comparison will highlight (for example in red) the pixels (picture elements) that have been changed from white to black or from black to white. The highlighting will show the differences between the old and the new versions of the document. The person doing the re-authoring or re-engineering of the document can then decide if the differences are material. If the differences are material, then the person can make changes to the native application format of the documents (by changing the documents using the new version of the native application) to correct the errors of interpretation. (The documents themselves must be changed to accommodate the new version of the application, because there is no way to change the new version of a native application.)

As a document moves from the native format to a computer-generated raster format for printing, there are other intermediate formats that may be created. These intermediate formats take the form of programs that, when executed in a computer, produce a page image. One intermediate format is a graphic markup language in which the structure of a document is defined separately from the textual content of the document. Examples are SGML (Structured Generalized Mark-up Language), HTML (Hypertext Mark-up Language), and XML (eXtensible Markup Language). There are also PDLs (Page Description Languages), examples of which are HP.com PCL (Printer Control Language) and Adobe.com PostScript. Adobe has enhanced the format of PostScript to adapt it to document preservation. The enhanced PostScript format is called PDF. (PDF includes at least five different formats, as described later in this paper.) These intermediate formats also provide benefits for document preservation both individually and synergistically, when combined with simultaneously preserved additional formats for a given document. Because these multiple document formats are key to providing actual access to preserved documents, these multiple formats, and their synergistic reinforcement, are important elements in the digital chain of custody.

The Bits

Preserving the bits is relatively easy. All that is necessary is to procure reliable media and record the bits using an ECC (Error Correcting Code). ECCs are built into all digital recording devices. Error correcting codes record additional information on the media so that errors, that occur when bits are recorded, can be corrected.

To avoid the problem of physical destruction of the media, multiple copies of the media are made and stored at multiple locations. Because ECCs are very important to successful preservation of documents, ECCs cannot remain hidden as a transparent feature of media writers. Records managers must maintain a knowledge of ECC operation and the specific ECC and density of errors (bad bits) on the physical media units used to preserve documents. ECCs, and the quality of physical media, are important elements in the digital chain of custody.

The System

Having addressed the document format problem, and the bit preservation problems, it is now necessary to address the problem of the system that inputs the documents, outputs the media, and then provides access to the documents on the media. Unfortunately, there is considerable evidence that all software disappears over time, for a myriad of reasons. To avoid this, it is best to plan for all documents, and document metadata, to be recorded on the media so that in the future, new software can be written to interpret the documents and metadata that was previously organized and stored.

Planning for a software conversion in the future is not a radical departure from the current state of affairs in the document management world. Many document management users have undergone one, and even more than one, migrations of their documents from one vendor's software to the software from another vendor, or even from one version to another version of the same vendor's software. In these migrations, the documents themselves, and their metadata, are separated from the first system, and inserted into the second system. What is suggested here is that this migration be preplanned, and that the documents and metadata be stored, at the outset, for the purpose of supporting future migrations. In this way, the first system that provides access to the documents is viewed as the first, of many, systems that the documents and metadata will be migrated to.

Thus, the solution, to the continuing cycle of the obsolescence of records management software, followed by document migration, is to plan for the cycle. The cycle appears to be inevitable, over long periods of time.

The static approach to document storage is most easily followed with documents that have a permanent or long term retention period. Dynamic documents, especially during the collaborative document creation period, do not fit well into this static model.

The static model makes backup much easier, because each document and its metadata do not have to be backed-up over and over again, the way documents must be backed-up when using the baseline and incremental backups built into document management systems designed for dynamic documents. This means that a records manager can identify the specific unit of media that a specific document is located on. This makes it possible to understand the backup process, and where documents are that must be destroyed or protected. Because the backups are static, it is very easy to organize multiple copies of an entire electronic records center at multiple locations, to protect against the physical destruction of media units.

Because the media that the documents may be written on may change over time, as media advances occur, the data is formatted and written to a virtual unit of media, called a fascicle in this paper. The fascicles are then written to a physical unit of media. This transporting of a virtual media unit is also useful when documents are migrated to a new online system. The fascicles are copied from media such as a DVD (Digital Versatile Disc) to a magnetic disk. The fascicles remain the same, but after migration the fascicles are available at online computer speeds. The metadata from the fascicles is then loaded into a relational database to provide document access.

In addition to providing portability across media types, fascicles can be digitally sealed using an

electronic signature. The electronic signature includes an encrypted checksum of the contents of the fascicle. This checksum is checked when the contents (documents) of the fascicle must be verified. If the fascicle has been surreptitiously or erroneously modified, the previously stored checksum and the checksum recently recomputed (for verification) do not match, and an unmodified fascicle can be retrieved from a second location, providing a verifiably true copy of the desired documents. Fascicles and their electronic seals are important elements in the digital chain of custody.

The system in this static document management model is called a TransFormat (TF) system in this paper. This is because the TF system is designed to work with multiple formats of each document, and because the TF system assists document managers in moving from old native formats to new native formats. Because the TF system is embodied in the static documents and metadata stored in the fascicles, the TF system also assists document managers in managing multiple TF system migrations: from on version (and / or vendor) of a TF system to another version (and / or vendor) of a TF system, over long periods of time. The TF system and the quality of fascicles it produces are important elements in the digital chain of custody.

Permanent TransFormat (TF) Records Management

Permanence is achieved through long-lived electronic records. This is similar to the time-proven reliability and permanence of file folders in a box. Records can often be recovered from boxes even after the organizations that arranged the records have disappeared. The format of TF electronic records must be kept simple, and their logical and physical arrangement must also be kept simple. Simplicity is required because, beyond a certain point, people and organizations give up and abandon systems (including TF systems), when resources are limited.

This certain point, this level of complexity, at which people and organizations give up, was described by George Miller in his paper, written in 1956, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". (See Bibliography for text of the paper). This paper states that a person can consider a maximum of seven different things at one time. If an eight item is added, failure occurs. By grouping items, more complexity can be handled. For example, by grouping 49 elements into 7 sets of 7 elements each, the larger problem can be made tractable. We can consider 7 groups at one time, but we cannot consider 49 elements, all at the same time. The problem is that there is a large stair step of effort required (to group elements) every time the number of elements, or groups of elements, exceeds seven. It is this stair step that is avoided by abandoning a TF system (or any system) before the systems complexity requires making the effort to ascend to the next level of grouping. This limit of seven is also the basis for span of control, in personnel management, and the fact that the height of an organization's pyramid is determined by the number of employees and the complexity of the work being done at each level.

TF system managers cannot expect their applications software to survive more than a decade. Beyond a decade, the simplicity of the records and the record's arrangement must provide the permanence. To avoid complexity, while providing complex services, TF systems serve (as a server system) the permanent and long term (referred to collectively in this paper as permanent) electronic records, that the TF systems contain, to external document management systems and other information systems. These external systems can have features that are very complex, without requiring that the complexity be transferred to the TF system.

Differences Between TF Systems and Other Systems

Traditional RIM Systems

TF systems store RIM (Records and Information Management) metadata, such as retention schedules and box (physical record) location, like traditional RIM systems. Unlike traditional RIM systems, TF systems also store electronic records. Like RIM systems, TF systems are designed to manage all of the permanent records of an organization and to include ephemeral (short and temporary records) in the TF system retention schedules. Unlike traditional RIM, TF system comprehensive records management extends beyond paper and microform records to permanent digital records. Image enabling of traditional RIM systems has come from the document management world, which deals with high value, active, current documents. The demands of the complex structures used in document management systems (that come with image enabling) place considerable strain on the budget and management time of RIM organizations.

When people save everything, they do not have time to think about what to save, or to think about how to arrange what they have saved. The result of saving everything is the same with electronic records as it is with paper records: a big mess. The administration of a TF system, like a traditional RIM system, imposes a requirement to sift and winnow the records and records series, so that when the events surrounding the creations of the records are not still fresh in the minds of TF system users, the users can still find the records they are seeking.

Document Management

TF systems store documents, as do document management systems. TF systems do not store the multiplicity of versions of documents created during the collaborative writing of the documents. TF systems do not manage the structure of subcomponents or objects, which may be included in documents by dynamic online reference via Internet or intranet hyperlinks and other techniques. TF systems manage all of the permanent documents in an organization, while document management systems manage only the documents stored in the document management systems.

Because documents stored in a TF system and their reference addresses (hyperlinks to URLs, Universal Resource Locators) do not change, documents that are managed as records in a TF system are ideal components of documents being constructed in document management systems. Stated another way, documents stored with unstable (non-ISO 9000 (See Bibliography) records management system based) hyperlinks should be flagged, and the unstable hyperlinks listed for each document.

TF systems, like traditional RIM systems, are expected to function normally after budget reductions. This is in contrast to the absolute requirement for continuous, and even augmented, funding for document management systems, which must be met or the document management systems often cease to function entirely.

Defining a Permanent TransFormat (TF) Records Management System

TransFormat Records Management systems are:

1. Designed to create electronic records that can last forever. (Almost all RIM systems manage some documents with a permanent retention period.)
2. Designed to manage paper, microform, and digital records (i.e. all records), optimized for permanent records, capable of storing long term records.
3. Designed to conform to ISO 9000 (See Bibliography) compatible quality audits.
4. Responsible for the proper handling of all of an organization's records.
5. Designed to survive budget cuts gracefully.
6. Designed to benefit from any increase in funding, no mater how small.
7. Designed to benefit from any money spent on an automation project, no matter when funding for the project is cut off. (This is unlike document management systems that often provide no value if the entire budget, and all cost overruns, are not fully funded.)
8. Designed to be used by (called by) elaborate, better funded document management systems, and other systems, that use the TF system to access permanently stored documents.
9. Designed with the expectation that there is a high probability that all document management systems that use the TF system will become unstable, flounder, and finally fail catastrophically over time; in ten, one hundred, or one thousand years.
10. Designed to not lose information on the decommissioning of document management systems that use the TF system.
11. Designed knowing that record viewing applications will change as vendors go out of business, that the operating system version that current applications run on will eventually be unavailable, that the hardware

that the operating systems can run on is only designed to last ten years, and that new hardware rarely supports old operating systems.

12. Provides a procedure for erasing permanent documents.

13. Designed to be managed by a RIM professional.

TF System Design Requirements

Simple Access

RIM cannot afford the emulators discussed by archivists for permanent document access. (See the papers in the Bibliography on "Ensuring the Longevity of Digital Documents" and "Preserving Information Forever".) RIM must depend on simple access to records without the operation of software. Boxed paper records provide this simple access. Punch card formatted data provide this simple access, with the addition of metadata. Metadata is data that explains, for example, which name is the first name and which name is the last name. On a punch card: "John Jay" can mean "Jay John" or "John Jay". Using XML (eXtensible Markup Language) (See Bibliography) metadata tags, "<name:first>John</name:first>" and "<name:last>Jay</name:last>", the data is unmistakably "John Jay", at any point in the future. The raster images produced by scanning are similarly easy to interpret as document images. For scanned documents, a document is made up of one or more scanned image files and the related metadata file.

To locate a file in a TF system, the metadata is searched and the corresponding document image file is retrieved and presented. The searching and presenting require a proprietary TF system software application. If the application is lost due to lack of funding, the stored documents and metadata can be recovered in their entirety, and entered into another proprietary TF system for searching. This recovery is exactly the same as the process currently used to migrate documents from one document management or RIM system to another. The only difference is that, in a TF system, every document is stored in a format that is specifically designed to facilitate the movement to another proprietary TF system. This requirement for future movement is the method used to guarantee that future migration of digitally stored documents will be successful.

The Archivists' Elegant, but Expensive, Emulator Approach to Future Rendering of Documents

Archivists are planning to use emulators to recover the document location and presentation functions of proprietary software. (See the papers in the Bibliography on "Ensuring the Longevity of Digital Documents" and "Preserving Information Forever".) Emulators are required because a given computer generated file, such as a Microsoft Word file, can only be rendered (presented), as intended by the file's author, by the exact version and build number of the software that the author used to write and proof the file (document). That version and build number of the software application will only operate exactly as the author used it on the operating system (such as Microsoft Windows) version that was used by the author. The operating system version must be run on the same version of the hardware (PC, Personal Computer) used by the author. As the software and hardware versions being used by future document users become less and less like the versions used by the authors, the rendering of the document becomes less and less like the original document, and ultimately, the software cannot be run and the document can not be used at all. Therefore, archivists plan to use software emulators to exactly recreate the author's original hardware and software environment in order to render the document perfectly.

The Muddle Through Approach (for Recovering Complex Document Formats)

The elegant emulators of the archival community will someday (in the fullness of time) allow the use of any software package from any computing era. RIM cannot afford to wait. The following is the TF system solution to storing computer generated documents that were not raster scanned. Building on the simplicity of raster document images, the application of human editorial effort preserves most of the usefulness of computer-generated files that are in formats that have become obsolete.

A Hierarchy of Formats and Format Complexity

Native application files, such as Microsoft Word files, are stored in their native format and can be displayed by the native application. These files can also be stored in a generic document format such as SGML (Structured Generalized Mark-up Language), which preserves both the text and layout information. The generic document format can be converted to a vector based outline font format such as Adobe PDF which describes all graphic elements as a line outline in a two dimensional Cartesian space (For a technical presentation, see Movement-Rotation-Scaling in PostScript and PDF in the Appendix.). The outline font format is converted to a raster (by a RIP, a Raster Image Processor, that is part of every raster printer) for printing by most types of printers, such as a laser or ink-jet printer. Each of these formats contains a progressively less rich format, containing progressively less of the original document's structure. Conversely, each of the less rich formats is progressively easier to preserve and present. By storing all four of these formats, the maximum benefits of each can be preserved. The fidelity of future interpretations of these formats can be judged as follows (following a muddling through process):

The simplest and easiest to preserve document format is the raster format. This raster format is the basis for validating future emulators, and the simpler software based viewers that are available even today for obsolete document formats. The raster images created by future emulators and viewers can be compared (visually and by computer) to the raster images saved at the time a document is stored in a TF system. The highlighted differences can be used to validate the quality of the emulator or viewer. The validation will always contain an element of the personal judgment of the person operating the system and judging the quality of the emulator or viewer. Similarly, applications carry the promise of being able to interpret files created by previous versions of the application. The raster images produced by these updated files can be compared to the raster images saved at the time the document was stored in a TF system to validate the quality of the application's update of the originally stored native file formats.

CAD, GIS, and Databases

CAD files, at the highest level of structure, include simulations based on three dimensional solid models of structures. Below this are the three dimensional solid models. The models can be reduced to three dimensional wire frame models. From these, two dimensional projections can be created and annotated. (These two dimensional projections are the familiar engineering drawings or blueprints.) These two dimensional models are similar to the word processor files described above. Databases (DB) and GIS systems follow a similar decomposition of structure. Database tables can be converted to a comma delimited flat file format and then stored using the same array of formats that word processor documents are stored in. Both CAD and databases can be preserved with extensions of the muddling through method described above. The native format office, DB, CAD, and GIS documents will always be needed because the native format is the best for future editing (modification).

Storing Multiple Formats of a Document to Improve Future Interpretation

By saving multiple formats of the same version of a document, with a method of assuring that all of the formats were created at the same time, and from the same version of the document, future reconstruction of the rendition of the various formats of the document can be supported.

For word processor documents that are retained permanently, the four formats in a multiformat rendition restoration container would be the original native format (such as Microsoft Word) (with an estimated lifetime of 3 to 9 months), a generalized text format such as SGML, a vector based outline font version such as Adobe PDF (Portable Document Format) (both with an estimated lifetime of 30 to 50 years) and a raster format (such as Adobe PDF Hidden Text format) (with an estimated lifetime of 5 hundred to 1 thousand years). Each of the formats has specific benefits, and the formats have additional benefits as a group. For example, the native format facilitates updates, while the raster format allows a quick review of errors created by migrating to new versions of the native format.

Logical Container for multiple formats

In Figure 1., 1 shows a logical container, which may be sealed by the electronic signature of the person placing the multiple formats of the same document in the container. The logical container links multiple formats of a single version of a document for storage as a single unit. 2 represents the protected interior space of the container. 3, 4, 5, and 6 represent multiple formats of the document (stored at the same time) for which future rendition restoration is contemplated or desired.

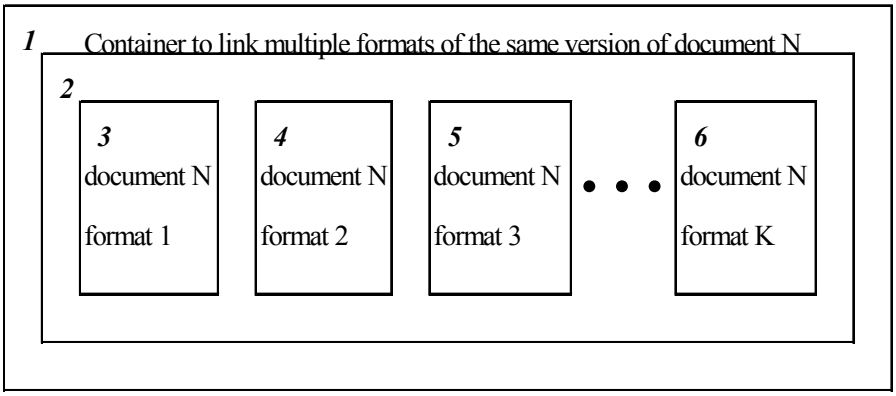


Figure 1. A logical container, linking multiple formats of a document

Example of Long Term Preservation of Records

We are getting used to being in the 21st century and to being in the 3rd millennium, but here in the Las Vegas area, (where this paper was presented to ARMA 2000) there are records managers that are preparing their records for survival into the 13th millennium. The Yucca Mountain Project Office is working on the Yucca Mountain Site Characterization Project [http://www.ymp.gov] of the US Department of Energy's Office of Civilian Radioactive Waste Management. The purpose of the Yucca Mountain Site Characterization Project is to determine if Yucca Mountain, Nevada, is a suitable site for a spent nuclear fuel and a high-level radioactive waste repository. These materials, including plutonium, are a result of nuclear power generation and national defense programs and will remain highly radioactive for thousands of years. The availability of records is planned for the first ten thousand years after closure of the repository.

Five PDF Formats

The following section illustrates five of the Adobe PDF (Portable Document Format) formats: 1.) The outline font technology found in the PostScript PDL (Page Description Language) from Adobe Systems. 2.) The raster scanned Original Image format without the text output of OCR (Optical Character Recognition). 3.) PDF Original Image with Hidden Text from (OCR) output, 4.) Normal Original Image, where the OCR output replaces scanned raster glyphs (character images) that were recognized by the OCR program (usually not acceptable for permanent digital images), and 5.) The raster image generated (from a PDL outline font page description), by a RIP (Raster Image Processor) in a laserprinter or ink jet printer. (See the PostScript Language Reference Manual in the Bibliography.)

Figure 2. shows a letter shape, 'L', produced by a list of coordinates that define the vectors that outline each letter in an outline font. Today, the shape of almost every graphic element of every printed image is produced using outline fonts.

The straight lines between the points listed above in Figure 2. are called vectors. Vectors can also be at an angle, as shown with the letter 'N' (below in Figure 5). Curved lines, such as the circle used for the letter 'O' are also a form of vector.

Figure 3. shows filling in the outline for the letter 'L' with large pixels. This makes the body of the letter black so that when the letter 'L' is printed, it can be seen.

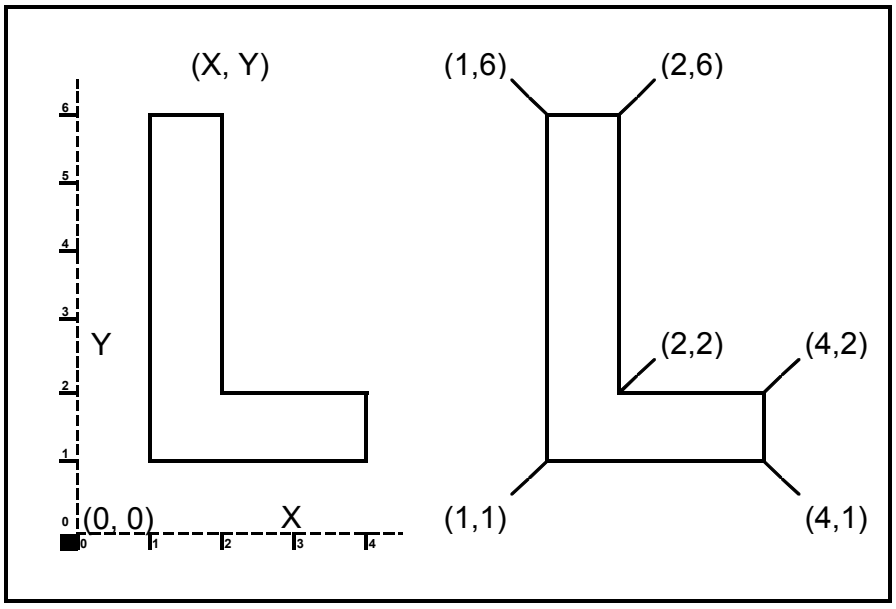


Figure 2. An 'L' produced by the following outline vector list coordinates: {(1,6); (2, 6); (2, 2); (4,2); (4,1); (1, 1)}

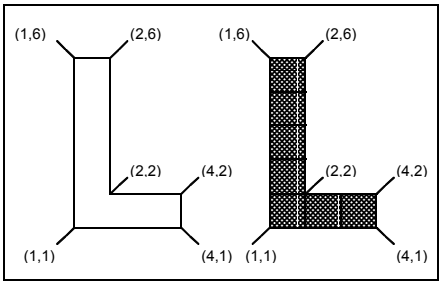


Figure 3. Filling-in an 'L' outline font with large pixels, creating a computer generated 'L' character glyph

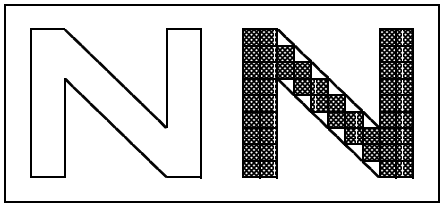


Figure 4. Filling in a second 'L' with smaller pixels

For the computer generated letter 'L' glyph in Figure 4., pixel size is not much of an issue.

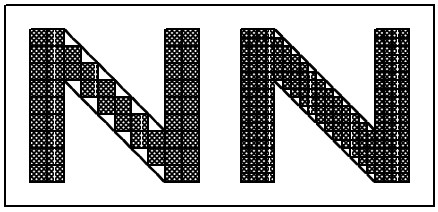


Figure 5. Filling-in an 'N' with large pixels

In Figure 5., large pixels leave a jagged edge on the diagonal of the computer generated 'N' character glyph.

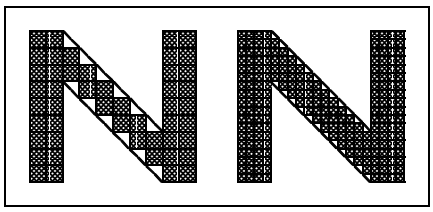


Figure 6. Filling in a computer generated 'N' character glyph with smaller pixels

In Figure 6., using a smaller pixel size improves the appearance of a 'N' glyph (by reducing the jaggedness of the diagonal stroke) created using an outline font.

These 'filled-in' character outlines (computer generated glyphs) of the 'L' and the 'N' represent the technology of outline fonts invented in the 1970s by John Warnock, now President of Adobe systems [http://www.Adobe.com]. All laser and ink jet printers use this technology to generate the raster bitmap images of the pages that the printers print.

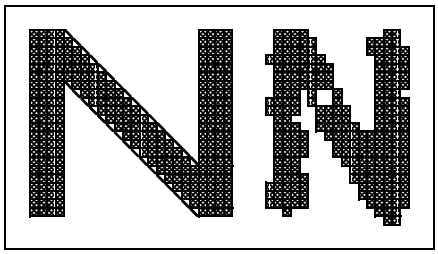


Figure 7. Computer generated glyph of an 'N' beside a raster scanned glyph of an 'N'

Note that the computer generated 'N' in Figure 7 looks much better than the raster scanned 'N' in Figure 7.

Note that the computer generated 'N' in Figure 7 looks much better than the raster scanned 'N' in Figure 7.

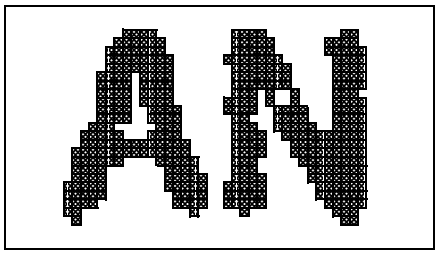


Figure 8. Raster scanned glyph of an 'A' placed next to raster scanned glyph of an 'N' of the same type font

Note that the addition of the raster scanned 'A' makes the raster scanned 'N' look much better in Figure 8.

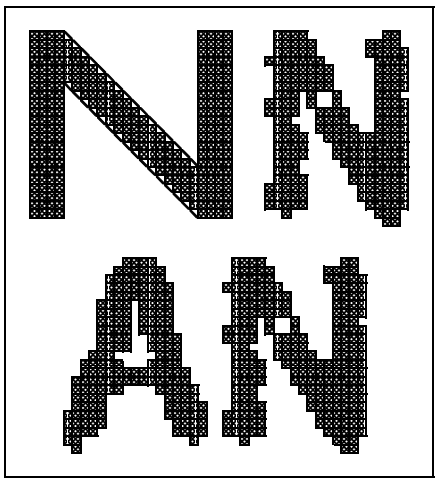


Figure 9. Computer generated glyph and scanned glyph

The two scanned 'N' glyphs in Figure 9. (One scanned 'N' from Figure 7. and one from Figure 8.) are identical.

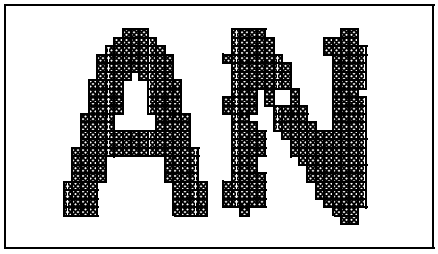


Figure 10. A computer generated 'A' glyph, that just matches the font that the 'N' glyph was scanned from

The 'A' glyph looks great (note its regularity and symmetry) and improves the appearance of the 'N' glyph in Figure 10.

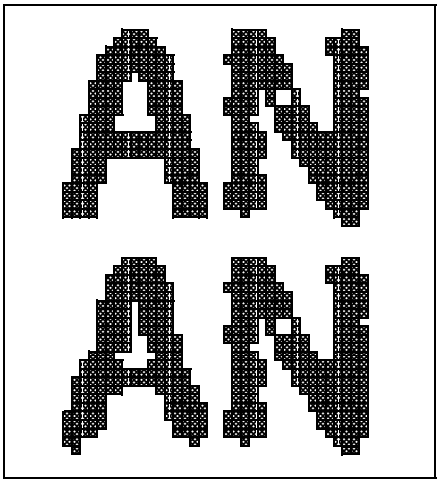


Figure 11. The two 'N' glyphs are identical, but the top 'N' glyph, next to the computer generated 'A' glyph, looks much better.

When the computer generated 'N' was placed beside the raster scanned 'N' in Figures 7 and 9., the raster scanned 'N' was made to look worse because the two 'N's were not in the same type font. When the type font was recognized for the scanned 'A' glyph, and the computer generated a matching glyph for the scanned 'A' glyph's font (which was substituted for the scanned 'A' glyph in Figures 10 and 11), the 'A' improved the appearance of the scanned 'N' glyph because the fonts of the 'A' and the 'N' glyphs matched and the overall harmony of the font caused the appearance of the scanned 'N" glyph to improve.

This illustrates the visual improvement to a page when computer generated glyphs (in the right font) are substituted for the scanned glyphs, creating the PDF Normal format. Even if some letters are not recognized by the OCR software, the recognized characters can be used to improve the appearance of the page by replacing some of the scanned glyphs. This is the technology of the Adobe PDF Normal format. Unfortunately, replacing the scanned glyphs changes the page image and this leads to questions in court about the modification of stored records

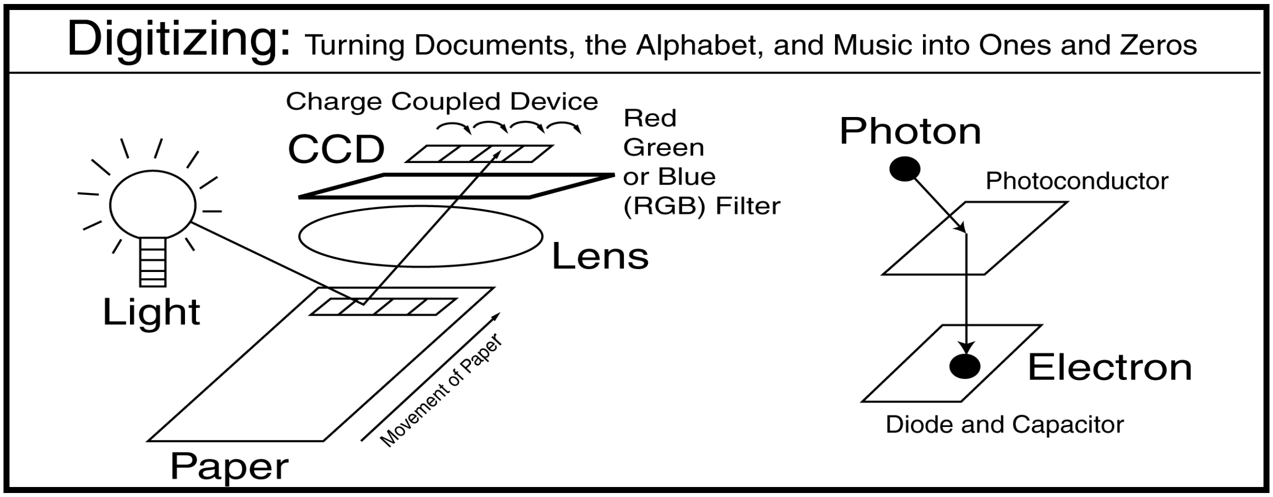


Figure 12. A portion of an image from the document *How Digitizing Works* is used here to illustrate outline fonts and graphics.

```
===================== Start of EPS (Encapsulated PostScript) =================
==================== PDL (Page Description Language) Program =================

%%Creator: Adobe Illustrator(R) 8.0
%%AI8_CreatorVersion: 8
%%For: (Steve Gilheany) ( )
%%Title: (22021v026  How Digitizing Works for Non-Technical Manager.eps)
%%CreationDate: (3/19/00) (11:13 AM)
%%BoundingBox: -222 -840 1088 775
%%HiResBoundingBox: -221.4863 -839.75 1087.6953 774.5938
%%DocumentProcessColors: Black
%%DocumentFonts: Helvetica

========================== [ 3 pages of text elided* ]  =======================

            } bind def
            mark
            /setcustomcolor where not
            {
                        /findcmykcustomcolor
                        {
                                    (AI8_CMYK_CustomColor)
                                    6 packedarray
                        } bind def
                        /findrgbcustomcolor
                        {
                                    (AI8_RGB_CustomColor)
                                    5 packedarray
                        } bind def
                        /setcustomcolor
                        {
                                    exch
                                    aload pop dup
                                    (AI8_CMYK_CustomColor) eq
                                    {
                                                pop pop
                                                4
                                                {
                                                            4 index mul
                                                            4 1 roll
                                                } repeat
                                                5 -1 roll pop
                                                setcmykcolor
                                    }
                                    {
                                                dup (AI8_RGB_CustomColor) eq
                                                {
                                                            pop pop
                                                            3
                                                            {
                                                                  1 exch sub
                                                                  3 index mul
                                                                  1 exch sub
                                                                  3 1 roll
                                                            } repeat
                                                            4 -1 roll pop
                                                            setrgbcolor
                                                }
                                                {
                                                            pop
                                                            4
                                                            {
                                                                  4 index mul 4 1
                                                                    [fixed] roll
                                                            } repeat
                                                            5 -1 roll pop
                                                            setcmykcolor
                                                } ifelse
                                    } ifelse
                        }
                        def
            } if
            /setAIseparationgray
```

```
============ [ 81 pages of text elided* ]  ============

1 0 0 1 398.8359 708.1401 0 Tp
0 Tv
TP
0 Tr
0 0 0 1 k
/_Helvetica 18.3526 17.0863 -4.1292 Tf
114.0882 100 Tz
(Photon) Tx 1 0 Tk     <<=========== Photon ===========
(\r) TX
TO
0 To
1 0 0 1 485.0859 591.1401 0 Tp
0 Tv
TP
0 Tr
(Electron) Tx 1 0 Tk    <<========== Electron =========
(\r) TX
TO
1 Ap
464.1475 597.75 m
464.1475 594.3247 460.9805 591.5493 457.0723 591.5493 c
453.167 591.5493 449.999 594.3247 449.999 597.75 c
449.999 601.1733 453.167 603.9497 457.0723 603.9497 c
460.9805 603.9497 464.1475 601.1733 464.1475 597.75 c
f
436.6592 695.5898 m
436.6592 692.1655 433.4932 689.3896 429.5859 689.3896 c
425.6797 689.3896 422.5127 692.1655 422.5127 695.5898 c
422.5127 699.0142 425.6797 701.7905 429.5859 701.7905 c
433.4932 701.7905 436.6592 699.0142 436.6592 695.5898 c
f
0 To
1 0 0 1 92.1323 541.9849 0 Tp
0 Tv
TP
0 Tr
(P) Tx 1 40 Tk    <<============ Paper ==============
(aper) Tx 1 0 Tk
(\r) TX
TO
0 To
1 0 0 1 267.2729 646.1987 0 Tp
0 Tv
TP
0 Tr
(Lens) Tx 1 0 Tk    <<============= Lens ===========
(\r) TX
TO
0 To
1 0 0 1 146.5229 698.6987 0 Tp
0 Tv
TP
0 Tr
/_Helvetica 19.2467 17.9187 -4.3303 Tf
108.7885 100 Tz
(CCD) Tx 1 0 Tk     <<============= CCD =============
(\r) TX
TO
0 To
1 0 0 1 61.4785 618.4546 0 Tp
0 Tv

=============== [ 6 pages of text elided ] ====================

Adobe_typography_AI5 /terminate get exec
Adobe_cshow /terminate get exec
Adobe_level2_AI5 /terminate get exec
%%EOF

============== End of EPS  PDL Program ==============

[*Count of pages elided is for lettersize, single column in this font and size.]
```

Figure 13. A portion of the document description (written in PostScript) for the image from *How Digitizing Works* in Figure 12

Any Graphic Symbol or Shape can be Created Using Outline Font Technology

a b c d e f g h i j k l m n o p q u r t u v w x y z

Just like a character, any graphic image consists of an outline that is filled in with a solid color. The following row (top) of graphic symbols is an example of a set of graphics that is provided as a Latin alphabet typeface, with one graphic image per letter. (These symbols can be cut and pasted into a word processor and the spell checker will mark them as misspelled text.) In this typeface (Mini Pics Lil Vehicles from Image Club Graphics [http://www.ImageClub.com]), a lower case 'a' looks like a car, a convertible, which is the first symbol on the left in the row of symbols below. If the embedded fonts of the outline font version of PDF are working, then the symbols in the top row look like vehicles. If the embedded fonts are not working, then the first row looks like the Latin alphabet and only the second row, which is a raster image, looks like vehicles.

Compression of Raster Scanned Images

All raster-scanned documents are stored and transmitted in compressed format. All compression formats are assumed to be lossless or used with a lossless setting, except MPEG (Moving Picture Experts Group), unless otherwise stated. Lossless or non-destructive compression (as opposed to lossy or destructive compression) does not change the document. That is, a decompressed document is identical to the original document before compression was done. Lossless compression is often needed to meet legal requirements for document storage. The most common form of one bit (per pixel), bitonal (The two tones of color are two shades of gray, which are black and white.), lossless compression, which is used in TIFF G4 and Adobe PDF (Portable Document Format), is the CCITT G4 (Group 4, see Bibliography) facsimile compression format. Before using any other form of compression, it is often useful to evaluate the cost savings of moving to the less common format. See also the section entitled "Identifying a Format of Record for Each Document Stored" (after Figure 15).

Chain of Document Production, From Authorship (or Scanning) to Printing

There are two means to produce raster images. One is to scan the images. These images can be stored in an Original Image PDF format file or the essentially equivalent TIFF or CCITT G4 formats. Files stored in a PDF format are more likely to be decodable because Adobe monitors the use of its trademarked 'PDF' format and attempts, by legal means, to stop the sale of products that produce badly formatted PDF files.

The second method of acquiring raster images for storage is to use outline font technology to generate the raster images using a RIP (Raster Image Processor) such as the RIP in Adobe Acrobat used to 'print-as-image'.

The following is the production chain from authorship to the printing (or storage) of the raster images: First an author uses a wordprocessor to create a document in a native wordprocessor format, such as Microsoft Word. Then the author issues a print command (by clicking on the print icon) and the Microsoft Word application program writes a PostScript (if the destination printer is a PostScript printer) PDL (Page Description Language) program to describe the document's printed pages. This is the PostScript print file that is sent to disk or to a printer. At the printer, a RIP (Raster Image Processor) interprets the PostScript PDL and generates a raster image, which is sent to the laser in the printer for printing. The following illustrates how the pixels are delivered to the paper.

Mechanism of a Laser Printer

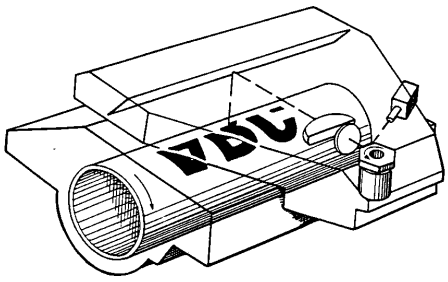


Figure 14. Mechanism of a Laser Printer

The dotted line in this picture represents a ray of light (laser beam) traveling from the laser to the drum of a laser printer.

The light leaves the laser and is reflected from a rotating hexagonal (six sided) prism (mirror). The rotation is shown by the curved arrow. The laser beam is then reflected by a second, fixed, mirror, onto the drum of the laser printer. As the prism mirror rotates the laser beam moves across the drum from end to end. (This effect can be simulated and tested with a laser pointer and a small mirror.) The laser is turned on to reproduce black pixels and off to reproduce white pixels. The drum rotates (shown by a curved arrow) to place the laser printed rows of pixels side by side, painting a raster image on the drum.

The drum of a laser printer operates in exactly the same way that a photosensitive drum in a xerographic copier operates. In a copier, in the places where light falls on the drum, the light discharges the surface of the drum, keeping the drum from attracting black toner to the drum. The unexposed portion of the surface of the drum (that received no light) attracts toner, which is then transferred to paper as the drum rotates and the paper passes under the drum. After the toner is transferred to the paper, the toner is fused to the paper by heat and pressure, creating a copy. (Laser printers in which the laser light causes the drum to attract toner are called black writers. On some printers the laser light causes the drum to stop attracting toner. These laser printers are called white writers.)

Because the toner is almost pure carbon, the xerographic copy, or laser printed copy, will last as long as the paper. This print life is over 3 hundred years if the print is made on acid free paper.

Intercepting the PostScript PDL Print File

The PostScript print file can be intercepted by printing to Adobe Acrobat. This can be done because Adobe Acrobat can be installed in a computer in such a way that Acrobat appears to be a printer. By selecting Acrobat as the printer, the PostScript PDL print file is automatically sent to Acrobat, which converts the PostScript file to a PDF file. This initial PDF file is an outline font based file. The outline font file can be converted to a raster file by using an Adobe Acrobat 'print-to-image' command to print the PDF as a raster image (generate a raster image) or by opening the PDF file in Adobe PhotoShop and exporting (by generating a raster image) the image as a raster image.

Permanent Virtual Fascicles

A TF system stores its data and metadata in permanent virtual fascicles, which are based on the idea of dividing information into fixed size pieces (fasciculation). Fascicles were used in the middle ages (That is why the multiple forms of the word 'fascicle' are in spell checkers. The word is over 5 hundred years old.), before the invention of moveable type printing, to divide books (in manuscript form, that is, handwritten) into portions (fascicles) whose size was selected based on the volume of copying a hired scribe could produce in one evening. While ignoring the meaning of the data, fascicles provided an efficient and reliable (within the constraints of manual technology) means of copying and maintaining the information (in fascicular form) (fasciculated) (stored, and copied, fascicularly).

TF systems use fascicles to divide data into equal size units to facilitate transmission through time, just as fixed size packets to transmit data through the Internet. Fixed size pixels are used for all types of digital imaging to represent all sizes of images. (The digital ortho-photo of the City of Los Angeles will contain over 48 billion six-inch-square-pixels (48 GigaPels).) As in both Internet packets and imaging pixels, the fact that the size of each part of the (fascicled) data element is fixed is more important than the meaning or context of the data contained in the data element.

Unfortunately, today, all of the processes in database or document management systems (the process of online storage, the process of online storage backup, and the process of long term preservation) use data arrangements on media that are optimized for the immediate need of each process. For this reason, the data arrangements are not optimized for the other processes. Further, none of the data arrangements are optimized for simplicity so that non-technical managers can understand and manage the processes.

Also unfortunately, online data storage uses many levels of abstraction: including RAID (Redundant Array of Inexpensive Disks), multiple levels of buffers and caches, and the invisible (transparent) migration of files around a SAN (System/Storage Area Network). Backup of online storage always assumes dynamic data, and treats all stored data as residing in a single, infinitely expandable, memory space, that must be backed up as a single unit. Long term preservation attempts to store all documents individually, and as a result has to deal with a very large number of variable size records and their multitude of physical and logical storage locations.

TF systems, with their use of fixed size fascicles, optimize online storage, backup, and long term storage needs as a system. In addition, TF systems are designed to be easily understood and managed by RIM professionals who are accomplished managers, skilled in the management of business systems and organizations. Because TF systems are relatively new, they have been optimized (designed, engineered) to take advantage of current computing economics, including magnetic disks that cost 10 US dollars per Gigabyte, single fibers that can transmit a Petabit per second (and provide, without repeaters, on each fiber, 1 billion T-1-like video channels of 1 megabit per second each: in development: [http://www.Omni-Guide.com], the coming 1 US dollar per disk blank DVDs, and the 16 X DVD readers that cost less than 100 US dollars and can restore a 4.5 Gigabyte fascicle in 10 minutes (with 100 DVD readers, at a cost of 10 thousand US dollars, 450 Gigabytes can be restored in 10 minutes, 2.7 Terabytes in 1 hour.)

In Figure 15., 1 and 2 depict a logical fascicle container constructed by an electronic signature. The fascicle contains the files that contain the documents/data and the sentinel files (document metadata) and the metadata (12) describing the fascicle and fascicle system (TF system) that explains how to use the fascicle in the system of fascicles. Also in Figure 15., 1 connotes a lid or structured entryway into the fascicle. 4, 6, 8 and 10 are the stored records (each of which can be a logical container of multiple formats of a single document, as shown in Figure 1). 5, 7, 9 and 11 are the associated sentinel records (files) containing the metadata for the associated document/data files. Sentinel files contain metadata that defines a data access database for records stored in one or more fascicles. The sentinel file metadata can itself be recorded in something like XML (eXtensible Markup Language) (See Bibliography), which would provide a means of recording metadata about the metadata. The TF system data access database can be reconstructed from the fascicles. (See also Figure 22., Fascicle Access Table, which illustrates the structure of the TF system database that is designed to hold the fascicular metadata.)

Because each data element, and corresponding data element sentinel file, is placed in a fascicle, and each fascicle will be sealed with a digital signature, none of the data element or data element sentinel files can be changed or removed

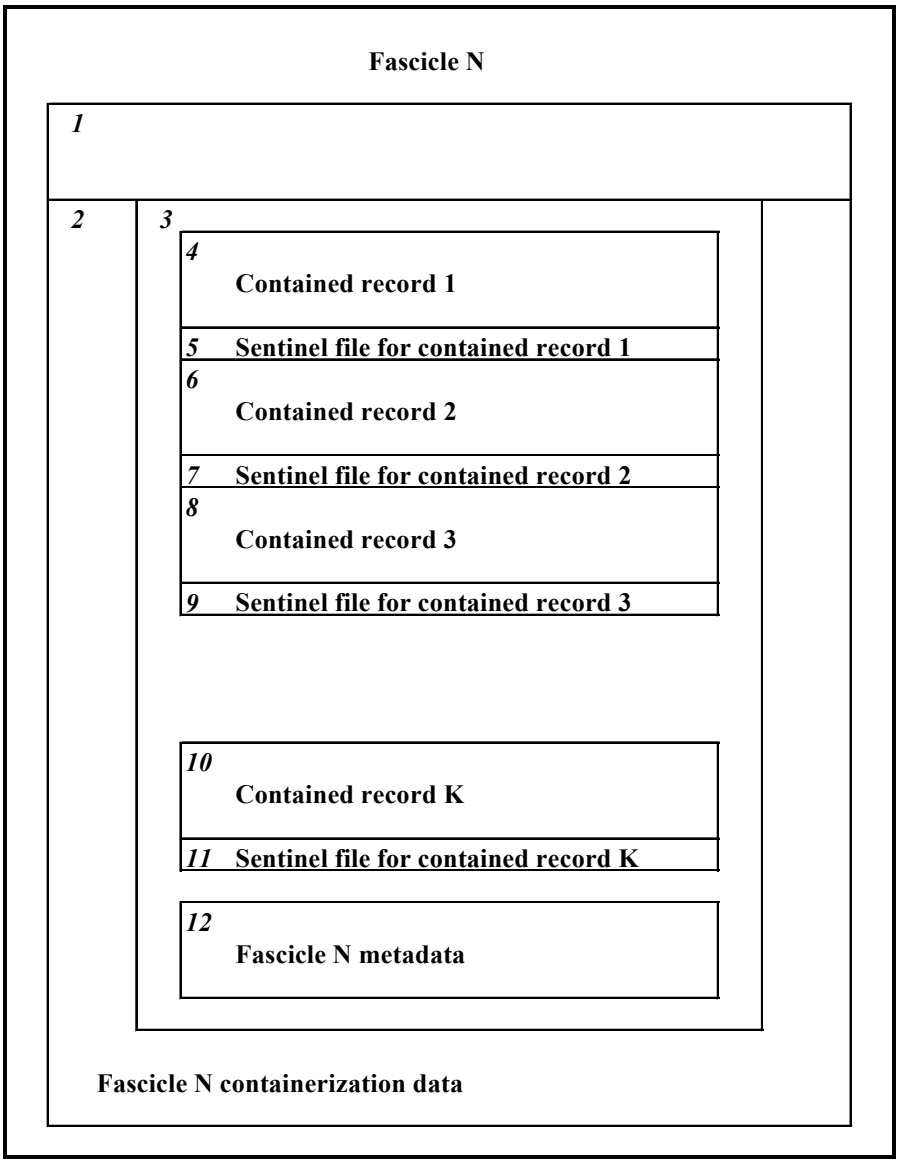


Figure 15. Permanent Virtual Fascicle

from the fascicle without detection. The fascicle can, however, be move freely around a SAN (System or Storage Area Network), LAN (Local Area Network), or an intranet (that may span the Internet). The fascicle can also be moved between computer storage media that operate at different speeds. [Without a digital seal, information on a disc, even a WORM (Write Once, Read Many) disc, can be changed easily by the simple technique of copying the disc and changing the stored data as it is copied.] Because only the checksum for a file is encrypted in an electronic signature block (when a file is electronically signed), the contents (document inside) of each file remain readable, even if the electronic signature key is lost. (Encrypted documents can also be electronically signed.)

By the choice of ECC (Error Correcting Code) and sufficiently stable media, the longevity of a stored fascicle can be made arbitrarily long, long enough to exceed the life of the universe as defined in the big bang/big crunch theory (85 billion years). Said another way, fascicles are designed (engineered) to store data forever. This is relevant because it allows non-technical managers to draw on all types of technical evaluations in determining that a fascicle will survive for the desired length of time, given any reasonable requirements. (A common configuration of magnetic disks for document imaging systems and TF systems is RAID (Redundant Array of Inexpensive Disks). RAID uses a simple, but effective and easy to understand form of ECC called parity. RAID parity is explained in paper 22022, RAID (Redundant Array of Inexpensive Disks) at [http://www.ArchiveBuilders.com]

Because fascicles facilitate the process (and greatly lower the cost) of making backups and of duplicating copies of backup disk, it is reasonable to make at least 7 copies of each fascicle for

storage in different locations (in order to ensure survival of at least one verifiably good copy of every fascicle).

For some media, for example the Norsam Rosetta media [http://www.Norsam.com] (see Bibliography), the fascicular metadata can be written in such a way that it forms a readable miniature raster image on surface of the media that can be read with a microscope, eliminating the need for any computer software to decode the fascicle. (If the pixel registration of the miniature image can be recovered, there will be no Nyquist Sampling theorem based (see Bibliography for Nyquist's paper, written in 1928) generational loss when the image is redigitized, each pixel will be read exactly as written, areally synchronously.)

Managing a TransFormat (TF) Records Management System

A TF system is designed to be as forgiving of major blunders as are boxes of paper records. The permanent fascicles cannot be changed (without detection and replacement by unchanged copies of the fascicle). This eliminates the long term effects of bad software, bad software configurations, and unreliable TF system management, for those records stored in fascicles before the problems occurred. Records stored in fascicles after the problems occurred will be affected by the problems.

As described above, TF systems are designed to be managed by RIM professionals.

A TF system is designed to be managed by asking simple questions. Questions such as:

1. How many fascicles do we have in the TF system?
2. Do we have 7 copies of every fascicle?
3. Where do you want to move these online fascicles in the TF system? And, why do you want to move them? (I know that you want to move them because you have requested permission to alter the fascicle access table, and only I have permission to alter the fascicle access table (Shown in Figure 22.).
4. Show me how you restore the TF system from the stored fascicles. How long does the restoration take?
5. Show me the tenth document in this fascicle. Show me an image from each of the formats of the document.
6. How much does it cost to add enough online disk storage to the TF system to store ten more fascicles?
7. How many fascicles have had no online access for more than a year, two years, a decade? (We can remove the fascicles, that have not been accessed, from online storage and convert those fascicles to offline storage on a manual basis if it is cost effective to do so.)
8. Can you show me the difference between the originally stored raster image for this document and the new raster image, generated from the original native format of the document, using the new viewer we are about to buy?
9. How long does it take to reload all the TF system software to begin viewing document from fascicles again?

Identifying a Format of Record for Each Document Stored

For each document submitted to a TF system (in multiple formats for storage), one of the document's formats should be identified as the format of record (official format). The best choice for the document format of record is the raster format. Raster scanned, or rasterized (using Adobe Photoshop) [www.Adobe.com] text files will probably last 5 hundred to 1 thousand years. This is because the Group 4 compression format (See CCITT G4 in Bibliography), used in the PDF and by TIFF (Tagged Image File Format) formats, is very simple and is viewed as a very light level of encryption by hackers who can easily break the code and display the document images. (TIFF was created by Aldus Inc. in the 1980s and Aldus was subsequently purchased by Adobe on August 31, 1996. Adobe now issues TIFF tags and does not release information about proprietary TIFF tags or proprietary TIFF formats.). See also the section of this paper entitled "Compression of Raster Scanned Images" (following Figure 13). The TF system manager need not identify the format of record for a submitted document, or convert a file format for a submitted document to the file format of record. In fact, the TF system administrators should not convert any documents from one format to another. The users submitting the documents should submit the documents in all of the formats recommended by the TF system administrator. It is also the user's responsibility to ensure that the digital document of record is also the legal document of record, which is in use by the TF system user's organization.

Apprising Users of Anticipated Format Conversion Problems

TF system administrators must be knowledgeable about document format change errors, and about new bugs and old bugs that are fixed in conversions between different versions of document viewers and native applications. In addition, it is the TF system administrator's professional responsibility to make users and their organizations aware of problems in document reproduction, conversion, and updates. All users should be aware that a conversion requires proofreading which is the equivalent of reauthoring or re-engineering the document. Any future migration, involving reauthoring or re-engineering, beyond simple copying, must be done by users, not by the TF system administrator.

An example of automated conversion is the difficulty of going hard metric automatically. (This is a good example of converting current documents into a possibly all-metric world of the (far) future.) The subtlety of metric conversion cost JPL (the Jet Propulsion Laboratory) 327.6 million US dollars on September 23, 1999 when the Mars Climate Orbiter miscalculated where the planet Mars was, due to a metric conversion problem, and ran into the planet. [http://Mars.JPL.NASA.gov/msp98/orbiter] [http://Mars.JPL.NASA.gov/msp98/orbiter/fact.html]

Protecting Documents from Misunderstandings Over Long Periods of Time

Protecting documents in fascicles also protects documents from misunderstandings over long periods of time. Another government agency, the US Department of the Interior, when asked if there was an inscription on Hoover Dam requesting that visits to other planets be recorded on the dam, replied that there was no such document. (Hoover Dam provides an example that is local to Las Vegas, where this paper was presented at the ARMA 2000 Conference (Association of Records Managers and Administrators, International)). Then the Interior Department provided the following transcript of an inscription on the dam (which is an attempt to communicate with persons far in the future) (The inscription is on the (non-Hollywood) Star Map (the text is in bronze letters, set into the terrazzo sidewalk) in front of the statuary entitled "Winged Figures of the Republic" from *Sculptures at Hoover Dam,* US Department of the Interior, Bureau of Reclamation, 1978): "This dam is a major structure of the times. That astronomical date line of the day of its dedication (21:30 local apparent time on 30 September 1935), imparted to future times by this monument and star diagram, is established in consequence of these theories, facts, and conclusions. When, in the course of time, the composition of our world and those other worlds in space shall be more fully known, record it here for future men {people} to see and, having seen, to speculate, investigate, and carry on the search." [http://www.HooverDam.com/service] (The plan here was that; if you want your message to last a long time, write your message on a long lasting structure, like Hoover Dam.) (Like the City of Los Angeles, described in the TF system installation below, the Hoover Dam area is working on its bridges, having concluded that the 726 foot thick concrete of the dam is no longer adequate as a roadway. (At least it is not wide enough at the top, and the approaches are narrow, steep, and winding.) [http://www.HooverDamBypass.org]

Converting the TF System Fascicle Access Table to a New SQL Database

The SQL (Structured Query Language) database based fascicle access table, for a TF system, is like a PC (Personal Computer): complex, but if necessary, you can just use another PC for your application if your PC or your PC vendor is unavailable. Similarly, you can use a SQL database from an alternate vendor if your SQL database or SQL database vendor is not available. Any SQL database can be used to reconstruct the fascicle access table from the metadata in the sentinel records on the fascicles. (For the SQL standard, see National Committee for Information Technology Standards [http://www.NCITS.org] which participate on behalf of the United States in the information technology international standards activities of the ISO.)

Computer Science has had 60 years to make computers more complex. The increase in complexity is exponential, so that the rate of increase of complexity is faster today than it has ever been. Facilitating switching between SQL database vendors is an example of how TF systems provide TF system administrators with a

means of dealing with this ever growing complexity.

Chain of Custody

In paper and microfilm based records centers, the chain of custody includes keeping track of which people had a key to the records center, and whether or not the door was locked when the records center was not occupied. With fascicles, the chain of custody is based on the management of the electronic signature used to electronically seal the fascicles. This is a digital custodial activity. Once sealed, the fascicles do not have to be physically protected, except to ensure the survival of at least one copy of each fascicle in an unaltered state for replication and document retrieval. Copying is a digital custodial activity, but format conversion and document migration are not, because format conversion and document migration are acts of authorship or re-engineering, not custodianship, and must be carried out by cognizant authors and engineers (the users of a TF system). This is why the structure of organizations does not include these non-custodial digital activities within records management, and it is why records management budgets do not include them. Non-custodial digital activities are not within the purview of records management.

Examples of Document Conversion Problems with this Document

A few conversion problems are illustrated by the documents included in the Appendix. These documents also provide estimates on the cost of magnetic disk and microprocessor capabilities over the next 10 years and the expected digital storage requirements for various types of documents. The native formats for the documents in the Appendix (along with their print parameters and sequencing) will be placed on the Internet, under ARMA 2000, in the handouts section of [http://www.ArchiveBuilders.com], so that TF system administrators can experiment with some difficult-to-convert documents. Figure 25., The Evolution of Intel Microprocessors, was created as a Microsoft Excel spreadsheet for formatting, but has been edited, after conversion to a Microsoft Word table, so that the capability to reformat the table, beyond adding rows, has been severely limited by the conversion and subsequent modification in Word. Figure 23., Projecting the Storage Costs of Magnetic Disks Over the Next Ten Years, was created in Microsoft Excel, and must be maintained in Excel because the values in the cost cells are computed based on a specific cost decline per year. The insertion of the Excel spreadsheet into Word provides a very useful shrink-to-fit function that takes into account the maximum length of text in each column. Unfortunately, when the width of the cells decreases, in the shrink-to-fit function, the height of the cells increases, so the original table has to be unreadably shrunk (in the vertical dimension) to make the table come out correctly in Word. So, the original must be maintained in Excel, but the original is unreadable in the 'transfer-from-Excel-to-Word' format.

Figure 24., Digital Image Sizes, illustrates a bug in Microsoft Word 2000. When first released, the page 'x of y' pagination produced page numbers of '1 of 1', '2 of 2', etc. instead of '1 of 2', '2 of 2', etc. The first service pack (SP1) fixed this problem, but, because PCs are bought over time, and both versions of Word 2000 are labeled Word 2000 (but have different build numbers) (The build number of the first version of Word 2000 is build 9.0.2720, located in the about submenu of the help screen of Word 2000), moving a document around an organization has unpredictable results with respect to this pagination problem. Figure 24. also may contain digital constructs, that when printed on some PostScript 2 printers, may create some digital artifacts, and that may be the result of a conflict over the interpretation of typesetting metadata between Microsoft (manufacturer of Word) and Adobe (manufacturer of Acrobat PDF). The problem appears on the last page (page 6 [A-16]) just above the section labeled "Paper, Trees, …" The number '4,294,967,296' is scrunched up, as is the word 'MegaByte' two lines above it and the words 'layers, and' on the line below it. Digital artifacts are often specific to particular combinations of hardware and software, but are not detected on most configurations of hardware and software. [By making the PDF Hidden Text raster format of the document the document format of record, the TF system user will be encouraged to located and correct as many subtle glitches (like these) as possible, before finalizing the document and submitting the recommended multiple formats of the document for preservation.] Because these problems affect the appearance of documents that users receive from TF systems, the TF system administrators are responsible for these problems.

Figures 23, 24, and 25 all have titles in reverse type with a black background and white letters. On Figure 24 the type is 48 points high. The type is smaller on Figures 23 and 25. On the smaller type, specifying reverse print causes the type to be printed black on black. (As a workaround, the type color was switched to automatic with a 10 percent tint background.) This problem may be due to a conflict between Microsoft (Word 2000, build 9.0.2720, help menu/about) and Adobe (Distiller 4.0, build 4.0 0312 08:06) on the meaning of a command in a special case.

In this document, which was created in Microsoft Word, Figures. 2. through 11. actually jump on top of one another when the cursor is backspaced from the character just before most of the images. (Figures 1. through 11., and Figures 15. through 22., are drawn in Microsoft Excel 2000.)

Example of a Document Migration Problem

In document migration, where a document is opened in a newer (or older) version of an application, the document format may change. This is a particular problem with computed fields (spreadsheet cells) (also called procedural coding statements). Microsoft is about to start a conversion to 64 bit computing to take advantage of the Itanium (and subsequent) 64 bit processors that will be available from Intel next year. [http://www.microsoft.com/PressPass/press/2000/Jul00/ItaniumPR.asp] (See also Figure 25., The Evolution of Intel Microprocessors) (The conversion to 64 bit computing may be as demanding as the conversion from 16 bit to 32 bit computing that the computer industry has recently completed.) The 64 bit processors can address 2 to the 64th power (2**64) bytes of memory. [This is much more than the 4 GigaBytes of memory that is the limit for the amount of RAM (Random Access Memory) that all 32 bit versions of Microsoft Windows can address easily today. The 4 GigaByte limit makes it impossible to keep large databases, or even application suites such as Microsoft Office, in RAM to provide access that is 100 thousand times faster (instant apparent access) than disk based data access.] When the formula =2**64 is entered in a cell of an Excel 2000 spreadsheet, the result is 18,446,744,073,709,600,000. This is different than the correct value, which is 18,446,744,073,709,551,616 (The correct value was computed in Excel 2000 with the use of an Excel IF() function). (A multiple of 2 can end in 2, 4, 6, or 8, but never in 0 (in base 10), so the first number, which ends in 0, must be wrong.)

Microsoft Excel gives the erroneous result with no warning that the value may be wrong (even though the error is well know because the actual Microsoft implementation for Excel requires extra coding to round off the answer (put zeros at the end)). In the future, Microsoft may fix this bug (stop rounding the numbers), with no notice that the bug has been fixed. This will change the value in the spreadsheet if it is opened in the new version of Microsoft Excel. If the value was saved as a number (a non-computed value, which is a declarative value) the number would be more likely to remain unchanged (and wrong) when opened by future versions of Excel. (Raster images are also declarative.)

Capturing and Preserving Data and Metadata from Incoming Media

A TF system manager removes the records and metadata from all incoming media as soon as possible and places the records and metadata in TF system fascicles. This is what the US National Archives does as well, without the electronic seals or media error correction monitoring that could be built into a TF system. See the National Archives and Records Administration Center for Electronic Records, Electronic Records Information for Archivists, Records Managers, and RIM Personnel, at [http://www.nara.gov/nara/electronic/rmirmpge.html], which describes: 1.) copying of records onto technologically current media every ten years, and 2.) an annual statistical sampling to identify any loss of data.

Preparing to Copy Fascicular Media

Like all storage and communications media, CD and DVD discs have the property that bits stored on them fade. Every day, some of the stored bits fade away. CDs and DVDs have an error correcting code (ECC) that can correct (replace) the lost bit values with their corrected values. Eventually, there are too many lost bits to be corrected. This is the basis for the estimated lifetimes of CD and DVD media. Rather than using an estimate, the ANSI/AIIM (American National Standards Institute / Association for Information and Image Management ([http://www.ANSI.org] [http://www.AIIM.org]) MS59-1996 media error monitoring and reporting standard, which complements the ANSI X3.131, media error hardware interface, provides a means of directly counting the number of bad bits (the

raw error rate, which is usually between 1 bad bit in 10 thousand and 1 bad bit in 100 thousand) on a given CD or DVD. This gives a disc-by-disc reading on when to copy the data on each disc, and indicates exactly which discs will actually last (protect the data) for the disc's projected lifetime (up to 100 years). Until commercial, end user implementations of MS59 are available for checking discs, many users are following a practice of copying CDs and DVDs every five years, regardless of the nominal warranty period.

A Time Horizon for Long Term Preservation of Electronic Records

Unless your software developers and support personnel are interested in preserving information for more than 1 or 2 years, your TF system (or any other system) will not preserve information more than 1 or 2 years. Almost all developers and support personnel have a 1 to 2 year time horizon. No matter what requirements you give developers or support personnel who have a 1 to 2 year time horizon, their implementation will have major migration problems in 1 or 2 years. This is why transferring managed documents to fascicles is important. The fascicles will survive system problems.

Preserving Paper Documents

Scanning is still expensive. If an amount equal to the lowest available cost of scanning (of the most scannable documents) was set aside for storing the physical documents, the amount of money would be sufficient to establish an annuity that would provide for the perpetual storage of the hardcopy paper documents that were scanned. (Using the lowest available commercial scanning cost of about 5 US cents per page, about 75 US dollars would be required to scan the documents in a 2,500 page box. 75 US dollars could easily provide an annual annuity of 3.75 US dollars per box, which would cover a 3 US dollar charge per year at a monthly per box storage rate of 25 US cents.)

Backup

Because TF system records (on fascicles) do not change, no baseline, and corresponding (complex) incremental backups (and media rotations), are necessary. All that is required is that as new fascicles are filled, the new fascicles are digitally sealed, and at least 7 duplicate copies of each of the new fascicles are made for offsite storage. The offsite storage should be at 7 or more spatially diverse (different locations) and managerially diverse (different vendors, not under the same management or ownership) sites.

By eliminating the complex incremental backups, it is much easier to locate and delete documents that have reached the end of their retention period, and to locate documents that had a permanent retention when stored, but that now must be destroyed because of a change in law or a mistake. Records can be deleted from fascicles by rewriting the fascicles, but because the rewriting process can allow for changes to any record in the fascicle (putting the contents of the entire fascicle at risk), the procedure must be carefully defined and well documented when carried out. To maintain integrity, a TF system must have a written policy for deleting any record.

Also avoided, by using static fascicles, is a 'keep forever' cycle of baseline backups that is necessary to avoid a worm, where a little of the database is nibbled away, day-by-day (by the software error worm, or by a virus), undetected, over a period of years. An example of a programming error worm is where, periodically, all records with a count of less than one thousand are moved to offline storage, and records with a count of less than or equal to one thousand are deleted. Each time this erroneous procedure is carried out, one more record is deleted than is stored. One record in one thousand is lost. (This problem is an example of the generic 'fence-post' problem, where, to have a fence of 4 sections, 5 fence-posts are required. And, because the fifth fence post is often overlooked, new bugs are often created.)

Disaster Plan

A fascicle based TF system is designed to survive in hibernation for years as stored fascicles, just as stored boxes of records survive for years in storage. In the less severe case of simple disk failure, a RAID array can transparently survive a single disk failure. In the case of dual disk failure, the data on the failed RAID array can merely be copied from stored fascicles to newly installed disks and the TF system will be restored.

The TF system installation described below for the City of Los Angeles has the additional requirement that the system be available during and after disasters, such as earthquakes, because the documents stored in the system are used to mitigate the effects of disasters. The planned (but not yet implement) solution is to have three or more ASP (Application Service Provider) based duplicates of the TF system on the Internet. (At least two of the three ASP sites should be at least 400 miles (650 kilometers) from the San Andreas Earthquake Fault. In addition to spatial diversity, managerial diversity (independent ownership) is also important in the ASP fault tolerant plan.) The non-volatility of TF system records (on fascicles) greatly simplifies this replicated design. Under normal circumstances the multiple TF systems share the user access load. If a site is lost, inquiries are merely directed to the remaining operating sites.

By having the system automatically fail-over to another ASP site on the Internet, system users see no indication that a disaster or system failure has occurred. This makes it possible for users to avoid doing anything they are unfamiliar with, during a disaster, when it is difficult to accurately complete even familiar tasks. (A common error when restoring a backup tape, under disaster conditions, is to erase the backup tape. This then requires access to the predecessor of the backup tape in the media rotation scheme. This predecessor tape is known as the grandfather tape.) Users gain access to the Internet using preplanned disaster procedures, and from that point on, all access to stored records (via the duplicated, ASP based, TF systems) is exactly the same as under non-disaster circumstances.

Commonly held beliefs on the scale of foreseeable disasters are at odds with reality, even in the minds of many disaster planners. For example, the last big earthquake in Southern California (the Los Angeles metropolitan area) was in 1857, almost 150 years ago. It was the 8.3 magnitude Ft. Tejon Earthquake. This earthquake ruptured the San Andreas Fault from central California to the Cajon Pass, a distance of over 225 miles (350 kilometers). The maximum ground displacement was 30 feet (9 meters). (See "Collocation Impacts on the Vulnerability of Lifelines During Earthquakes with Applications to the Cajon Pass, California", FEMA-226 (United States Federal Emergency Management Agency), February 1992, 104 pages, prepared by INTECH, Inc., Potomac, MD, [http://www.FEMA.gov]) Even more urgent is the fact that such earthquakes have occurred every 132 years, on average, over the past 1,400 years (1857 + 132 years = 1984). From paleoseismology, these big earthquakes occurred in the years AD (anno Domini, in the year of the Lord) 671±13, 734±13, 797±22, 997±16, 1048±33, 1100±65, 1346±17, 1480±15, 1812, and 1857 (Sieh, *et al.*, 1989, *Jour. Geophys. Res*., 94, 603-623). The USGS (United States Geological Survey) [http://www.USGS.gov] forecasts that there is an 85 percent chance that a major earthquake (>7.0 magnitude) will occur somewhere on the San Andreas Fault in Southern California within the next 30 years.

Policies, Procedures, and Logs

Like an ISO 9000 [http://www.iso.ch/9000e/9k14ke.htm] (See Bibliography) certified organization, a TF system needs written policies, procedures, and logs of operation that provide a means of verifying the chain of custody of records (auditing). Versions of the policies and procedures, the versions of the TF system website, and their associated version control system, are part of the log of TF system operation. All of this information is managed as part of the TF system record series containing TF system metadata records.

System administration and user instructions should be created using software help tools. All software and text documents should be managed in a version control system. The actual documents will be stored in fascicles, just as all TF system maintained records are stored in fascicles, so that when the software help tools and version control systems are no long available, the stored records can be transferred to the then current software help tools and to the then current version control system.

Charge-Back for System Usage

One of the creeping capabilities to be kept out of the TF system is charge back. This function can be handled by one of the more complex systems (such as a document management system) that access permanent documents in the TF System.

GASB 34 Changes

Recently (June 1999), the US GASB (Governmental Accounting Standards Board) issued GASB Statement No. 34, *Basic Financial Statements—and Management's Discussion and Analysis—for State and Local Governments* [http://www.rutgers.edu/Accounting/raw/gasb/repmodel/index.html] which, among other topics, puts into effect changes that require government entities to account for the physical infrastructure, and its state of repair (and the quality of its maintenance). This accounting requirement could be extended to include the loss in value that

infrastructure components suffer when the plans (records) for the components are lost, or when plans are not drawn up correctly in the first place. Work order changes caused by lack of information, or erroneous information, can amount to ten percent or more of a project's value, often amounting to millions or even tens of millions of US dollars on a single project. Lost plans affect the specific project that becomes undocumented due to the loss, but lost plans also decrease the value of the general infrastructure, creating brown fields of undocumented uncertainty that generally affect the costs of all infrastructure changes in the area. (Projects may run into forgotten undocumented substructures that must be avoided through expensive project change orders, or modifications to existing structures must be unnecessarily robust to accommodate the possibility that the structures (or related structures) are not as strong as they appear to be on the surface.) For large governmental entities, an infrastructure valuation change (decrease) of ten percent, based on the lack of plans, the poor quality of records, or the poor quality of the maintenance of records, could easily exceed one billion US dollars. The magnitude of the value that could reasonably be placed on missing or inadequate plans could reach tens of billions or even hundreds of billion of US dollars in the United States. Internationally, the newly identified cost of lost or badly managed records could be much higher. These figures could increase the focus on the need for quality records management and on the need for the newly available capabilities of TF systems.

GASB 34 could be further enhanced by building in a recognition of the fact that any improvement in record keeping, or improvement in the quality of plans, provides a double increase in the value of a government's infrastructure. The first increase in value is the increase in the value of the structure for which the plans or record keeping are improved. The second increase in value is the increase in the value of the infrastructure in the area that is affected by the brown field status of the structure that has some degree of undocumented uncertainty. GASB 34 could be extended to encourage governments to charge, to a special cost-of-poor-record-keeping account, costs related to recreating as-built plans, the costs for work order changes to work-around and to document undocumented problems and fixes, and for rebuilding structures with newly drawn plans. These costs could be counted as assets twice, once for the structure that is now better documented, and once for the surrounding infrastructure that is now less impacted by the structure's brown field status that resulted from undocumented uncertainty. This special cost-of-poor-record-keeping account would collect together the costs of poor document production, or reproduction, and the costs of poor record keeping. These special accounts would help to establish the actual magnitude of the value of good plans and good record keeping.

While it is possible that the general level of funding for TF systems may increase, it is important to note that the design criterion that TF systems survive budget cuts gracefully is a very useful one, even during times of high volume funding. No matter what the funding level, TF systems (and all systems) should be designed to deliver value for every increment of funding, and to survive any and all funding reductions while providing full value for whatever magnitude investment is actually made.

Coming Document Management Industry Changes

TF systems support document management systems as higher, more complex systems that use (call) the TF systems. Users can access documents stored in TF systems through a document management system. TF systems also shield the documents stored in the TF systems from changes in document management systems, and from changes among vendors supplying document management systems.

The Microsoft document management system, code named Tahoe, is now (summer of 2000) being demonstrated in beta, tightly coupled to Office 2002 (projected designation) (code named Office 10), which is being demonstrated now in beta and is due out in the Fall of 2001. Also available in beta, and due out in the fall of 2000, is Microsoft Windows ME (Millennium Edition). Windows ME manages document scanning for Windows users, and Tahoe will manage their documents. Microsoft has a long history of bringing order and homogeneity to markets (and causing a consolidation of vendors). Microsoft establishes standards (Microsoft standards). This was the case with the Microsoft based (hosted) AutoCAD system in the early 1980s, with Microsoft Word in the 1990's, and with the Microsoft Windows operating system. Microsoft is now describing how Tahoe will bring order to document management. [http://www.microsoft.com/presspass/press/2000/Oct00/MECPR.asp] Tahoe will be fully integrated with the Microsoft Office suite, and Microsoft may even have a version that comes free with Office, just as Outlook (which comes free with most PCs) prepares Microsoft customers for the more comprehensive mail, calendar, and document management features of Exchange. [http://www.microsoft.com/presspass/press/2000/Oct00/ExchangePR.asp] A further integration of Microsoft products was mentioned by Steve Ballmer, President of Microsoft, speaking in Long Beach, California, at a Microsoft TechNet Briefing, (Microsoft event number 25316, [http://events.microsoft.com]), on August 12, 1999, where he said that in the next release (of both the Microsoft Windows operating system and of Microsoft applications), after the 2000 release, Microsoft planned to merge the Windows 2000 file system, the Microsoft SQL Server database, and Exchange (the equivalent of Outlook on servers).

If Tahoe document management features are included as a free addition to Office, then the 120 million existing Office customers will have Tahoe features available as soon as they upgrade to Office 2002. If Tahoe features are added to Outlook, then almost everyone buying a new PC (or getting a 15 US dollar replacement disk for the Microsoft software on their PC) will have access to the Tahoe document management features. [http://www.Microsoft.com]

In addition, Microsoft plans to merge Microsoft IE (Internet Explorer) with the Microsoft Network (MSN) Portal, thereby creating MSN Explorer (now in second beta) to assist customers in locating Microsoft software, such as Office and Tahoe, for rent over the Internet. If the Microsoft software is not free (As Microsoft Word is, as part of the Microsoft Works package that is given away free with the purchase of most new PCs purchased with the Microsoft Windows operating system), the software can be rented for just the cost of a single use. (Microsoft is committed to ubiquity. Microsoft has contracted for the right to use MainSoft.com products and professional services to port Windows Media Player, Internet Explorer, and Outlook Express to various forms of Unix, including Linux. Microsoft is gaining the experience it would need to port the Microsoft Office suite to Unix, including Linux.) Microsoft has invested in Corel [http://www.microsoft.com/PressPass/press/2000/Oct00/CorelPR.asp]. This will help Microsoft bring WordPerfect and Corel Linux into the Microsoft.net fold. [http://www.zdnet.com/zdnn/stories/comment/0,5859,2636002,00.html] [http://www.zdnet.com/enterprise/stories/main/0,10228,2640025,00.html]

Microsoft is planning to expand the market for document management. This will expand the market for TF systems while consolidating the market for document management systems.

TF System Website Maintenance

TF systems, like all modern software, will provide intranet and perhaps Internet access to records. Maintenance of the TF system website that provides this access should be integrated with the metadata maintenance for general records management. For example, the descriptions of records series and their finding aids should be loaded onto the website automatically, from the TF system metadata. If the website information (metadata) is maintained separately from the TF system metadata, then there will be two databases that are supposed to contain the same information. As is always the case, according to Murphy's Law, if any piece of information is in two databases, it will be different in the two databases. Copying information from one database to another doubles the maintenance burden, at least. All displays of information should be computer retrieved from one master copy of the information.

Application of TransFormat (TF) Records Management at the City of Los Angeles

The Bureau of Engineering (BOE) [http://www.CityofLA.org/BOE/index.htm], Department of Public Works, City of Los Angeles, designs the City of Los Angeles' infrastructure and buildings, maintains the record of those designs, and manages the City's GIS system, which is being expanded to tie together the City's engineering records, and searches for those records. The BOE, located in the heart of the extremely photogenic, on-location filming district known as the Central Core, where many of the scenes in the movie Blade Runner were shot, is installing a fascicle based system to provide Internet/intranet access to approximately one million engineering drawings and maps. The fascicles will be approximately 4.5 GigaBytes in size to facilitate movement to DVD. Multiple fascicles could be stored on the 1 thousand year

ion milled nickel media from Norsam.com, which have now been delivered commercially. The iridium version of the ion-milled media is planned to last over 1 billion years. (The 1 thousand year Norsam nickel disks must be stored at or below 300 degrees Centigrade (at or below about 550 degrees Fahrenheit).)

For online access, the images will be stored in a JEMSdata.com RAID (Redundant Array of Inexpensive Disks) configuration in a 19 inch rack that can store one hundred twenty 73 GigaByte Seagate.com fiber channel hard drives, that have a transfer rate of 400 MegaBytes per second, and that can use a dual fiber channel 2 (FC2) interface and dual controller configuration for fault tolerance. The fiber channel controllers can control 120 drives so that the disk drives can be purchased on an as needed basis, as scanning is done, with up to 8 TeraBytes of RAID storage per 19 inch rack. RAID sets can be any size, up to 120 drives per RAID set, and hot spares can be shared across all RAID sets within the 120 drives on the dual fiber channel controllers. The fiber channel fabric that allows failover between servers in a clustered server configuration also supports simultaneous connections to the Sun Solaris, IBM AIX, Windows NT 4.0,Windows 2000, and the Apple Macintosh operating systems (The design on which City's configuration was based was originally optimized for non-linear video editing.). The fabric allows access to a maximum of 128 thousand of the 8 TeraByte RAID racks for a maximum configuration of 1 ExaByte. JEMSdata.com is also a source for DVD writers. Aperture cards are being audited for arrangement using a Cardamation.com punch card reader and scanned using a Tameran.com aperture card scanner. ZumaCorp.com and IA-Info.com have provided microform scanning assistance.

Engineering drawings, maps, and other documents will be indexed using the City GIS (Geographic Information System), creating a spatial index to the documents. In addition, polygons defining the extent of projects or other geographically extensive indexed items can be added, using all of the polygonal attributes available in the GIS system from ESRI.com (Environmental Systems Research Institute). Also planned for use in document indexing are existing paper and microform indices that have been scanned in, and relational databases which may include the use of Hansen.com specialized municipal management databases. Scan-on-demand service and day-forward scanning will make it appear that all documents are available on-line, long before the backfile conversion has been completed. LizardTech.com and ERmapper.com are sources for software to facilitate the transmission of large format images over the Internet. In addition, storage of CAD (Computer Aided Design) files, digital orthophotographs, remote sensing images, word processor and other office suite documents, inspection photographs, and sewer inspection videos (including the enormous volumes of emergency sewer inspection videos required after major earthquakes) is also planned. By continuing to follow established procedures, all microfilming will be kept current so that all engineering drawings and maps will be in microfilm format when the backfile conversion is completed. All engineering drawings and maps will then last at least 5 hundred years in microform. (See Bibliography for requirements for a 5 hundred year microfilm life expectancy.) Also, the management commitment required to have a high integrity microform system also supports the creation of a high integrity digital records management system (TF system).

The use of all of this technology might have caused the management commitment, and the existence of a project champion, to be overlooked, even though the commitment and the project champion are the most important parts of the project for ensuring success. Mr. Clark Robins, with 35 years of service to the City of Los Angeles, has built many bridges, both literally and figuratively, in his leadership role in the CAVRS (Computer Access to Vault Records System) TF system. Clark was in charge of the 18 million US dollar, ten year project that created the City of Los Angeles land-base (entry of all land parcels in the City) for the City GIS. Clark has always had excellence as his goal, whether it was in establishing the GIS land base or in his recently completed project to seismically retrofit and meticulously restore twelve historic bridges across the Los Angeles River and three other historic bridges that included the Shakespeare Bridge that provides access to the neighborhood where Walt Disney lived. The bridges, which are among the largest and most beautiful concrete arch bridges in the world, were built as part of the City Beautiful movement from 1909 to 1932. The retrofitting and restoration of the 15 historic bridges, a 78 million US dollar project, was part of a larger project that Clark managed to seismically retrofit 118 bridges for 146 million US dollars.

Most of the historic bridges were built by Merrill Butler, Engineer of Bridges and Structures for the City of Los Angeles from 1923 to 1963, who left the City's service only two years before Clark Robins arrived to build and rebuild bridges and to build a GIS and records management system. Clark has received many awards for the retrofitting and restoration of the fifteen historic bridges, as the head of the City of Los Angeles Structural and Geo-Technical Engineering Division. This last summer (Summer 2000), Clark's bridges and their restoration were documented by the Los Angeles River Bridges Recording Project of the Historic American Engineering Record [HAER http://www.cr.NPS.gov/habshaer], National Park Service (NPS) [http://www.NPS.gov], United States Department of the Interior (DOI) [http://www.DoI.gov] and will be permanently available on the Internet as part of the United States Library of Congress [http://www.LoC.gov] American Memory Project [http://memory.LoC.gov] after April 2001. Hopefully the City's documents will someday be linked to the holdings of the Archives of the Indies (Archivo General de Indias) in Seville, Spain [http://www.mcu.es/lab/archivos/se2000_spain/se20001.htm] that include many documents related to the founding of the City of Los Angeles in 1781.

We talk about paradigm shifts, and new paradigms, to become familiar with the important, and world defining, concept of paradigms. TF systems are a new paradigm. The records of historic structures (such as the HAER records), taken together, illustrate the holistic unity and diversity of eastern philosophy, of what can only be seen, in western philosophy, as total integration and inter-dependence of our enterprise in a free world. The Historic American Engineering Record (HAER) documents record an accommodation and celebration of diversity of thought, ideas, beliefs, and expression, that sprang from the paradigms of paradigms that make up our diversity. From this accommodation and celebration of diversity, we may see the wisdom of accommodating our world's paradigms of paradigms: in which each person, in their own unique paradigm, accommodates, recognizes, and celebrates the existence of every other person's unique paradigm. This accommodation, recognition, and celebration is essential to the advancement of the world into an otherwise uncertain future. This is why we seek and preserve knowledge, manage documents, store records, build archives and museums to preserve the records and artifacts of our deeds and discoveries, and build libraries to provide for the diffusion of knowledge. And, it is why we celebrate the freedom of the Internet as we work our way up from discernable differences (of voltages on wires, of light/no-light on optical fibers, of pit/no-pit on optical discs or ion milled nickel or iridium surfaces), to bits, to data, to information, to knowledge, and hopefully, to wisdom (as only people, and not computers, are able).

Success requires a strong commitment to excellence, perseverance, and good-natured encouragement. In short, success requires an effective project champion.

Appendix

Bibliography (Most titles followed by URLs (Universal Resource Locators) are available free on the Internet)

Adobe Systems, *PostScript Language Reference*, Third Edition, Addison-Wesley, Menlo Park, CA, 1999, ISBN 0-201-37922-8 [http://www.adobe.com/products/postscript/pdfs/PLRM.pdf]

Adobe Systems, *PostScript Language Reference Supplement,* Adobe PostScript 3, Version 3010 and 3011, Product Supplement, 1999, 28-007 [http://partners.adobe.com/asn/developer/PDFS/TN/PS3010and3011.Supplement.pdf]

Adobe Systems, Inc., *Adobe Type 1 Font Format,* Addison-Wesley, Menlo Park, CA, 1990 ISBN 0-201-57044-0 [http://partners.adobe.com/asn/developer/PDFS/TN/T1_SPEC.PDF]

Adobe Systems, *Type 1 Font Format, Supplement,* 1994, Technical Specification #5015, PN (Part Number) LPS5015 [http://partners.adobe.com/asn/developer/PDFS/TN/5015.Type1_Supp.pdf]

Adobe Systems, *Portable Document Format Reference Manual*, Version 1.3, Addison-Wesley, Menlo Park, CA, 1999, ISBN 0-201-62628-4 [http://partners.adobe.com/asn/developer/acrosdk/DOCS/pdfspec.pdf]

Adobe Systems, *Display Postscript Manual*, 1993 [http://partners.adobe.com/asn/developer/PDFS/TN/DPS.refmanuals.PSW.pdf]

Adobe Extreme, brochure (integrates PDF and Postscript Printing), 1998, Adobe Systems [http://www.adobe.com/products/extreme/pdfs/extremewp.pdf]

Ried, Glenn C., *PostScript Language Program Design*, Addison-Wesley, Menlo Park, CA, ISBN 0-201-14396-8 1988

Adobe Systems, *PostScript Language Tutorial and Cookbook* Addison-Wesley Menlo Park, CA, 1986 ISBN 0-201-10189-0

Adobe Systems, *PostScript Language Document Structuring Conventions Specification*, Version 3.0, Adobe Systems, 1992, PN LPS5001 [http://partners.adobe.com/asn/developer/PDFS/TN/5001.DSC_Spec.pdf]

Document Imaging and Document Management, raster glyphs (Figures 2. through 11.) and Figures 1. and 12. through 26. are from this 3 day class, which is taught quarterly in the UCLA Extension program, by Stephen J. Gilheany, of ArchiveBuilders.com, and are used with permission. [http://www.UnEx.UCLA.edu/catalog] (key words for class search: "Document Imaging Document Management") [http://www.ArchiveBuilders.com]

Jeff Rothenberg, 'Ensuring the Longevity of Digital Documents', *Scientific American*, January 1995, Vol 272, No 1, pages 42-47

Preserving Information Forever and a Call for Emulators, How Digitizing Works; Sizing a Document Management System: Image Size Estimates for All Types of Digitized Documents, Document Management System Search Techniques: the More, the Merrier, Paper Sizes and Paper Weight: Metric and US Standards, Disaster Planning for Document Management, COLD, COOL, COM, Greenbar, and Your Bank Statement, The Use of Future Digital Data Sources in Land Use Planning Documents, How RAID (Redundant Array of Inexpensive Disks) Works, The Next Three Years on the Internet, How the Internet Works, DVD Does Not Stand for Digital Video Disk, Microsoft Evolution: The 3.1 Flavors of Windows 2000 Become the Microsoft Environment: The History of Microsoft and Its Products, (and 17 illustration and 732 PowerPoint slides) by Stephen J. Gilheany, [http://www.ArchiveBuilders.com]

"Keeping Merrill Butler's dream alive, restoring Los Angeles' bridges", by Lewis MacAdams, *Los Angeles Conservancy News*, March/April 1999, page 1, 4 [http://www.LAConservancy.org]

"Shakespeare Bridge's new Life Celebrated", by Matea Gold, *Los Angeles Times*, April 20, 1998, page B-1, LATimes.com

George Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", *The Psychological Review,* 1956, 63:81-97. The text is available on the Internet at: [http://cogprints.soton.ac.uk/documents/disk0/00/00/07/30/cog00000730-00/miller.html]

ISO 9000 International Quality Standard (International Standards Organization) [http://www.iso.ch/9000e/9k14ke.htm]

Matrix, Warner Brothers, 1999; *Blade Runner*, Warner Brothers, 1982 [http://www.WB.com]

Harry Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans.*, AIEEE, Vol. 47, April 1928, pp. 617-644.

The *Rosetta Stone* is part of the collection of the British Museum, London, collection number EA 24. "[The Rosetta Stone] ends by saying that it is to be made known (in March, 196 BC, Before Christ) that all the men {people} of Egypt should magnify and honor [King] Ptolemy V, and that the text should be set up in hard stone, at multiple locations, in the three scripts which the Rosetta Stone still bears today (hieroglyphic, Demotic, and Greek)". Thus, the Rosetta Stone contains its own metadata, and a single document reproduced in three formats that are locked (sealed) together in stone, for the purpose of causing the message to last a long period of time; like a permanent virtual fascicle. The Rosetta stone even specified a spatially diverse pattern of storage, which increased its physical longevity. The following provides a Rosetta Stone history: [http://www.thebritishmuseum.ac.uk/egyptian/ea/gall/rosetta.html] The following provides an explanation of the Rosetta Stone text: [http://www.thebritishmuseum.ac.uk/egyptian/ea/further/rosettasay.html]

CCITT Group 4, The CCITT (Comité Consultatif International pour le Télégraphe et le Téléphone) (International Telegraph and Telephone Consultative Committee) is now a part of the ITU (International Telecommunications Union) [http://www.ITU.int] The G4 ITU recommendation T.6 (11/88), Facsimile coding schemes and coding control functions for Group 4 facsimile apparatus, is on pages 48-57 of the CCITT Blue Book, Volume VII - Fascicle VII.3, Terminal Equipment and Protocols for Telematic Services, Recommendations T.0 - T.63, ISBN 92-61-03611-2

For *SGML* (Structured Generalized Markup Language), *HTML* (HyperText ML), *XML* (Extensible ML), and *CGM* (Computer Graphics Metafile): see OASIS (Organization for the Advancement of Structured Information Standards) [http://www.OASIS-open.org] W3C (World Wide Web Consortium) [http://www.w3.org/XML]

Microfilm life expectancy of 500 years - *ANSI/NAPM IT9.1-1992 Imaging Media (Film)-Silver-Gelatin Type-Specifications for Stability* gives the maximum concentration of residual thiosulfate in microfilm that will allow for a microfilm life expectancy of 500 years. (American National Standards Institute [http://www.ANSI.org], National Association of Photographic Manufacturers [http://www.techexpo.com/tech_soc/napm.html]). For film storage requirements, see Kodak.com at [http://www.kodak.com/cluster/global/en/consumer/products/techInfo/e30/e30Contents.shtml]

A full text searchable, and a hyperlink clickable copy of this paper, along with updates and background material for this paper, *Permanent Digital Records and the PDF Format: Defining a Permanent TransFormat Records Management System, A Hierarchy of Record Storage Formats, Five PDF Formats, and Document Copying/Migration* presented at Session T-204, 10:30 AM to 11:45 AM, Tuesday, October 24, 2000, ARMA 2000 (Association of Records Managers and Administrators, International) [http://www.ARMA.org], October 23 to 26, 2000, in Las Vegas, Nevada, USA, is available at [http://www.ArchiveBuilders.com] (AM, ante meridiem, before the meridian, PM, post meridiem, after the meridian, w.r.t. (with respect to) the point in time when the sun is directly overhead at your local meridian, as modified by the creation of time zones in the United States, which was managed by William F. Allen, secretary of the General Time Convention and the American Railway Association (ARA) and, managing editor of the Official Guide of the Railways (text prepared for the Internet by Carsten Möller using the text from a Santa Fe Railroad publication by Carlton J. Corliss, *The Day of Two Noons,* from the Association of American Railroads (AAR) a successor to the ARA, 1952 [http://www.fremo.org/betrieb/timezone.htm] [http://www.AAR.org]), and occurred at Noon, in each time zone, November 18, 1883 (Mr. Allen's accomplishment is commemorated by a large bronze plaque In the waiting room of Union Station, In Washington, DC) and worldwide on November 1, 1884, by the International Meridian Conference in Washington, DC. [http://physics.nist.gov/GenInt/Time/world.html] [http://physics.nist.gov/time] The introduction of time zones represents a completed (in the distant past), and forgotten, paradigm shift. (When a paradigm shift is complete, the old paradigm disappears.) This can be seen because no one has a noon mark on their kitchen floor any longer (showing where, when the Sun is vertically overhead, the rays of the Sun pass vertically through a slit and illuminate the noon mark which is drawn (registered) on the meridian local to the person's kitchen, establishing the local solar time for the kitchen), as was common at the time of the introduction of time zones. No longer is an array of clocks intended to show the proprietary times of multiple railroads (which the railroads usually took from the solar time of their headquarters city). An array of clocks now shows the time in important time zones. When time zones were introduced, they came with the paradigm of artificial time, rather than solar time, which was tied directly to the Sun) (Similarly, by about the year 1500, the end had come to the convention of drawing maps so that maps were right reading when the maps were oriented with the East at the top.)

Glossary

Glyph – the image of a character rendered in pixels.

Raster – the scanned image created by a kinescope (a CRT, Cathode Ray Tube, such as that used in computer displays)

Pixel – (PICture ELements) or pels (Picture ELements), an image sample area that is almost always square. Arranged in a grid, pixels form a raster image. A scanned page of a paper or microform document creates a digital image that is a raster of pixels. The RIP (Raster Image Processor) in a printer produces a raster of pixels from the PDL (Page Description Language) files (PDL files are actually interpretable computer programs.) sent to the printer for printing.

The pixels most commonly used to represent images as a computer file are of a uniform size and shape. The pixels do not overlap (they are non-imbricated), and they abut (touch) adjacent pixels on all four sides. All of the pixels of an image use the same digital format to express the numeric value for the portion of the image that each pixel represents. For scanned textual documents, each pixel is represented as a single binary bit that has a value of either one (white) or zero (black). Some years ago almost all pixels in computing were standardized on a square shape. Hexagonal pixels have been used in modeling economic geography. Rectangular pixels are still used in video and continue to cause problems when converted to square pixels for use in computing. For document imaging, only square pixels are used.

Text image – the content of a text record, often the contents of a page of text.

Vector Based Outline Fonts (Standard for Printing)

(For a less technical presentation, please skip this section.)

In the PostScript Language (and in the PDF Language which is derived from PostScript), all outline fonts and graphics are made up of vectors. Vectors are mathematical equations that describe a line or path from one point to another point. Vector based images are not limited to straight lines. Any curve can be represented. For example, the curves, or circles, that makes up the inner and outer edges of the image of the letter 'O' are examples of vectors. The equation {X squared + Y squared = R squared} forms a circle of radius R. This can be used to form the inside circle in the character 'O'. The outside circle of the character can be formed by the equation {X squared + Y squared = (R + T) squared}. In this equation, T is the thickness of the black stroke that forms the 'O'. This equation produces a stroke of a uniform thickness, but most typeset 'O's have a stroke of varying thickness. Strokes of varying thickness are created by using third order equations that include an X cubed and a Y cubed component. Figure 16. Crafter Edge 'O' is an example of a typeset capital 'O' modeled on a 14 Point ITC Korinna Regular uppercase 'O' in 300 dpi (dots (pixels) per inch) laser resolution, ITC (International Typeface Corporation [http://www.ITCfonts.com])

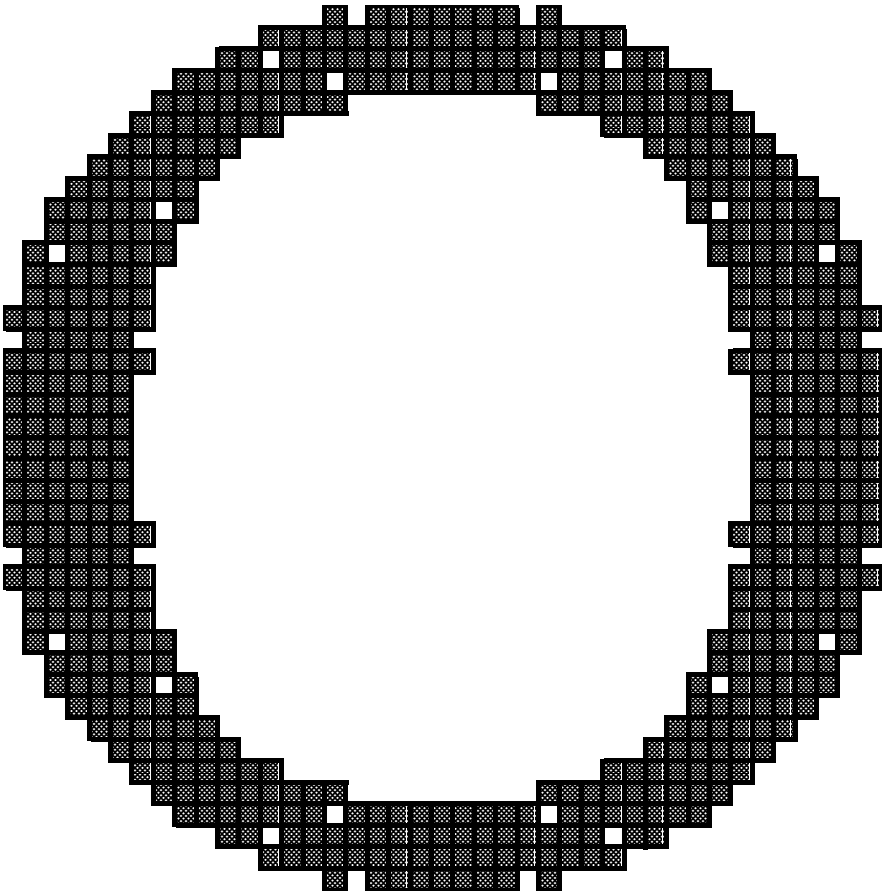


Figure 16. Crafted Edge Character: 'O'

The jagged edges of the 'O' glyph are designed to manage the physics and statistical uncertainty of the electrostatically (Electrostatics are difficult to manage; lightning is an electrostatic process.) based laser printing process. The laser beam is 'more-or-less' round, not square, and is fuzzy around the edge. Toner grains are much smaller than the printed pixels, but toner clumps together and cannot 'stick' to a single pixel. Toner 'sort-of' follows the edge of the glyph. The goal is to captain the ship (of the electrostatic laser printing system) along the polygons (edges) of the glyphs' outline fonts (character and symbols), and hope for the best.

Movement-Rotation-Scaling in PostScript and PDF

(For a less technical presentation, please skip this section.)

Vector based outline fonts (and graphics) can make use of any size pixel (for different printer resolutions). This makes the page images, composed of vector based images of graphics and characters, printer independent. Beyond providing printer independence, vector based images also have the desirable feature that they can be moved around a page, rotated, and scaled (enlarged or reduced) by simple multiplication and addition operations performed on the vectors that define the graphic elements. More complex morphing/distorting techniques can also be applied mathematically.

The simplest transformation is to multiply all graphics on a page by 2, in both the X and Y dimensions. This operation enlarges the image (page) by 100 percent. Multiplying all graphics on the page by 1/2, in both the X and Y dimensions, shrinks the image (page) by 50 percent.

The following is a presentation of moving, rotating, and scaling arithmetic functions for image transformation.

PostScript (and Acrobat PDF) transformations are represented by a 3-by-3 matrix (as seen on Page 187 of the PostScript Language Reference, Section 4.3.3, Matrix Representation and Manipulation, see Bibliography.).

$$\begin{bmatrix} a & b & 0 \\ c & d & 0 \\ t & u & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t & u & 1 \end{bmatrix} \quad \begin{bmatrix} a & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} \cos(\text{Theta}) & \sin(\text{Theta}) & 0 \\ -\sin(\text{Theta}) & \cos(\text{Theta}) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

General Matrix | Movement Matrix | Scaling Matrix | Rotation Matrix

This 3-by-3 matrix is multiplied by each of the vectors of the outline of an image. The vectors of the outline of an image are the paths that connect the vertices of a polygon and that form the outline (or polygon) that defines the image.

In PostScript, this 3-by-3 matrix is represented by the six position array object [a, b, c, d, t, u], because the three values in the last column of the matrix (0, 0, 1) are constants

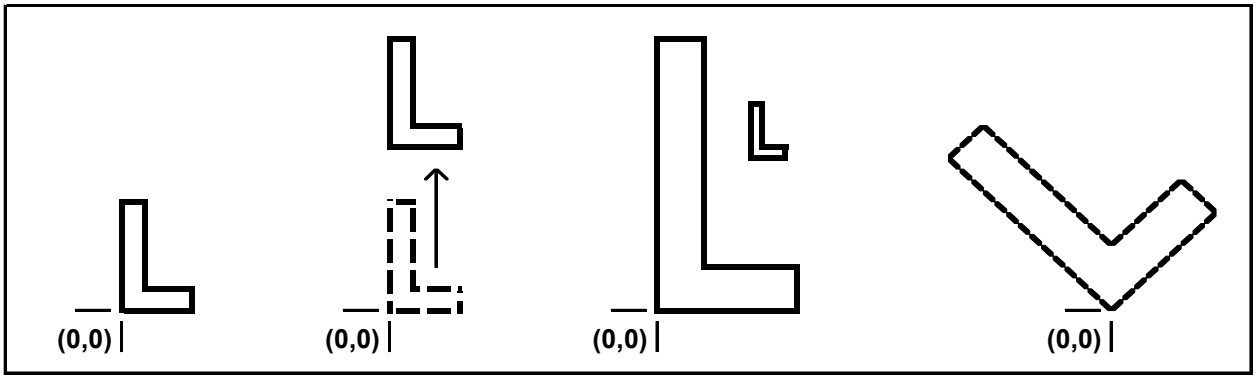
General Transformation	x' = ax + cy + t y' = bx + dy + u	The coordinate pair (x, y) can be transformed into another coordinate pair (x', y') by multiplying the (x, y) coordinate pair by the general transform matrix. (This represents a simultaneous movement, scaling, and rotation.)

In Adobe PostScript and PDF, the following transformations are supported:

Movement (Translation)	x' = x + t y' = y + u	When the movement matrix is used, a spatial translation or movement (of 't' in the 'x' dimension and 'u' in the 'y' dimension) is effected.

Scaling	x' = xa y' = yd	The scaling matrix produces scaling (enlargement or reduction of the image size) by a factor of 'a' in the 'x' dimension and by a factor of 'd' in the 'y' dimension.

Rotation	x' = cos (Theta) x + sin (Theta) y y' = -sin (Theta) x + cos (Theta) y	Use of the rotation matrix produces counterclockwise rotation of the image about the origin by an angle Theta (θ).



Original Glyph Size and Placement | Upward Movement (Spatial Translation) | Enlarge 100 Percent Shrink 50 Percent (Scaling) | Rotate Counterclockwise Through a 45 Degree Angle

Figure 17. Illustration of Movement-Rotation-Scaling in PostScript and PDF

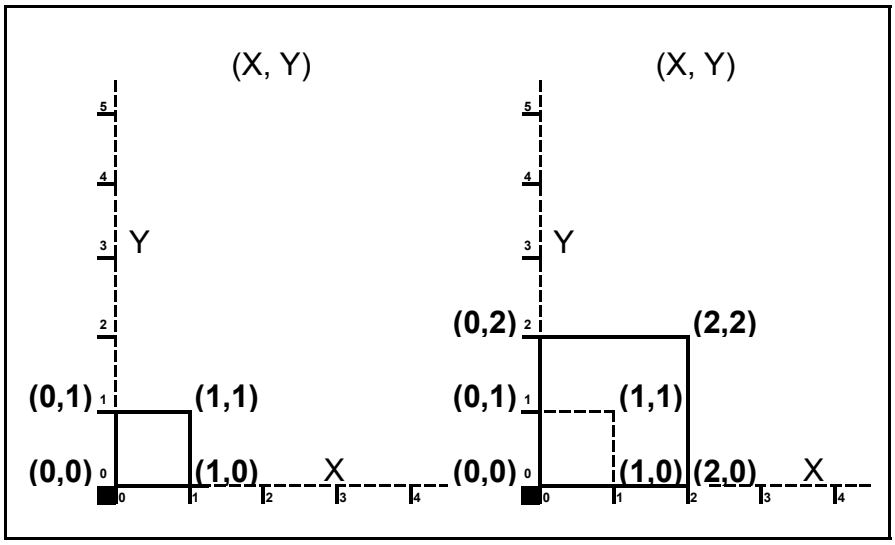


Figure 18. Scaling up of a symbol for a period

In Figure 18. the symbol for a period that is defined by the polygon defined by the vectors {(0,1), (1,1), (1,0), (0,0)} is multiplied by (2,2), doubling its size in both the X and Y dimension to {(0,2), (2,2), (2,0), (0,0)}

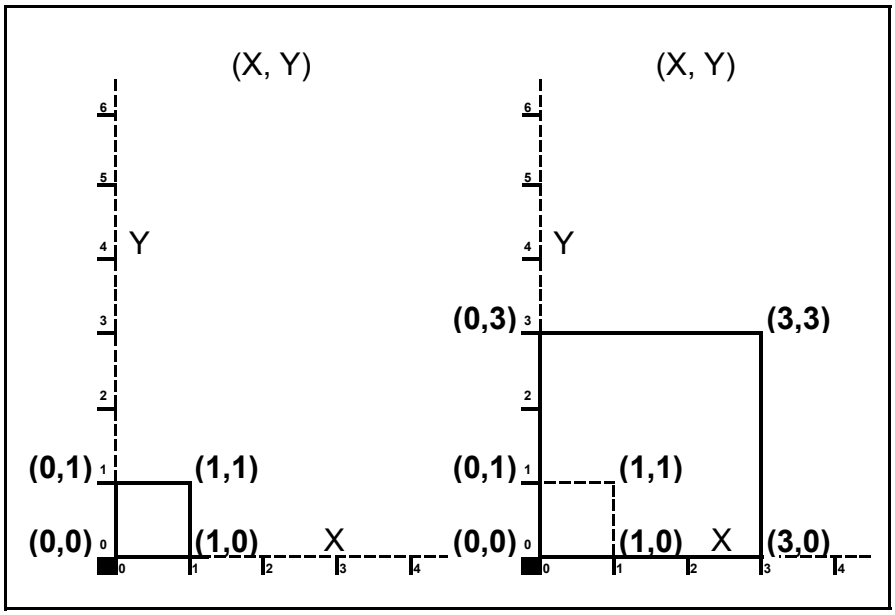


Figure 19. Scaling up of a symbol for a period

In Figure 19. the symbol for a period that is defined by the polygon defined by the vectors {(0,1), (1,1), (1,0), (0,0)} is multiplied by (3,3), tripling its size in both the X and Y dimension to {(0,3), (3,3), (3,0), (0,0)}

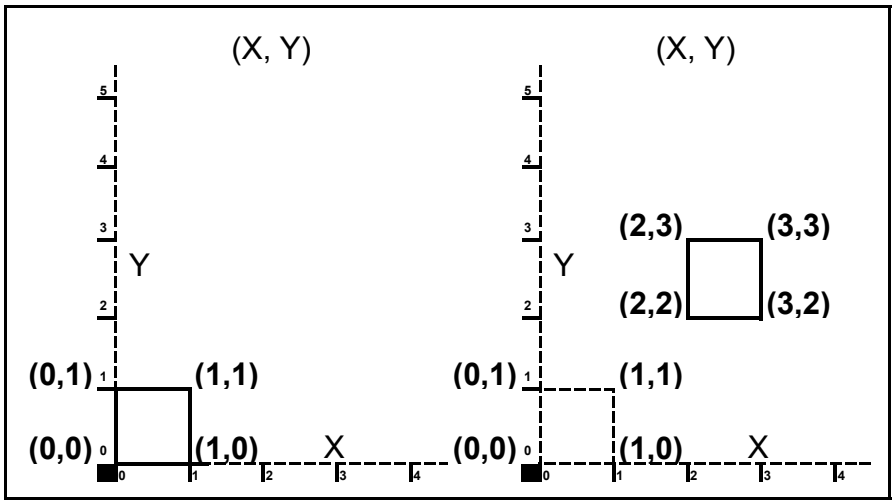


Figure 20. Moving a symbol for a period up and to the right

In Figure 20. the symbol for a period is moved (translated) plus two units in the X direction (to the right) and plus two units in the Y direction (up). By adding (2, 2) to the polygon defined by the vectors {(0,1), (1,1), (1,0), (0,0)} the polygon is moved to {(2,3), (3,3), (3,2), (2,2)}

Combining Operations

An important mathematical attribute of these transformations is that if several movements, scalings, and/or rotations are desired, in succession, on the same image or image set, the operations can be performed on just the equations, and the resulting concatenated (composite) equation (or function) can be applied to the outlines of the image. (This is called concatenating the functions or operations.) Creating a composite transformation equation and applying it once is much more efficient that applying each individual transformation equation in sequence to all of the outline vectors in an image or image set.

An example of applying all three functions at once (movement, scaling, and rotation) is Figure 26., Spiral Infinity. The Spiral Infinity figure, created in 1981, is essentially impossible to create without the mathematics behind the PostScript (and PDF) page description languages (PDLs) and is therefore an exemplar for PDLs.

All halftoning, color, and raster image management in the PDF, PostScript, and other PDLs have been omitted because of their complexity. (For PostScript, these topics are covered in the *Postscript Language Reference*, Third Edition, see Bibliography.) Some of the complexity can be transferred to a continuous tone or bi-tonal raster image that should last forever. Other parts of the complexity fall into the category of an ephemeral format that will not last for the foreseeable future.

The Third Dimension in the Equations

The third dimension, 'Z', is assumed to be zero for the above transformations. When the third dimension, 'Z', is not assumed to be zero, a series of these types of transformations on 3D objects results in a computer generated movie or even a complete 3D world, as was the cyberspace premise in the movie 'Matrix'. Beyond movies and cyberspace, these same equations are the equations used for the real buildings, bridges, and cityscapes that are designed, built, and documented with 3D as-built GPS (Global Positioning System) [http://Tycho.USNO.Navy.mil/gpsinfo.html] based (with solid state inertial navigation for interior imaging) optical and ground penetrating radar scans and stereo images, using CAD (Computer Aided Design) systems and GIS (Geographic Information Systems) systems. (With the integration of CAD and GIS, 64-bit computing and data values are required to maintain CAD precision across GIS dimensions. The 64-bit solution carries with it the need, which will require constant attention, to avoid the illusion of pseudo-accuracy.) With low cost storage and 64 bit data values, GIS systems can finally move away from a symbolic foundation to a 3D solid model foundation. This will allow the creation of a 3D solid model repository for the built (and natural) environment. (While a symbolic foundation is often adequate for the study of geography, the sharing of the GIS systems by all disciplines makes the transition to a 3D solid model foundation a change of paramount importance.) Movie studios and medical/legal/architectural animators employ digital archivists to manage records based on these equations for use in future movies and animations. Southern California now has several digital archivists (records managers) managing these high value digital assets. With the 'Z' dimension, these equations extend to3D simulations, automatic generalization in cartography, and data visualization using VRML (Virtual Reality Modeling Language) and the eXtensible 3D (X3D) specification, which is extending VRML97 ISO/IEC 14772-1:1997 (ISO Standard), using the eXtensible Markup Language (XML). Web3D Consortium [http://www.VRML.org].

These formulas provide an excellent practical application of mathematics, and would therefore facilitate the study of mathematics as well as the physical world through chemical, mechanical, geologic, and celestial animations. These animations can be controlled by a user with a joystick creating virtual reality exploration. With one more step, a virtual reality index to City documents can be created. Images from videos of a mockup of such a virtual reality index for land and building documents of the City of Los Angeles are available from the UCLA Urban Simulation Team website: [http://www.ust.ucla.edu/ustweb/ust.html].

In the virtual reality model of Los Angeles, users can use a joystick to fly over, and drive through, the City of Los Angeles to locate and gain access to simulated City records. By giving the simulated dirt a transparent setting, fly-unders would also be possible (in addition to the more traditional fly-overs), providing access indexing for the City's substructure. The joystick model also allows flying forward and backwards in time, with buildings popping out of the ground as a City grows. This was first described in 1956 by Arthur C. Clark, in his book, *The City and the Stars*, (page 46-58, Harcourt, Brace & World) in which he combined a calendar and a three dimensional model to catalog and display changes in the history of a city. Visitors could use a 4D projection of a simulacrum of the city to move a cursor through the city in any direction or through time. Any scale or perspective could be selected, and by moving smoothly, one could fly over, move through, and watch the growth and decline of different parts of the city over time. As a start on this model, the UCLA Urban Simulation Team has a model that starts in second century (AD) Rome that shows the progression of buildings on selected sites.

Another application for these simulations is downstream from Hoover Dam, where the random walk of the Colorado River around its alluvial cone (at a randomly peaked angular velocity) causes a periodic filling and drying (in the picohertz (1 thousand year) range) of the Salton Sea (Sink) [http://www.lc.usbr.gov/~saltnsea/ssrest.html] [http://ca.water.usgs.gov/gwatlas/basin/terminal.html] [http://edcwww.cr.usgs.gov/earthshots/slow/Imperial/Imperialtext] [http://ceres.ca.gov/watershed/geographic/colorado.html] Use of the graphics equations above would make this natural cycling more clear to the people seeking to restore the Salton Sea (Sink). The equations would show the concentric and evenly spaced iso-lines, with their common center at the mouth of the river's valley, that show an equipotent surface with respect to flowing water (and a cone with respect to the gravitational / centrifugal equipotent surface of the earth's geoid.) The graphic equations would also help those seeking to restore the Los Angeles River to see that the corresponding random walk of the Los Angeles River around its alluvial cone makes contention for possession of the United States Army Corps of Engineers built Los Angeles River Channel [http://www.lalc.k12.ca.us/target/units/river/tour/index.html] [http://ceres.ca.gov/wetlands/geo_info/so_cal/los_angeles_river.html] an illusory quest for a concrete, but fictitious, permanent river channel: a modern version of tilting at windmills (Trying to transform, and possibly put at risk, what is otherwise a really great place to train bus drivers, and what is also a really great place to film movies about futures we hope do not happen.). A simulation, using these equations, would show that any of the infinite number of channels available on the alluvial cone could be used to create one or more newly restored natural river channels of the Los Angeles River.

The Long Term from the 1930s

The astronomical star chart on Hoover dam points to another periodicity that is woven into our culture, but is relatively unknown in a specific sense. This behind-the-scenes periodicity brought us the dawning of the 'Age of Aquarius' (as the vernal equinox moves into the sign of Aquarius) celebrated in the musical *Hair*: a wished for age of harmony and understanding (something like the connections between the mathematics of typesetting and our newly computer based/constructed physical world). The dawning of a new astrological age is linked the fact that the signs of the zodiac are currently off by more than a month and still moving (including moving the vernal equinox into the constellation of Aquarius). The periodicity here is the precession of the earth's axis, which rotates once every 26 thousand years. In the more than two thousand years since the signs of the Zodiac were first identified, the earth's axis, and the signs of the Zodiac, have precessed more than a month. (A month is one twelfth of a year, and 2 thousand years is a little less than one-twelfth of 26 thousand years, the period of the precession of the signs of the Zodiac, a periodicity that is just over 1 picohertz.) [http://csep10.phys.utk.edu/astr161/lect/time/precession.html] [http://cse.ssl.berkeley.edu/lessons/indiv/beth/beth_intro.html] (See also Figure 21. Inverse Table of Periodicity) This table would have been of great use to the engineer who once suggested the use of disk drives that were 2 milliseconds faster in order to speed access to data (documents) that had last been reviewed more than a century ago. Comparing centuries and milliseconds can be a problem! Records managers must plan for all of the technological changes that might occur during the retention period of each document covered by a retention schedule.

Customary Units	Number	Equivalent Customary Units	Number of Common Units	Common Units	Number	Hertz Range	Power of 10
electron frequency (at C, the speed of light)	2.454	picometers	1,200,000,000,000,000,000	hertz	1.20	exahertz	18
fiber optic wavelength (carrier frequency)	1,500	nanometers (=1.5 um)	230,000,000,000,000	hertz	230.00	terahertz	12
microprocessor clock rate (cycle time)	1	billion clock cycles/sec.	1,000,000,000	hertz	1.00	gigahertz	9
Computer RAM (Random Access Memory)	50	nanoseconds	20,000,000	hertz	20.00	megahertz	6
magnetic disk access time (12,000 RPM)	5	milliseconds	200	hertz	200.00	hertz	0
jukebox access (picker) (1 to 5 seconds)	1	second	1	hertz	1.00	hertz	0
second (1 hertz = 1 cycle per second)	1	second	1	seconds	1.00	hertz	0
minute	60	seconds	60	seconds	16.67	millihertz	-3
hour	60	minutes	3,600	seconds	277.78	microhertz	-6
day	24	hours	86,400	seconds	11.57	microhertz	-6
week	7	days	604,800	seconds	1.65	microhertz	-6
month	1/12	year	2,629,800	seconds	380.26	nanohertz	-9
year	365.25	days	31,557,600	seconds	31.69	nanohertz	-9
year	1	year	1	years	31.69	nanohertz	-9
decade	1	ten years	10	years	3.17	nanohertz	-9
century	1	hundred years	100	years	316.90	picohertz	-12
millennium	1	thousand years	1,000	years	31.69	picohertz	-12
precession of the Zodiac	1	rotation	26,000	years	1.22	picohertz	-12
million years	1	million years	1,000,000	years	31.69	femtohertz	-15
billion years	1	billion years	1,000,000,000	years	31.69	attohertz	-18
period of the universe (postulated)	1	period	85,000,000,000	years	372.80	zeptohertz	-21

Figure 21. Inverse Table of Periodicity

Crossover from Permanence to Capacity: A Short Story

Now that bits have been superimposed on the wave that electrons are by University of Michigan professor Philip Bucksbaum (See "Quantum laser turns electron wave into memory" By R. Colin Johnson *EE Times* [www.EETimes.com] 08/31/00, 2:39 PM ET.), one can speculate on the information carrying capacity of an electron. The deBroglie wavelength of any particle, including electrons, is equal to Planck's constant / momentum. Momentum = mass * velocity (C assumed in this case). Wavelength = 6.626 * 10**-34 J * sec / 9*10**-31 kg * 3*10**8 m / sec = 2.454*10**-12 m. Frequency = C / wavelength = 3*10**8 m / sec / 2.454*10**-12 m = 1.222 * 10**18 Hertz = 1.222 ExaHertz. If 8 bits could be modulated per baud (a 56 kilobit/second modem does better than 16 bits per baud [sometimes, theoretically]), then one electron could carry 1 ExaByte. The electrons in a mole of carbon (12 grams) (6.022*10**23 atoms) (12 electrons per atom) for a total of 12 * 6.022*10**23 electrons = 7.2 * 10**24 electrons. 12 grams of carbon could then store 7 *10**42 Bytes or 7 million trillion YottaBytes.

The short story: The Big Bang / Big Crunch theory postulates that the Universe explodes and then collapse on itself in an 85 billion year cycle (A periodicity of about 375 zeptohertz.). Imaging that a civilization had lasted 1 billion years (or even 50 billion years) in the last cycle. The scientists were very accomplished, but they could not stop the coming Big Crunch / Big Bang which occurred about 15 billion years ago. The scientists were only able to superimpose the records of their civilization on the definition of what electrons would become after the big bang. To see if this was the case, all we would have to do is to read the data carried in the wave of any electron.

Carrying this full-circle, Richard P. Feynman, starting with a description of document management (books), linked molecular models (A person's ability to walk, talk, and write is based on molecular manipulation.) and the molecular machines of nanotechnology (defined by Eric Drexler starting about 1980) [http://www.Foresight.org] [http://www.nano.gov] in his 1959 talk: 'There's Plenty of Room at the Bottom'. (See Caltech's *Engineering and Science*, February 1960) [http://www.zyvex.com/nanotech/feynman.html] Norsam.com writes such records, in a simple, easy to read raster format (The Norsam Rosetta media format can be read optically, with a microscope.), with an ion milling machine using a 7 to 50 nm (nanometer) spot size (about 400 thousand to 3 million dpi).

Feynman's 1959 description of document reproduction proceeded as follows: ". . . there is a device on the market, they tell me, by which you can write the Lord's Prayer on the head of a pin. But that's nothing; that's the most primitive, halting step in the direction I intend to discuss. It is a staggeringly small world that is below. In the year 2000, when they look back at this age, they will wonder why it was not until the year 1960 that anybody began seriously to move in this direction. *Why cannot we write the entire 24 volumes of the Encyclopedia Britannica on the head of a pin?* Let's see what would be involved. The head of a pin is a sixteenth of an inch across. If you magnify it by 25,000 diameters, the area of the head of the pin is then equal to the area of all the pages of the Encyclopedia Britannica. Therefore, all it is necessary to do is to reduce in size all the writing in the Encyclopedia by 25,000 times. Is that possible? . . . one of the little dots on the fine half-tone reproductions in the Encyclopedia. . . . is still 80 angstroms [8 nanometers] in diameter---32 atoms across, in an ordinary metal. In other words, one of those dots still would contain in its area 1,000 atoms. . . . and there is no question that there is enough room on the head of a pin to put all of the Encyclopedia Britannica. . . . but let's consider all the books in the world. . . . let us say that there are some 24 million volumes of interest in the world. . . . Now, instead of writing everything, as I did before, on the *surface* of the head of a pin, I am going to use the interior of the material as well. . . . Suppose that . . . Each letter represents six or seven ``bits'' of information; . . . Suppose, to be conservative, that a bit of information is going to require a little cube of atoms 5 times 5 times 5---that is 125 atoms. . . . it turns out that all of the information that man has carefully accumulated in all the books in the world can be written in this form in a cube of material one two-hundredth of an inch wide--- which is the barest piece of dust that can be made out by the human eye. So there is *plenty* of room at the bottom! Don't tell me about microfilm!" (Feynman probably wanted to replace the microfilm in Vannevar Bush's Memex: 'As We May Think', *Atlantic Monthly*, July 1945.) [http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush.shtml] [http://www.feedmag.com/html/document/97.09bush/97.09bush_master.html] ('As We May Think' is often cited by Ted Nelson, inventor of hypertext, as the progenitor for hypertext and the Internet. [http://www.sfc.keio.ac.jp/~ted/XU/XuPageKeio.html])

Professional organizations, such as ARMA (Association of Records Managers and Administrators, International) [http://www.ARMA.com] and AIIM (Association of Information and Image Management, International), [http://www.AIIM.com] serve to connect the different groups involved in shepherding ideas (like the ideas above) from their inception, through development, first implementation by early adopters, and the long sifting and winnowing process that yields the few tried and true techniques that can be honed for adoption by all organizations worldwide. Very important to this process are the professional publications and meetings sponsored by these organizations. Professional publications provide an opportunity to give structure to ideas so that the ideas can be presented and commented on. Professional meetings provide ample opportunity for dialog: one-on-one, in small groups, and in larger groups taking part in interactive presentations. In large international conferences, people move from group to group, cross-pollinating the discussions and ideas, so that a gathering of thousands plays together as a single professional team, working to advance knowledge and its use throughout the world. The knowledge thus gained is then diffused throughout the world as participants return home from the gathering and link their organizations to the dialog.

Fascicle Access Table

(For a less technical presentation, please skip this section.)

The following describes how metadata is taken from the fascicles and used to construct an access table (usually (physically) residing entirely in semiconductor RAM (Random Access Memory) for speed) (also usually (logically) residing in a SQL (Structured Query Language) table) to provide access to the data (records and records metadata) in the fascicles when the fascicles are placed on-line on magnetic disk. The access database structure in Figure 22. is used to access data in fascicles after the fascicles have been copied to magnetic disk and a RAM based fascicle access SQL database table has been assembled from all of the metadata in all of the sentinel files in all of the fascicles.

The data and metadata (from the fascicles) in the assembled fascicle access database table is still physically stored in fascicles that have been moved to fast access media. It is a copy of the metadata that is loaded into a standard SQL database to provide fast access for searching and delivery of records.

Any other database or document management system can access the fasciculated data through a TF system fascicle access database by calling the TF system. These other databases or document management systems can also provide access to ephemeral data (that is not stored by a TF system in fascicles) by using complex and dynamic data structures.

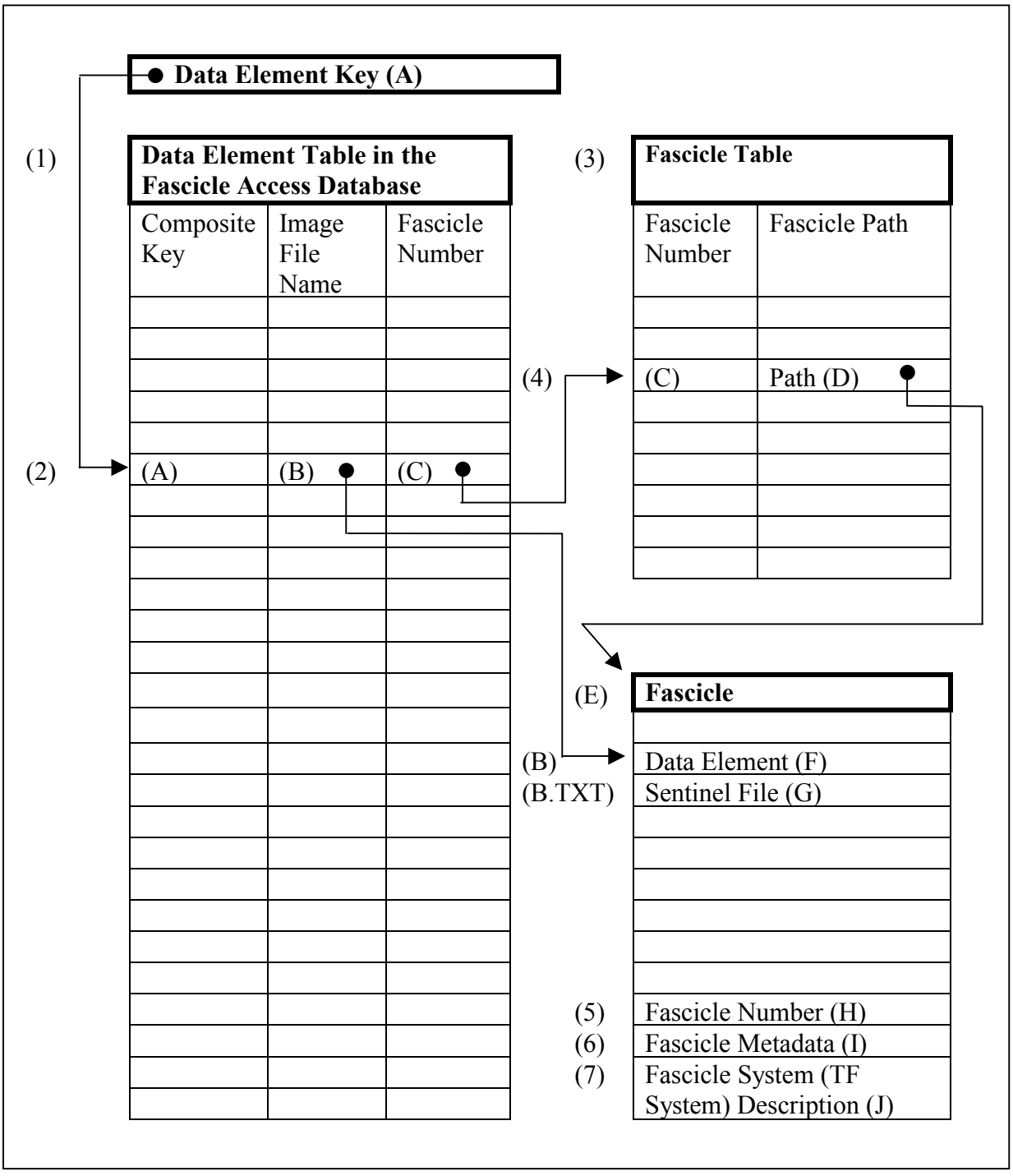


Figure 22. Fascicle Access Table (reconstructed database table for access to fasciculated data)

In Figure 22: The external key (A) for the data element (F) (e.g. a document stored (preserved) in a fascicle) is used to locate the row (2) of the data element table (1) which contains the data element file name (B) and the data element fascicle name (C) for the data element (F). The data element fascicle name (C) is used to locate the data element fascicle path (D) in the data element fascicle table (3). The data element file name (B) is then used to locate the data element file (F) in the data element fascicle (E). A parallelly named sentinel file (G) in the fascicle contains the database metadata information for the data element file. The parallelly named sentinel file (G) makes the data element fascicle self-defining. Because the fascicle is self-defining, it is not dependent on any application, operating system, or hardware. Each fascicle also carries fascicle defining metadata (H and I) and TF system defining metadata (J). (See also, Figure 15., Permanent Virtual Fascicle, which provides a graphical representation of the data and metadata records stored in a permanent virtual fascicle.)

Colophon

This paper is derived, revised, and expanded from a paper presented at ARMA 2000 (Association of Records Managers and Administrators, International). In some versions of this paper, these white papers are included in the PDF format version of the paper. For ARMA 2000, the Appendix include the following white papers as figures on the following pages:

Figure Numbers	White Paper Number	Title
23	22011	Projecting the Cost of Magnetic Disk Storage Over the Next 10 Years
24	22009	Digital Image Sizes
25	22016	Evolution of Intel Microprocessors: 1971 to 2003
26	22028	Spiral Infinity: Exemplar of PostScript Outline Fonts

Note to Readers

Updates and More Detailed Descriptions

When using the information in this article, please check the website www.ArchiveBuilders.com for updates. The version number of this article is just before the page number below. The website also has articles that provide more details on some of the terms and concepts in this article.

Comments

Please let us know how you like this paper, or if you had any questions. What would you like to see in the future? For more, and the most recent version of this article, please visit our web site at www.ArchiveBuilders.com.

Please send your comments via email to SteveGilheany@ArchiveBuilders.com. Tel: +1 310-937-7000. Fax: +1 310-937-7001. Also, please let us know where you saw this article.

Acknowledgements

This paper is a derivative work based on a paper presented at ARMA 2000 on October 24, 2000 (ARMA International, the Association of Records Managers and Administrators, www.ARMA.org).

Reprinted from Archive Planning, Volume 5, number 1, 2001, Archive Builders' analysis newsletter for document management.

See www.ArchiveBuilders.com.

Note to Editors

Paper 22025v186

We will continue to update these articles as we get comments. Please contact us for the most current version before you publish. Also, please request permission to publish the article. Permission will be given freely for most purposes.

Steve Gilheany
Archive Builders
1209 Manhattan Ave.
Manhattan Beach, CA 90266

Tel: +1 310-937-7000 Fax: +1 310-937-7001
SteveGilheany@ArchiveBuilders.com

Bio

Steve Gilheany, BA in Computer Science, MBA, MLS Specialization in Information Science, CDIA (Certified Document Imaging System Architect), AIIM Maser, and AIIM Laureate, of Information Technologies, CRM (Certified Records Manager, ARMA) has twenty years experience in document imaging and is a Sr. Systems Engineer at Archive Builders.

Author

Steve Gilheany is a Sr. Systems Engineer at Archive Builders. He has worked in digital document management and document imaging for twenty years.

His experience in the application of document management and document imaging in industry includes: aerospace, banking, manufacturing, natural resources, petroleum refining, transportation, energy, federal, state, and local government, civil engineering, utilities, entertainment, commercial records centers, archives, non-profit development, education, and administrative, engineering, production, legal, and medical records management. At the same time, he has worked in product management for hypertext, for windows based user interface systems, for computer displays, for engineering drawing, letter size, microform, and color scanning, and for xerographic, photographic, newspaper, engineering drawing, and color printing.

In addition, he has nine years of experience in data center operations and database and computer communications systems design, programming, testing, and software configuration management. He has an MLS Specialization in Information Science and an MBA with a concentration in Computer and Information Systems from UCLA, a California Adult Education teaching credential, and a BA in Computer Science from the University of Wisconsin at Madison. His industry certifications include: the CDIA (Certified Document Imaging System Architect) and the AIIM Master (MIT), and AIIM Laureate (LIT), of Information Technologies (from AIIM International, the Association of Information and Image Management, www.AIIM.org), and the CRM (Certified Records Manager) (from the ICRM, the Institute of Certified Records Managers, the official certifying body for ARMA International, the Association of Records Managers and Administrators, www.ARMA.org).

Contact:

SteveGilheany@ArchiveBuilders.com
Tel: +1 310-937-7000 Fax: +1 310-937-7001

For more information, courses, and papers:

http://www.ArchiveBuilders.com