DIGITALITZACIÓ I ARXIVAMENT A LLARG TERMINI DE MATERIAL FOTOGRÀFIC

Lukas Rosenthaler
Imaging & Media lab. Universitat de Basilea. Suïssa

INTRODUCCIÓ

L'arxivament a llarg termini dels materials fotogràfics sempre ha estat una tasca difícil. La majoria de processos fotogràfics no han estat dissenyats amb la longevitat en ment. Altres objectius, com ara la reproducció del color, la facilitat de revelatge i els costos, entre d'altres, han tingut més importància. Per tant, la fotografia és intrínsecament inestable i *tot* el material fotogràfic es deteriorarà amb el temps. Mentre que les fotografies en blanc i negre tenen una esperança de vida d'uns 100 anys (fins que els primers efectes del deteriorament es fan visibles), normalment la fotografia en color es deteriora més ràpidament. Fins i tot el material fotogràfic modern, que és més estable químicament, es deteriorarà amb el temps. El ritme de deteriorament pot variar i depèn principalment de les condicions d'emmagatzematge, però també dels materials emprats, els mètodes de revelatge, la manipulació, etc. Hi ha un tipus de material fotogràfic que té una esperança de vida molt més llarga: el microfilm. Tant el microfilm en B/N com en color[1] tenen una longevitat de més de 500 anys si es tracten correctament.

Com que molts objectes desats en arxius són elements únics irrecuperables en cas de danys o pèrdua total, qualsevol manipulació d'aquests elements sempre suposa grans riscos de danys o pèrdua total. Sovint, la manipulació també és poc pràctica i feixuga. Per aquest motiu, moltes fotografies de col·leccions i arxius s'estan digitalitzant. En general, l'objecte digitalitzat perd tota la «materialitat»: es tracta d'una representació immaterial d'una part de l'objecte original. Una fotografia digitalitzada només pot captar el contingut visual de la fotografia original, mentre que la majoria d'aspectes del material original, com ara les propietats físiques (gruix, característiques de la superfície, olor, etc.), normalment es perden. Sovint, les metadades que descriuen l'objecte original només són un substitut insuficient de la materialitat.

La majoria de projectes de digitalització s'han emprès o concebut sense l'arxivament a llarg termini en ment. Aquests projectes s'han iniciat per millorar l'accés als actius a través de bases de dades i d'Internet, però tan bon punt arriben les primeres dades digitals, sorgeix la qüestió de l'arxivament: el procés de digitalització és lent, costós i feixuc, per la qual cosa normalment no es tornarà a repetir en un futur immediat. A més, convindria evitar qualsevol manipulació addicional de les fotografies originals per tal de minimitzar el risc de danys. Com a conseqüència, qualsevol persona implicada en processos de digitalització s'enfronta, de cop i volta, al problema de l'arxivament a llarg termini de les dades digitals.

Per tal de complicar la situació, la majoria de fotografies que es creen avui en dia són «digitals de naixement», la qual cosa significa que tenen un origen digital. Però no només les fotografies, sinó també les pel·lícules, les animacions per ordinador i els vídeos que utilitzen tecnologia moderna produeixen «originals» de naturalesa digital. Aquest fet augmentarà la pressió per trobar solucions per a la preservació digital a llarg termini «de manera més aviat immediata».

La nostra experiència diària ens demostra que les dades digitals semblen ser molt volàtils i inestables. Qualsevol persona que treballi amb ordinadors de qualsevol magnitud ha tingut l'experiència negativa de perdre dades: un document de Word que de sobte és il·legible, un suport d'emmagatzematge extern al qual no es pot accedir, etc. Sembla que «arxivament a llarg termini» i «digital» siguin conceptes diametralment oposats. Tanmateix, les característiques inherents de les dades digitals comprenen la possibilitat d'un emmagatzematge il·limitat en el temps.

DADES DIGITALS I INFORMACIÓ

INTRODUCCIÓ

Les imatges digitals no existeixen: només hi ha ***dades digitals*** *que* ***representen*** *una imatge!* Aquest fet axiomàtic sempre s'ha tingut en compte a l'hora de parlar de les imatges digitals. Les dades digitals que representen una imatge sempre s'han de convertir amb algun mitjà tècnic (pantalla LCD, un projector, una impressora, etc.) en una distribució analògica de llum, que els éssers humans percebran com a imatge. La representació digital d'una imatge —és a dir, la «imatge digital» — es pot *crear*, *manipular* i *analitzar* mitjançant algorismes informàtics, però l'ordinador no pot *veure* una imatge.

En un sentit ampli, les dades digitals es poden definir com tot allò enregistrat utilitzant un codi basat en símbols en un suport. Un codi d'aquestes característiques utilitza un conjunt finit *S* de símbols:

$$S = \{s_1, s_2, .., s_n\}, n \geq 2$$

En són un bon exemple l'alfabet llatí, els jeroglífics egipcis, etc. Si *n* = 2 i, per tant, el codi tan sols utilitza dos símbols, s'anomenaria un «codi binari». Els codis binaris són els codis més simples de tots i es poden implementar fàcilment mitjançant maquinària informàtica.[2]

DIGITALITZACIÓ

La majoria de dades digitals són el resultat de la conversió d'un senyal físic analògic que pot variar en temps, espai i altres propietats. Per exemple, una càmera cinematogràfica en color enregistra la quantitat de llum que cau al sensor de llum com una funció de temps i espai dins dels límits del sensor i la freqüència (color) de la llum. El procés de digitalització converteix aquest senyal analògic en una sèrie de símbols permesos pel codi:

$$F(x, y, z, t, ...) \rightarrow s_i, s_j, s_k, ... \in S$$

Normalment, aquests símbols representen números relacionats amb l'amplitud física del senyal analògic. Per tal de comprendre aquest tipus de codi, cal saber el significat del codi. Per exemple, els números (53, 130, 211) en la posició 23.877 d'un fitxer poden representar la intensitat de la llum vermella, verda i blava, mentre que el número 0 fa referència a la intensitat de 0 i el número 255 a la intensitat màxima que el sensor de llum pot captar. La posició 23.877 indica que els valors s'han mesurat en una ubicació que es troba 3,75 mm a l'esquerra i 2,15 mm a la dreta de la cantonada superior esquerra del sensor. Per tal d'interpretar correctament aquests números, cal conèixer informació addicional, com ara les característiques dels filtres de color emprats per captar la llum a les àrees de vermell, verd i blau de l'espectre de llum o si els valors dels números enregistrats augmenten linealment amb la intensitat de la llum, etc. Normalment no es coneixen tots aquests paràmetres, si bé es poden fer suposicions raonables. Tanmateix, especialment quan hom treballa amb material històric únic, calen més coneixements per tal d'obtenir un resultat de digitalització valuós i útil.

En principi, el procés de digitalització inclou dues parts diferents:

1. Mostreig o tramatge

Les propietats físiques com ara el temps, l'espai, etc. s'han de mesurar en punts ben definits i diferenciats. En el domini del temps, aquest procés normalment s'anomena *mostreig*, mentre que la *freqüència de tramatge* defineix quant de temps transcorre entre dos mesuraments.

a. *Mostreig de l'espai*

A l'espai, normalment s'utilitza el terme *tramatge* i la distància entre dos punts on es mesura la propietat física dóna com a resultat la *resolució*. És evident que la qualitat del senyal mostrejat està directament correlacionada amb la freqüència de mostreig: com més elevada és la freqüència de mostreig, millor és la qualitat. La freqüència de mostreig o resolució mínima ve donada per la freqüència Nyquist, que es defineix com a dues vegades la freqüència més elevada que es produeix en el senyal. Si la freqüència de mostreig se situa per sota d'aquest límit, es poden produir artefactes irreversibles com ara distorsions o moaré. Un exemple molt conegut es pot observar en moltes pel·lícules de l'oest: les rodes d'una diligència de sobte semblen girar enrere (p. ex., durant una persecució) si la freqüència dels radis de la roda és més elevada que la freqüència de mostreig de la pel·lícula (que és de 24 imatges/segon).

b. *Mostreig de color*

En el cas de les imatges en color, també es mostreja la freqüència de la llum. Normalment només es prenen tres mostres (corresponents a vermell, verd i blau), però hi ha càmeres i escàners multiespectrals que utilitzen molts més punts de mostreig.[3] L'ús de només tres mostres és possible perquè la visió del color de l'ull humà també depèn de tres tipus de receptors de llum que són sensibles a la part vermella, verda i blava de la llum visible. Tanmateix, com que la corba de sensibilitat dels cons a l'ull humà normalment difereix de la resposta dels filtres de color emprats als sensors digitals, es poden produir artefactes relacionats amb el color. Aquest tipus de metameries poden aparèixer de dues maneres. En primer lloc, diferents colorants que tenen el mateix color per a l'ull humà poden ser diferents en la imatge digitalitzada. En segon lloc, els colorants que són diferents per a l'ull humà poden donar els mateixos números durant la digitalització i, per tant, seran indistingibles en la imatge digital. Com a conseqüència, ambdós efectes també depenen de la il·luminació. Això no obstant, es tracta d'efectes que normalment només són importants en la fotografia digital per a la digitalització de quadres, objectes, etc. La digitalització de fotografies en color és menys crítica, perquè la pròpia pel·lícula de color només es compon de tres capes, sensibles a les parts vermella, verda i blava de l'espectre de llum visible.

2. Quantificació

Cal convertir els mesuraments analògics realitzats en els punts de mostreig en un valor numèric en una representació simbòlica utilitzant un número finit de símbols. Així doncs, la precisió d'aquest tipus de conversió és *sempre limitada*. Amb molta més freqüència, la maquinària informàtica moderna utilitza un codi binari amb el número de símbols essent un múltiple de 8. El nombre de valors diferents que es poden representar ve donat pel nombre de dígits binaris (anomenats *bits*) elevats a la potència de 2 (n = 8 " 256 nivells, n = 12 " 4096 nivells, n = 16 " 65536 nivells). Hom diu que una quantificació que utilitza *n* bits té una *profunditat de bit* d'*n*.

Per tant, el resultat d'un procés de digitalització és una sèrie de símbols que *representen* el senyal analògic original. En teoria, la precisió d'aquest tipus de representació simbòlica no és limitada i es pot augmentar mitjançant l'increment de la freqüència de mostreig i les profunditats de bit. Tanmateix, la mida del codi digital resultant també augmentarà corresponentment. A la pràctica, les característiques mecàniques, electròniques i altres del procés de digitalització limitaran la precisió que se'n pugui obtenir.

CONVERGÈNCIA DE MITJANS

Les dades digitals poden representar de manera unificada la majoria de tipus de mitjans, ja sigui text, so, imatges, pel·lícules o altres tipus de mitjans nous com ara l'hipertext o les bases de dades relacionals, etc. Al final, tot plegat no és res més que un *flux de bits*.

ENREGISTRAMENT DE CODIS DIGITALS

A la pràctica, hi ha un altre obstacle: un codi digital només és un constructe idealista i lògic. En qualsevol cas, els símbols d'un codi s'han de representar *físicament* a través d'alguna propietat física que, en essència, sempre té una naturalesa *analògica*.[4] Alguns exemples d'aquestes propietats són el color de la tinta que constitueix la forma dels símbols escrits sobre un tros de paper visible, que pot consistir en el canvi de direcció d'un camp magnètic sobre la superfície d'un material ferromagnètic o en una marca gravada sobre la capa reflectora d'un disc compacte. Per tant, qualsevol procés d'enregistrament que converteix un codi digital en «permanent» deixa marques físiques *analògiques* en el suport d'enregistrament. Aquestes marques *representen* els símbols del codi digital.

Per tal de recuperar les dades, cal identificar aquestes marques analògiques i assignar els símbols digitals apropiats a les marques. Això es fa llegint el senyal analògic i aplicant un *procés de decisió*. En el cas d'un codi binari, aquestes decisions es poden prendre simplement llindaritzant el senyal físic, tot i que sovint cal processar la senyal i altres procediments més complexos com ara el reconeixement de la forma, etc. Així doncs, no hi ha res millor que l'«enregistrament digital» de dades. Tots els enregistraments «digitals» són analògics per naturalesa i, durant el procés de lectura, el codi digital es genera «al vol».[5] Si bé els *codis binaris* que requereixen només dos símbols o estats diferents són la codificació més habitual, hi ha molts exemples de codificacions que fan servir més símbols, la més important de les quals és el text escrit.

Tanmateix, en el món dels ordinadors, ja fa moltes dècades que hi predomina el codi binari que utilitza només dos símbols «0» i «1», «vertader» i «fals» o, quan la informació es representa a l'ordinador, «+3,3V» i «-3,3V». Es tracta d'un codi que és fàcil d'implementar perquè el procés de decisió binària és el procés més simple possible i, per tant, molt sòlid. No obstant això, a la pràctica es produeixen errors del procés de decisió que cal detectar i corregir mitjançant mecanismes especials. Sobretot, pel que fa a l'enregistrament permanent de dades digitals, els mètodes matemàtics anomenats Codis Correctors d'Errors (ECC, segons les seves sigles en anglès) hi juguen un paper important (vegeu més endavant).

LECTURA I DESCODIFICACIÓ DE CODIS DIGITALS

Com que les dades digitals s'enregistren utilitzant marques analògiques en un suport físic, el primer pas per llegir i descodificar dades digitals normalment requereix la identificació del mètode d'enregistrament. Si bé això pot semblar una tasca trivial ja que els mètodes d'enregistrament bàsics (òptic, magnètic) es poden distingir fàcilment, la clau està en els detalls. Mentre que la caixa i el factor de forma de les cintes modernes de la mateixa família (p. ex., DAT, LTO, DLT...) són iguals, els estàndards d'enregistrament han canviat significativament al llarg de diverses generacions. La taula següent mostra alguns dels paràmetres més importants de la família LTO:

	Any	Capacitat	Pistes paral·leles	Pistes escrites/passada
LTO-1	2000	100 GB	384	8
LTO-2	2003	200 GB	512	8
LTO-3	2005	400 GB	704	16
LTO-4	2007	800 GB	896	16
LTO-5	2010	1.400 GB	1280	

Evidentment, les generacions més recents tenen una densitat de dades molt més elevada, cosa que es tradueix en pistes magnètiques més estretes. Per aquest motiu, normalment les unitats de cinta magnètica només poden llegir i escriure cintes de fins a una generació anterior, i llegir cintes de fins a dues generacions anteriors. Les cintes més antigues no es poden llegir ni escriure.

D'això, se'n deriva que el primer pas a l'hora de llegir i descodificar dades digitals sigui la identificació del mètode d'enregistrament específic de les dades en un suport de dades i la obtenció d'una màquina que sigui capaç de llegir les marques físiques específiques. Com que habitualment la vida útil d'una tecnologia específica d'enregistrament és de pocs anys, el primer pas pot resultar difícil. Avui en dia, trobar una unitat de cinta que llegeixi una cinta LTO-1 o una unitat per llegir un disc Iomega ZIP antic pot resultar extremadament complicat.

Descodificar un codi digital és el procés d'extreure la *informació significativa* d'una seqüència de símbols. Per tal de llegir i comprendre un text escrit, no només cal identificar els símbols —els caràcters o signes—, sinó que també cal conèixer el llenguatge —la sintaxi i la semàntica— en el qual ha estat escrit el text. L'escriptura jeroglífica de la famosa pedra Rosetta —que conté el mateix text en dos idiomes (egipci i grec)—, finalment només es va poder llegir i comprendre gràcies a la utilització de tres escriptures (jeroglífica, demòtica i grega). Cada codi digital té normes sintàctiques i semàntiques explícites o implícites que cal saber per tal d'interpretar el codi correctament.

El significat d'un símbol sovint depèn de la *posició* a dins de la seqüència de símbols. Freqüentment, els símbols es combinen en grups per formar símbols nous (habitualment coneguts com a «paraules»), que al seu torn es combinen per formar unitats superiors («frases»). Els ordinadors que utilitzen els codis binaris normalment empren paraules d'una mida de 8, 16, 32 i 64 bits. Les normes que defineixen la sintaxi i la semàntica d'un codi digital s'acostumen a conèixer com a *format de fitxer*. Per tant, el coneixement del format de fitxer és fonamental per llegir dades digitals. Els formats de fitxer *oberts* donen un coneixement explícit i ofereixen una descripció detallada i completa de la sintaxi i de la semàntica. Aquesta descripció pot venir donada per una descripció textual en anglès bàsic (o qualsevol altre idioma comú) o per la font (amb sort comentada) d'un programa informàtic que llegeix les dades. Aquest tipus de descripció no està disponible per als formats de fitxer *de propietat*. Com a molt, un programa binari o biblioteca és capaç d'interpretar el codi digital dels formats de fitxer de propietat.

Per tant, llegir i comprendre un codi digital requereix tres passos diferents:

1. Cal identificar el mètode d'enregistrament.
2. Cal identificar el format de fitxer.
3. Cal aplicar les normes sintàctiques i semàntiques adequades del format de fitxer per tal d'interpretar el codi digital.

Si un fitxer de dades s'identifica com un fitxer d'imatge TIFF, però es desconeix l'especificació del format TIFF, la imatge representada per les dades no es podrà extreure. Així doncs, si no es pot identificar o es desconeix el sistema semàntic d'un codi digital, no es podrà extreure la informació continguda en les dades digitals.

Això fixa els prerequisits següents per a la recuperació de dades digitals:

1. Cal conèixer la propietat física emprada per crear les marques. Per als tipus de mitjans actuals, normalment es coneix la propietat física que s'ha fet servir per crear les marques. *Sabem* que un disquet té marques magnètiques i que un CD-R té marques detectables òpticament. Tanmateix, els futurs arqueòlegs digitals podrien tenir problemes per determinar quina propietat física s'ha utilitzat per crear les marques, especialment pel que fa a les noves tecnologies d'enregistrament emergents.
2. Les marques físiques del mitjà s'han de poder detectar i convertir en símbols. Si això no és possible, a causa de possibles danys i del pas del temps, caldrà considerar el mitjà com a «destruït» i no llegible.
3. Cal identificar i conèixer el sistema sintàctic i semàntic (format de fitxer).

Si alguna d'aquestes tasques no es pot dur a terme, les dades digitals ja no es podran llegir i la informació enregistrada es perdrà.

REDUNDÀNCIA, COMPRESSIÓ SENSE PÈRDUES I CODIS CORRECTORS D'ERRORS

L'obra fonamental de Claude Shannon sobre la teoria de la informació «A mathematical theory of communication»[6] conté dues afirmacions importants, que reproduïm tot seguit de manera simplificada:

1. Qualsevol codi en què la probabilitat d'ocurrència no sigui la mateixa per a tots els símbols conté redundància. En aquest cas, és possible trobar un nou codi que la minimitzi.
2. Si una via de comunicació introdueix errors als símbols transmesos, es pot trobar un nou codi que permeti corregir aquests errors.

La primera afirmació fa referència a la possibilitat de la *compressió sense pèrdues*, mentre que la segona afirmació tracta de les possibilitats dels *codis correctors d'errors*. La teoria de Shannon mostra que hi ha un equilibri entre eficiència (compressió sense pèrdues), d'una banda, i correcció dels errors (redundància), de l'altra. Molts codis —com ara el llenguatge escrit— contenen moltes redundàncies i d'aquí que siguin força tolerants amb els errors. Tanmateix, en els sistemes informàtics digitals es requereix una elevada eficiència, motiu pel qual sovint s'empren tècniques de compressió.

Resulta interessant saber que la majoria de dispositius informàtics d'emmagatzematge utilitzen internament alguna redundància per tal d'aconseguir capacitat de correcció d'errors. De fet, les unitats de disc, les unitats òptiques i les unitats de cinta magnètica no funcionarien sense la correcció interna d'errors. La conversió del senyal físic analògic en diferents símbols té una probabilitat força significativa de ser incorrecta. Fins i tot amb la correcció d'errors, un disc dur modern corrent té una probabilitat no zero d'errors: estadísticament, 1 de cada 10^{14} bits són erronis.[7] Aquests errors s'anomenen *errors de lectura no recuperables*. Estadísticament, això suposa un fitxer malmès cada 25 cops que es copia un disc dur de 400 GB. La tecnologia de CD-ROM i CD-R no funcionaria sense la correcció d'errors. De fet, un CD-R tindria una capacitat bruta de més d'1 GB, però un 33% de la capacitat bruta d'un CD-ROM o CD-R s'utilitza per afegir redundància per a la correcció d'errors.

SUMES DE VERIFICACIÓ

Les sumes de verificació són l'equivalent digital de les empremtes humanes. Per a una determinada seqüència de símbols digitals (p. ex., un flux de bits o un fitxer de dades), l'algorisme de suma de verificació ideal calcula una seqüència nova *única* de símbols que normalment és molt més curta que la seqüència original. Això no obstant, aquest tipus d'algorismes ideals no existeixen. A la pràctica, la probabilitat que dos fitxers diferents tinguin la mateixa suma de verificació és insignificant. Hi ha molts algorismes de suma de verificació disponibles que produeixen sumes de verificació de diferent longitud. Els tipus de suma de verificació més habituals s'indiquen en la taula següent, incloent-hi les sumes de verificació resultants del text codificat ASCII «*Ser o no ser: aquesta és la qüestió*»:

Algorisme de suma de verificació	*Suma de verificació*
MD5	eaf606c87569b2f97e230e792049833e
SHA-1	71d7726d2db38295ddea57c5dccd3be388fc0ab5
RIPEMD-160	c5fd0db228230b0f0813ace8376150527bf24588
WHIRLPOOL	ca4685900bbc481f3d8a1c71b17512aa4c62b4fb da19df8969da9052a2c54ea2e7073ba524183890 8cbc865a0d4ac6cc74e284f12f90f4dc90d9864d9 4fdd900

Per tant, les sumes de verificació són un instrument ideal per verificar si dos fitxers digitals són idèntics: si ambdós fitxers tenen la mateixa suma de verificació, han de ser idèntics.

Les sumes de verificació són un mitjà ideal per verificar la integritat d'un fitxer digital, ja que la suma de verificació serà completament diferent fins i tot si només se n'ha modificat un bit. Si la suma de verificació d'un fitxer s'ha emmagatzemat per separat del fitxer, es pot tornar a calcular i comparar amb la suma de verificació original en qualsevol moment. Si ambdues sumes de verificació són idèntiques, això significa que el fitxer no s'ha modificat; si no ho són, vol dir que el fitxer està malmès. Per tant, les sumes de verificació també s'empren per garantir que un procés de còpia de fitxer s'hagi realitzat correctament. Si els fitxers «original» i «copiat» tenen la mateixa suma de verificació, el procés de còpia s'ha fet sense errors.

PROPIETATS DE LES DADES DIGITALS

De la secció anterior «Dades digitals i informació», se'n poden deduir les propietats següents de les dades digitals:

Pèrdua de la noció d'un «original»

Per a la informació digital, la noció d'un *original* no té sentit. Les dades digitals es poden copiar sense pèrdues reproduint la mateixa seqüència de símbols de la seqüència «original». Les dues còpies no es podran distingir l'una de l'altra i, per tant, tampoc es podrà determinar quina és l'«original». Tanmateix, com que la representació física d'un codi digital sempre té una naturalesa analògica que pot provocar errors, el procés de còpia digital només es completa quan s'ha *verificat* que ambdues còpies son *idèntiques*, ja sigui mitjançant una comparació a escala dels símbols (o, en el cas de dades binàries, a escala dels bits) o utilitzant sumes de verificació.

Per tant, les dades digitals es poden copiar sense límits i no hi haurà pèrdua generativa.

Independència del mitjà d'enregistrament

Les dades digitals són independents del mitjà on s'enregistren sempre que els símbols es puguin desxifrar. Per exemple, un fitxer informàtic binari que representa una imatge amb el format JPEG es podria gravar en una pedra (si bé no seria massa pràctic per treballar-hi, sí que seria viable). Així doncs, les dades digitals es poden copiar d'un mitjà a un altre sense patir pèrdues.

Anul·lació de l'espai

Les dades digitals es poden transportar a través de l'espai a la velocitat de la llum sense necessitat de moure àtoms o matèria. Aquesta propietat permet *telecopiar* les dades digitals sense patir pèrdues a la velocitat de la llum.

Fonts de pèrdua de dades

Per tal de comprendre els problemes de l'arxivament a llarg termini de les dades digitals, cal avaluar les possibles fonts de pèrdua de dades:

1. **Error en llegir els bits**

   Si els símbols del codi ja no es poden identificar, es perden les dades enregistrades. Hi ha diverses causes de la incapacitat per identificar els símbols enregistrats (per «llegir els bits»):

a. ***Dany físic del mitjà***
Les marques físiques de l'enregistrament ja no es poden llegir com a conseqüència d'un mitjà malmès (envelliment, tractament brusc del mitjà, defectes, etc.).

b. ***Ja no hi ha dispositius de lectura disponibles***
El mitjà no es pot llegir perquè el dispositiu necessari ja no està disponible. Per exemple, les unitats de cinta per llegir cintes magnètiques DLT I ja no estan disponibles comercialment.

c. ***Error humà***
Les dades enregistrades en el mitjà s'han esborrat per un error humà, el mitjà s'ha etiquetat malament i mai no s'hi havien arribat a escriure les dades que suposadament sí que hi eren...

Normalment, la vida útil física d'un mitjà d'enregistrament és més llarga que la vida útil d'un sistema d'enregistrament específic, motiu pel qual l'envelliment del mitjà d'enregistrament, sempre que s'emmagatzemi adequadament, no acostuma a ser el factor limitador. Per exemple, segons John W.C. Van Bogart, els mitjans magnètics (cintes) tenen una vida útil d'uns 30+ anys.[8] Això no obstant, la vida útil del sistema d'enregistrament en conjunt depèn de la disponibilitat del servei tècnic per al maquinari i programari necessaris per part del(s) fabricant(s). Aquesta vida útil pot ser força curta (3-10 anys) i acostuma a ser el factor limitador de la vida útil de les dades digitals enregistrades en un mitjà.

1. Error en llegir el format de fitxer

Hi ha diversos motius pels quals un format de fitxer no es pot llegir:

d. ***Identificació del format de fitxer***
Si no hi ha metadades per indicar quin format de fitxer s'ha emprat per escriure les dades, identificar-lo pot arribar a resultar molt difícil. En l'actualitat es fan servir diversos milers de formats de fitxer.[9] Alguns dels més habituals es poden identificar fàcilment mitjançant l'anomenat «número màgic», compost pels 4-8 primers bytes d'un fitxer, però per a tota la resta de formats la identificació pot arribar a ser molt difícil.

e. ***S'ha perdut l'especificació del format de fitxer***
Si es pot identificar el format de fitxer, l'obstacle següent és trobar el programari que sigui capaç de llegir-lo i interpretar-lo. O bé encara hi ha disponible un programari que sigui capaç de llegir les dades o bé cal crear un programari nou per llegir el format. Tot sovint, la primera opció resulta difícil d'aconseguir, mentre que la segona exigeix que el programador tingui a la seva disposició l'especificació completa del format. Per tant, els formats de propietat, l'especificació del format dels quals no està disponible, no acostumen a ser adequats per a l'arxivament a llarg termini.

Taula de formats habituals de fitxers d'imatges amb número màgic

Tipus de fitxer	Extensió típica	Dígits hexadecimals xx = variable	Dígits ASCII
Format GIF	.gif	47 49 46 38	GIF8
Format FITS	.fits	53 49 4d 50 4c 45	SIMPLE
Format de mapa de bits	.bmp	42 4d	BM

Sistema de nucli gràfic	.gks	47 4b 53 4d	GKSM
Format rgb IRIS	.rgb	01 da	..
Format ITC (CMU WM)	.itc	f1 00 40 bb	....
Format d'intercanvi de fitxers JPEG	.jpg	ff d8 ff e0	....
NIFF (Navy TIFF)	.nif	49 49 4e 31	IIN1
Format PM	.pm	56 49 45 57	VIEW
Format PNG	.png	89 50 4e 47	.PNG
Format PostScript	.[e]ps	25 21	%!
Sun Rasterfile	.ras	59 a6 6a 95	Y.j.
Format Targa	.tga	xx xx xx	...
Format TIFF (Motorola big-endian)	.tif	4d 4d 00 2a	MM.*
Format TIFF (Intel little-endian)	.tif	49 49 2a 00	II*.
Format de mapa de bits X11	.xbm	xx xx	
Estructura de fitxers Gimp XCF	.xcf	67 69 6d 70 20 78 63 66 20 76	gimp xcf
Format Xfig	.fig	23 46 49 47	#FIG
Format XPM	.xpm	2f 2a 20 58 50 4d 20 2a 2f	/* XPM */

Tal com s'ha indicat abans, els formats de fitxer defineixen el *significat dels bits*. És a dir, el format de fitxer defineix la semàntica dels bits. Normalment, el format de fitxer es pot descriure mitjançant un text bàsic en algun idioma humà. Per exemple, l'especificació bàsica del format de fitxer d'imatge TIFF és un document de 121 pàgines en un anglès bàsic que descriu detalladament l'estructura i la semàntica del format TIFF. Com que un format de fitxer es pot emprar per a diversos subtipus d'objectes digitals (p. ex., el format TIFF es pot utilitzar per a molts tipus diferents d'imatges digitals), molts formats de fitxer inclouen les anomenades *metadades tècniques* que descriuen aspectes tècniques com ara la resolució, la profunditat dels bits, la interpretació colorimètrica... d'una imatge específica. Aquestes metadades tècniques són necessàries per descodificar i renderitzar adequadament el contingut d'un fitxer digital, motiu pel qual són parts integrals del format de fitxer. Les metadades tècniques no s'han de confondre amb les *metadades descriptives* que fan referència al contingut que representa el fitxer digital. Molts formats de fitxer permeten afegir, com a mínim, algunes metadades descriptives al contingut d'un fitxer.

1. Pèrdua de les metadades descriptives

La majoria de fitxers de dades digitals no tenen sentit si no van acompanyats de metadades descriptives. Una col·lecció de 10.000 CD-R desconeguts etiquetats com a «00001» a «10000», amb cada CD-R amb 50 fitxers anomenats «000.dat» a «049.dat», gairebé no té cap valor si no hi ha més

informació disponible. Les metadades poden ser tan simples com un nom de fitxer significatiu que pugui llegir un ésser humà i una etiqueta eloqüent dels suports de dades, o pot ser un complex esquema de metadades basat en XML. El fet que hi hagi algun tipus de metadades disponible té la seva importància. La millor pràctica consisteix a incloure algunes metadades en el fitxer de dades. Molts formats de dades (p. ex., fitxers d'imatges TIFF, etc.) permeten afegir metadades a la *capçalera del fitxer*, automàticament extraïbles a través d'un processament eficient.

Com a conseqüència, hi ha dos problemes bàsics en relació amb la longevitat de les dades digitals:

1. L'envelliment o el dany del mitjà on estan enregistrades les dades digitals.

2. L'obsolescència del maquinari, del programari i dels formats de fitxer. El desenvolupament accelerat de la tecnologia informàtica provoca que la vida útil dels sistemes (maquinari i programari) sovint sigui inferior a 5 anys. Si els formats de dades es trien amb cura, poden arribar a tenir una vida útil de 10 anys o més.

MÈTODES D'ARXIVAMENT A LLARG TERMINI DE DADES DIGITALS

Kenneth Thibodeau, director del «Programa d'Arxius de Registres Electrònics» de la NARA (*National Archives and Records Administration* o Arxius Nacionals i Administració de Documents dels EUA), identifica més d'una dotzena de mètodes diferents per a l'arxivament a llarg termini de dades digitals.[10] Es poden agrupar en cinc mètodes bàsics:

1. ***Res***
   No fer res, els futurs arqueòlegs digitals faran la feina per tu.

2. ***Museu informàtic***
   Arxivar tot el sistema (mitjà, perifèrics i ordinador, incloent-hi tot el programari) en condicions de treball.

3. ***Emulació***
   L'emulació i la virtualització de sistemes obsolets en els ordinadors actuals permet executar programari obsolet en ordinadors moderns.

4. ***Migració***
   Copiar (i convertir a un format de dades actual, si cal) les dades a un mitjà nou, si el mitjà anterior comença a envellir o si el format s'està convertint en obsolet.

5. ***Mitjà permanent***
   Enregistrar les dades digitals en un mitjà amb una longevitat molt elevada, ja sigui emprant un format autodescriptiu o preservant la informació sintàctica i semàntica (p. ex., com a metadades).

Tots aquests mètodes tenen avantatges i inconvenients, però sembla que el primer mètode —tenir confiança en la capacitat dels futurs arqueòlegs digitals— encara continua essent el prevalent, tot i que presenta inconvenients que són obvis.

No fer res

Es tracta d'un mètode encara estès per a la preservació de les dades digitals. El motiu és que tots els altres mètodes requereixen, com a mínim, una despesa important i coneixements especials. Si algun d'ambdós components o tots dos no estan disponibles, no fer res —simplement conservar els suports de dades i esperar que algú sigui capaç de llegir i interpretar les dades en el futur— potser sigui l'única alternativa a destruir

activament les dades. Tot i no ser la millor solució, conservar un suport de dades en un entorn adequat, com a mínim, potser podrà arribar a alentir el deteriorament del material del suport de dades i de les marques físiques que hi té escrites, i deixarà oberta la possibilitat perquè en un futur se'n puguin recuperar les dades.

Museu informàtic: arxivar el mitjà, el maquinari i el programari

L'opció de preservar no tan sols el mitjà, sinó també el maquinari i el programari, sembla ser una solució aparent al problema de l'arxivament a llarg termini de dades digitals. Tanmateix, mantenir maquinària informàtica complexa en condicions de treball és molt difícil. No només envellirà el mitjà, ja que la resta de components dels ordinadors i els seus perifèrics tampoc són estables en el temps. Els circuits integrats, les plaques de circuit, les unions soldades, etc. envelliran i en algun moment o altre deixaran de funcionar. A més, a mesura que l'equip es faci vell, cada cop serà més difícil trobar peces de recanvi, manuals tècnics i de reparació, i tècnics que encara siguin capaços de reparar tecnologia antiga. Per tant, si bé la maquinària informàtica conservada és important per a l'arqueologia digital, no és un mètode generalment aconsellable per aconseguir la longevitat de les dades digitals. No obstant això, en els casos en què es descobreixin mitjans d'enregistrament antics en un arxiu o propietat, tenir accés a maquinària informàtica antiga per tal de transferir les dades d'un mitjà d'enregistrament antic a un de nou pot arribar a suposar una «salvació».

Emulació

L'emulació de programari crea un entorn de programari que permet executar els programes informàtics en una plataforma diferent (arquitectura d'ordinador i/o sistema operatiu) a l'original per a la qual s'havien escrit. Per tant, escriure un emulador per a un sistema informàtic antic i obsolet permet executar els programes antics en un ordinador nou. Amb l'ajuda de l'emulador, els formats de fitxer emprats en el sistema antic encara es poden llegir en un sistema nou. L'emulació planteja els problemes següents en relació amb l'arxivament a llarg termini de les dades digitals:

- Com que l'emulació normalment no pot emular els dispositius perifèrics —com ara les unitats de disquets, les unitats de cinta magnètica, etc.—, els fitxers de dades i els programes s'han de migrar a mitjans moderns perquè el sistema amfitrió de l'emulació els pugui llegir.

- Caldrà conservar el programari d'emulació en si. Això és pot aconseguir mitjançant la migració a un nou maquinari i programari (bàsicament, reescrivint-lo per als ordinadors de nova generació) o mitjançant l'ús d'un concepte imbricat d'emulacions jeràrquiques (p. ex., Wordstar[11] executa una emulació de CP/M que s'executa en una emulació de l'OS-9 que, a la vegada, s'executa en un PC amb Window XP). Raymond Lorie va proposar un «Ordinador Virtual Universal» (UVC, segons les seves sigles en anglès) per tal de facilitar la migració d'emuladors.[12] La idea bàsica que s'amaga darrera de l'UVC és la definició d'una màquina virtual simple que es pugui implementar en el maquinari actual i futur. A més dels fitxers de dades digitals, el programari de renderització (és a dir, el programari que fa que la informació d'un fitxer de dades digital puguin utilitzar-lo els humans) s'ha d'escriure per a l'UVC i cal conservar-lo juntament amb les dades. Com que l'especificació de l'UVC no hauria de canviar, en el futur, el programari de renderització s'executarà en un UVC i es podrà utilitzar tant perquè els humans puguin emprar les dades com per transformar-les en un format de dades modern.

Com a conseqüència dels problemes que planteja l'emulació en general, podria ser una solució només per a casos en què sigui essencial conservar els propis programes informàtics. És el cas, per exemple, dels jocs o d'aquells casos en què el disseny d'un programa és essencial i cal conservar-lo. L'emulació també es podria emprar per llegir formats de dades de propietat no documentats que requereixin determinats programes de propietat que ja no existeixen en els ordinadors moderns. Tanmateix, com que l'emulació només permet executar un programa antic *dins* d'un entorn simulat, normalment no hi ha cap manera de traslladar la informació de l'emulació a la màquina moderna. És a dir, que generalment no és possible transferir les dades de l'emulació a un format útil a l'ordinador modern.

Migració

La migració utilitza la propietat de les dades digitals que es poden copiar sense patir pèrdues. Hi ha dos nivells de migració:

1. Migració del flux de bits

En aquest cas, només s'intercanvia el suport de dades mitjançant la còpia de les dades digitals d'un mitjà d'emmagatzematge d'una generació a la següent. Si el nou mitjà té una capacitat d'emmagatzematge diferent, caldrà un cert reencapsulatge dels fitxers de dades. En general, el procés de còpia només es completa si la comparació dels fitxers de dades a escala dels bits no detecta cap error.

2. Migració del format

En aquest cas, no només s'intercanvien els suports de dades, sinó que també canvia el format de les dades digitals (p. ex., de TIFF descomprimit a JPEG2000 descomprimit). Un procés de migració d'aquest tipus és bastant difícil:

a. Cal garantir que no es produeixi cap pèrdua de dades durant la conversió del format (p. ex., compressió amb pèrdues).

b. Generar la prova per saber que el procés de còpia ha reeixit és més difícil. El fitxer amb el nou format s'ha de tornar a convertir al format antic i després s'ha de comparar amb el fitxer «original». Aquesta comparació s'ha de fer a una escala lògica (p. ex., comparant valors de píxel en el cas d'imatges) i no a escala dels bits dels fitxers resultants. El motiu és que l'estructura dels fitxers pot ser diferent, fins i tot encara que representin exactament el mateix contingut (p. ex., dos fitxers TIFF poden ser diferents a escala dels bits, i tot i així representar imatges idèntiques amb contingut informatiu idèntic). Això és atribuïble al fet que sovint hi ha variants igualment vàlides de formats de fitxer comuns.

Amb els mitjans actuals, cal fer una migració del flux de bits aproximadament cada cinc anys. Les conversions de format només seran necessàries si un format esdevé obsolet; és a dir, si el nou programari deixa d'admetre el format. Tanmateix, una conversió de format pot tenir sentit si hi ha un nou format que presenti avantatges importants.

Mitjà permanent

Les dades digitals es podrien enregistrar en un mitjà permanent que tingui una longevitat intrínseca elevada. No obstant això, les dades s'han d'enregistrar d'una manera que sigui autoexplicativa i independent del format. A més, tornar a llegir les dades no ha de dependre d'un maquinari especialitzat, que a curt termini esdevindrà obsolet.

Segons ha desenvolupat l'Imaging & Media Lab,[13,14,15] una manera d'aconseguir aquest objectiu és mitjançant la codificació visual de les dades digitals sobre microfilm. Els bits es poden enregistrar com a configuració de bits formant una mena de codi de barres bidimensional. A més, la informació basada en text i les imatges analògiques es poden enregistrar en el mateix suport.

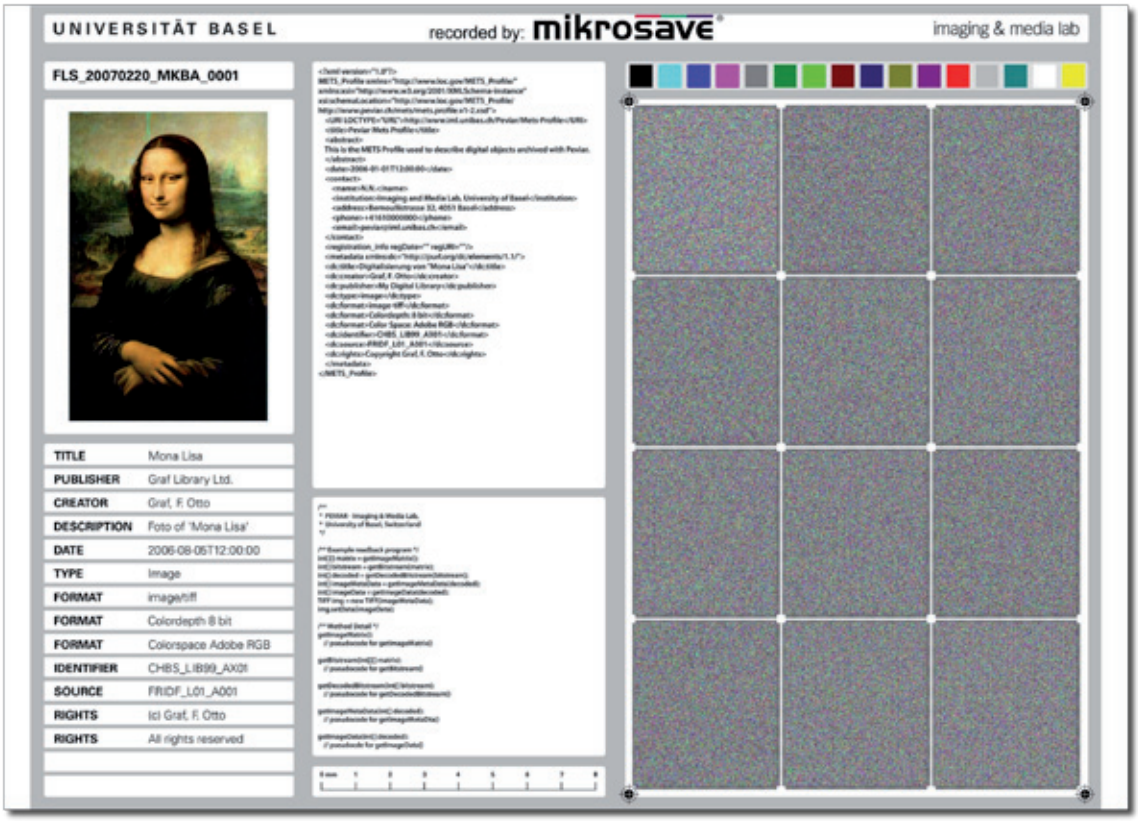

Figura 1: microfilm que té enregistrades una imatge analògica, informació basada en text (esquerra) i dades digitals binàries (dreta).

Un enregistrament digital basat en un microfilm d'aquestes característiques es pot tornar a llegir sense fer servir cap equip especial, simplement amb qualsevol càmera digital o escàner que tingui prou resolució per fer-ho. A més de les imatges analògiques que ajudaran a identificar l'objecte digital, la informació basada en un text pot contenir les instruccions sobre com descodificar la configuració de bits, així com informació sobre el format de fitxer. Un mitjà d'emmagatzematge d'aquestes característiques és del tot independent de qualsevol tecnologia específica, motiu pel qual no esdevindrà il·legible com a conseqüència de l'obsolescència tècnica. El microfilm té una esperança de vida de més de 500 anys.

EL MODEL DE REFERÈNCIA OAIS

El model de referència OAIS (*Open Archival Information System* o Sistema obert d'arxivament de la informació) per a l'arxivament digital a llarg termini es va establir el 2003 com a estàndard ISO (ISO 14721:2003). De fet, l'OAIS és un *model de referència* complex que intenta identificar i definir totes les tasques i els processos possibles en un arxivament digital a llarg termini. És important destacar que l'OAIS no representa una arquitectura real d'un arxiu, sinó que es tracta d'un *model* que ajuda a identificar i definir els components, les tasques i els processos dins d'una implementació real d'un arxiu. Bona part dels arxius digitals no implementen tots els processos disponibles derivats del model OAIS.

A més, el model OAIS fa algunes suposicions implícites sobre l'arxivament a llarg termini:

1. El mètode d'arxivament està basat en la migració. El model no és adequat per a altres mètodes d'arxivament a llarg termini.

2. Pressuposa grans institucions, amb un finançament garantit a llarg termini, que gestionin l'arxiu digital a llarg termini.

Tot i que el model OAIS resulta molt útil per identificar processos i dóna pistes sobre les millors pràctiques, és possible que només serveixi per a molts pocs casos com a camí per construir un arxiu digital a llarg termini.

Aplicació a la fotografia

Si bé moltes de les afirmacions i dels raonaments que s'han fet fins ara es poden aplicar a qualsevol tipus de dades digitals, la fotografia —ja sigui digitalitzada a partir de fotografies analògiques o digital de naixement— representa un tipus especial de dades digitals.

LA DIGITALITZACIÓ DE FOTOGRAFIES ANALÒGIQUES

Durant l'actual procés d'escaneig (digitalització) cal escollir amb cura i supervisar permanentment els paràmetres següents:

1. ***Resolució espacial***
   Cal adaptar la resolució espacial a l'original fotogràfic, és a dir, al contingut de la informació. En aquest sentit, la resolució espacial de l'emulsió fotogràfica és un paràmetre important, com també ho és la qualitat del sistema òptic emprat per fer la imatge original.

2. ***Resolució fotomètrica (reproducció de l'escala de grisos)***
   S'ha de reproduir per complet el grau de brillantor (contrast). En el domini digital, els nombres dels valors dels grisos (o, en el cas d'imatges en color, el nombre de valors per color primari) determinen el grau de precisió. Amb 8 bits, es poden representar 2^8 = 256 nivells diferents, amb 12 bits hi ha 4.096 nivells i 16 bits permeten 65.536 nivells de grisos diferents. El nombre de bits necessari ve determinat pel contrast de l'original. Això no obstant, aquests valors pràcticament no tenen cap valor sense una interpretació física: podria ser transmissió/reflexió (una escala lineal), densitat òptica (una escala logarítmica), una brillantor visual (CIE L*) o valors «sense calibrar». Per tant, cal un calibratge fotomètric de l'escàner, i el significat dels valors dels grisos o colors digitals s'han d'enregistrar amb la imatge. Les propietats dels diversos materials fotogràfics comporten les recomanacions següents per a una resolució fotomètrica mínima:

Positiu transparent (diapositiva)	≥ 12 bits
Negatiu transparent	≥ 14 bits
Impressió de reflexió, escala lineal	≥ 10 bits
Impressió de reflexió, escala logarítmica	≥ 8 bits

Per motius pràctics (arquitectura informàtica), l'emmagatzematge de les dades s'ha de fer en format de 8 bits o 16 bits (per a imatges en color són 3 x 8 bits = 24 bits o 3 x 16 bits = 48 bits):

Negatiu	16 bits
Diapositiva	16 bits
Impressió de reflexió	8 bits

Per a les impressions de reflexió, és possible reduir la representació interna de 10-12 bits amb una transformació logarítmica o una correcció «gamma» de 8 bits. En aquest cas, cal enregistrar la corba de transformació. El calibratge d'imatges amb valors de grisos es pot fer amb falques de grisos, mentre que el calibratge d'imatges en color requereix l'estàndard de calibratge per a escàner IT 8.7 que existeix tant per a transmissió com per a reflexió.

3. ***Calibratge de maquinari (escàner)***
   Els escàners tenen tres propietats que requereixen una atenció especial:
   1. L'anomenat «corrent de foscor» emet un senyal, fins i tot si el sensor està en una foscor absoluta.
   2. El soroll introdueix petites modificacions aleatòries en els valors de brillantor resultants. Especialment en àrees fosques, el soroll pot destruir amb eficàcia els petits detalls.

3. A més, la il·luminació pot ser que no sigui homogènia i que variï amb el temps.

Aquestes propietats que normalment no són constants durant la vida útil d'un escàner, cal tenir-les en compte per a un calibratge adequat i una supervisió periòdica de la qualitat.

La digitalització per a l'arxivament a llarg termini requereix un estricte control de la qualitat. D'una banda, el calibratge del maquinari —tal com es descriu més amunt— s'ha de repetir periòdicament. D'altra banda, cal un control visual *complet* de *cada* imatge digitalitzada a resolució completa. Les propietats següents s'han de supervisar mitjançant aquest control visual:

1. ***Nitidesa***
   El desgast mecànic, les vibracions, etc. poden modificar la geometria del sistema òptic de l'escàner i introduir-hi imatges borroses de manera sistemàtica.
2. ***Pols i brutícia***
   L'escàner (el vidre) es pot embrutar amb la pols i els residus dels originals.
3. ***Geometria***
   ¿Totes les imatges s'escanegen de manera correcta o es reflecteixen?
4. ***Errors d'escaneig***
   ¿La imatge és com s'esperava?

Aquesta supervisió hauria d'acompanyar permanentment el procés de digitalització per tal de detectar errors sistemàtics al més aviat possible.

Un altre aspecte és la completesa i la integritat de la col·lecció digitalitzada. Les col·leccions fotogràfiques que són prou valuoses per preservar-les constitueixen un conjunt que ha de romandre complet. Per a col·leccions grans, en què el procés de digitalització dura molt de temps, cal comprovar amb molt cura la completesa, perquè —per exemple— durant el procés es pot oblidar una imatge o es pot utilitzar dos cops un mateix nom de fitxer, etc.

FORMATS DE FITXERS D'IMATGES

Més enllà de les normes genèriques per a la preservació a llarg termini de les dades digitals (format obert i ben documentat), que també són aplicables als fitxers d'imatges digitals, la selecció curosa i correcta del format de fitxer hi juga un paper important. Afortunadament, les imatges digitals són objectes digitals relativament simples (en comparació, per exemple, amb les bases de dades relacionals). Tot i així, els fitxers d'imatges digitals poden arribar a ser complexos. Com que la representació digital d'imatges normalment genera objectes digitals entre grans (uns quants MB)[16] i enormes (200-300 MB), molts formats de fitxers d'imatges incorporen mètodes per reduir la petjada de les imatges digitals mitjançant tècniques de compressió. Cal distingir entre dos mètodes de compressió bàsicament diferents:

- ***Compressió sense pèrdues***
  Els esquemes de compressió sense pèrdues intenten reduir la redundància que normalment hi ha en les dades digitals. Els esquemes de compressió sense pèrdues no depenen de les propietats específiques de la imatge, ja que es poden aplicar a tot tipus de dades digitals. Com que les imatges acostumen a contenir una certa redundància (un valor de píxel en una determinada posició sovint és similar als valors de píxel adjacents), normalment es pot aconseguir una lleugera reducció de la quantitat de dades (entre un 30% i un 50%). La reducció serà més gran si la imatge conté grans àrees homogènies o àrees amb patrons repetitius. Per a les imatges que contenen grans àrees de patrons irregulars (soroll aleatori), la reducció serà molt reduïda. En alguns casos, la compressió sense pèrdues fins i tot pot inflar la mida del fitxer d'imatge. Un esquema típic de compressió sense pèrdues és l'algorisme LZW,[17] que es pot utilitzar en el format TIFF.

- ***Compressió amb pèrdues***
  La compressió amb pèrdues elimina informació que l'algorisme de compressió decideix que té poc interès. Així doncs, els algorismes de compressió amb pèrdues depenen directament de les propietats de les imatges i de les particularitats del sistema visual humà, motiu pel qual *modifiquen* la imatge de manera irreversible. Normalment només eliminen informació que es considera poc important per a l'observador. Tanmateix, si el factor de compressió és massa elevat o, encara més, si la imatge comprimida es manipula amb mètodes de processament d'imatges en una fase posterior (p. ex., millora del contrast), hi poden aparèixer artefactes. Per tant, la compressió amb pèrdues tan sols s'hauria d'aplicar a imatges que s'utilitzin *només per ser visualitzades*. Les imatges que s'han d'arxivar o que s'han de manipular en una fase posterior *mai* no s'haurien d'emmagatzemar en un format amb pèrdues! Els algorismes de compressió habituals (i els corresponents formats de fitxer) són:

  - ***JPEG***
    JPEG són les sigles en anglès de **J**oint **P**hotographers **E**xperts **G**roup, un esquema de compressió amb pèrdues consolidat que utilitza la *Transformada Discreta Cosinus*[18] (DCT, segons les seves sigles en anglès) com a base per a la compressió. La imatge es divideix en blocs de 8x8 píxels. La DCT es calcula per a cada bloc. D'acord amb el nivell de compressió, només s'utilitzen els coeficients més grans de la DCT. L'algorisme JPEG pot provocar artefactes molt particulars, que fan visible l'estructura de bloc de l'algorisme. La taxa de compressió normalment hauria de girar al voltant de 5-25. Les taxes de compressió més elevades sovint comportaran artefactes visibles.

  - ***JPEG2000***
    JPEG2000 és el successor de l'algorisme JPEG, però utilitza un enfocament totalment diferent. Es basa en la *Transformada de Wavelet*, que crea una piràmide de resolució de la imatge. Genera artefactes menys visibles, tot i que la imatge comprimida pot donar la impressió de menys nitidesa. També hi ha una variant sense pèrdues de l'algorisme JPEG2000.

    Una altra característica molt interessant de l'algorisme JPEG2000 és que no sempre cal llegir tot el fitxer d'imatge. Si només se'n llegeix la primera part, es pot visualitzar tota la imatge com si s'hagués comprimit amb una taxa de compressió més elevada. Per exemple, per tal de visualitzar una imatge en miniatura, només cal llegir el primer 5%-10% d'una imatge JPEG2000 (p. ex., variant sense pèrdues).

Normalment, els esquemes de compressió amb pèrdues no són recomanables per a finalitats d'arxivament. Els algorismes de compressió amb pèrdues *modifiquen* el contingut de la imatge i sempre poden introduir-hi artefactes i/o reduir la nitidesa de la imatge, cosa que provocarà una certa pèrdua de detalls. Tanmateix, si per determinats motius (p. ex., finançament, espai d'emmagatzematge, etc.) no és possible una compressió sense pèrdues, és millor aplicar les normes de l'arxivament a llarg termini a les imatges digitals amb compressió amb pèrdues que no pas no fer res.

La variant sense pèrdues de JPEG2000, que normalment redueix la petjada d'una imatge digital en més d'un 50%, pot ser una bona alternativa a la compressió amb pèrdues, fins i tot encara que aquest format no estigui gaire estès. Es tracta d'un format de fitxer d'imatge obert i documentat, si bé molt complex, que malauradament no està tan estès com seria desitjable.

Per tant, si no cal comprimir, el format TIFF és una bona elecció per a l'arxivament a llarg termini de fitxers d'imatges digitals. Si cal comprimir, JPEG2000 sense pèrdues o un esquema de compressió amb pèrdues, escollit amb cura, de la variant JPEG2000 pot ser una molt bona opció.

Informació de color i gestió del color

Una qüestió especial és la informació de color de les imatges digitals. Tal com s'ha indicat abans, les càmeres digitals normalment depenen de sensors que utilitzen tres filtres de color a la banda vermella, verda i blava de

la llum visible. En un cas ideal, aquesta informació hauria de ser suficient per reproduir la *impressió de color* que produeix un colorant en l'observador humà. La limitació a tres filtres és possible perquè l'ull humà també té tres tipus diferents de cèl·lules, que són sensibles a la banda vermella, verda i blava. De fet, però, els filtres dels sensors electrònics difereixen significativament de les corbes de sensibilitat de les cèl·lules de l'ull humà. Obviant qüestions com ara la metameria que requeriria molts filtres per suportar les imatges multiespectrals, el coneixement de les característiques de la càmera o de l'escàner, juntament amb les característiques de color d'un dispositiu de sortida, permetran un mapatge matemàtic dels valors de color per tal de crear una imatge de sortida que s'assembli a la imatge original tan com sigui possible des d'un punt de vista tècnic, amb la combinació específica de càmera i dispositiu de sortida. El procés de crear aquesta relació matemàtica s'anomena *calibratge del color*. Per fer un calibratge del color, cal calibrar tant el dispositiu d'entrada (càmera, escàner, etc.) com el de sortida (impressora, pantalla, projector) per tal d'establir-ne les característiques de color. Normalment, això es fa mitjançant la utilització de pedaços de color especialment dissenyats (p. ex., la carta de colors IT8) i colorímetres. Hi ha diversos sistemes comercials disponibles per fer-ho. Els perfils de color resultants (tant per al dispositiu d'entrada com per al de sortida) produeixen conjuntament una transformació matemàtica que mapeja els valors de color del dispositiu d'entrada al de sortida, de manera que l'observador humà vegi els colors iguals. Així doncs, la característica d'entrada —és a dir, *el perfil de color d'entrada*— s'ha d'emmagatzemar amb les dades de la imatge. L'ICC (*International Color Consortium* o Consorci Internacional del Color) ha establert un estàndard per als perfils de colors. Aquests perfils ICC es poden incrustar en molts formats de fitxers d'imatges com un element de dades opac (p. ex., TIFF, JPEG, PNG, EPS, PDF, SVG, etc. permeten incrustar perfils de colors). Cal destacar que el fitxer d'imatge en si no «sap» res sobre el perfil de colors. Simplement emmagatzema un fragment de dades i el transfereix a un programa de renderització d'imatges (p. ex., per visualitzar-lo en una pantalla o imprimir-lo) que, al seu torn, interpreta les dades del perfil de colors per renderitzar correctament els colors.

Per tant, cada imatge en color digital hauria d'anar acompanyada d'un perfil de colors que sigui específic per a cada dispositiu d'entrada. Tanmateix, sovint resulta més còmode transformar la informació de color a un dels perfils de colors estàndards estesos, com ara els perfils sRGB, AdobeRGB o ProPhoto. Cal destacar que, durant aquest tipus de transformació, es pot perdre certa informació de manera permanent, ja que la quantitat de colors (anomenada el gamut) que pot representar un perfil de colors específic pot ser més limitada que el gamut del perfil de colors original. Així doncs, utilitzar un perfil de colors estàndard té el gran avantatge que les dades del perfil de colors es poden ometre del fitxer d'imatge, sempre que s'inclogui la informació sobre el perfil de colors estàndard al qual responen les dades de la imatge.

CONCLUSIÓ

Actualment, en la majoria de casos, la longevitat de les dades digitals es pot aconseguir molt millor a través de la implementació d'un model de migració basat en les normes següents:

1. Redundància

Les dades s'han de conservar amb un nivell elevat de redundància. S'han de desar com a mínim tres còpies en un mínim de dos tipus diferents de mitjans d'emmagatzematge (p. ex., dues còpies al disc dur, una còpia en una cinta magnètica) en ubicacions geogràfiques diferents.

2. Sumes de verificació

Per a tots els fitxers de dades cal calcular sumes de verificació, que s'han d'arxivar amb els fitxers de dades. Això permetrà comprovar, en qualsevol moment, si un fitxer de dades ha canviat o conté errors com a conseqüència de l'envelliment.

3. Revisió

Cada 12-24 mesos cal revisar les dades i comparar les sumes de verificació. Si s'hi detecten errors, s'ha d'iniciar immediatament una migració.

4. Migració

Cal planificar per endavant les migracions, incloent-hi el finançament. Cal fer una migració del flux de bits cada cinc anys, aproximadament. Si un nou format de fitxer esdevé estàndard i la conversió pot fer-se sense perdre dades, és aconsellable fer una migració del format.

5. Documentació

Cal documentar detalladament cada pas i tots els mitjans s'han d'etiquetar correctament.

Si se segueixen aquestes normes, les dades digitals es poden preservar de manera indefinida. Tanmateix, cal tenir-ne cura constantment. Si això no és possible durant un determinat període de temps, les dades es perdran i és possible que només se'n pugui recuperar una part gràcies a l'arqueologia digital.

Una via alternativa pot ser l'ús d'un mitjà visual permanent, com ara un microfilm, per enregistrar dades analògiques, dades basades en text i dades binàries digitals. Aquest tipus de suport de dades és independent de qualsevol tecnologia específica i es pot tornar a llegir en el futur utilitzant dispositius simples de captació d'imatges. La interpretació dels bits es pot facilitar amb informació basada en text que contingui les instruccions.

Si bé el cost inicial d'emmagatzemar les dades digitals en uns microfilms pot ser més elevat, a llarg termini pot acabar resultant menys car que no pas el mètode de la migració. A més, a diferència del mètode de la migració, no requereix un flux continu de finançament.

NOTES

1_És a dir, Ilfochome Micrographic.

2_P. ex. S={0,1}, S={true, false}, S={+5V, -5V} o S={#, $}

3_Vegeu, per exemple: A. Ribes, H. Brettel, F. Schmitt, H. Liang, J. Cupitt and D. Saunders, *Color and Spectral Imaging With the Crisatel Acquisition System*, PICS03 The Digital Photography Conference (2003).

4_Aquesta afirmació no és certa en el món de la física quàntica. En aquesta àrea hi ha propietats físiques com ara un spin electrònic que té dos valors diferents: amunt o avall. Des d'aquest punt de vista, un spin electrònic pot representar un bit d'una manera veritablement digital.

5_A dins d'un ordinador també té lloc el mateix procés de llindaritzar. Els 0 i 1 es representen mitjançant el corrent elèctric o voltatge, i en algun lloc s'ha de definir un punt «basculant» que distingeixi entre ambdós estats.

6_Claude E. Shannon, «A mathematical theory of communication», Bell System Technical Journal (1948).

7_Aquest valor prové del full de dades d'un important fabricant de discs durs per a un disc dur IDE de 400 GB.

8_John W.C. Van Bogart, «Magnetic Tape Storage and Handling», National Media Laboratory (1995).

9_Vegeu per exemple PNDesign, «Data formats, file extensions database», a http://extensions.pndesign.cz

10_Kenneth Thibodeau, *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*, Conference Proceedings: The State of Digital Preservation: An International Perspective (2002).

11_WordStar va ser un processador de textos molt reeixit per al sistema operatiu CP/m desenvolupat el 1978.

12_R. A. Lorie, *Long-Term Preservation of Complex Processes*, Proceedings of the IS\&T Archiving Conference (2005).

13_Normand, C.; Gschwind, R.; Fornaro, P., *Digital images for eternity: color microfilm as archival medium*, Color Imaging XII:

Processing, Hardcopy, and Applications. Edited by Eschbach, Reiner; Marcu, Gabriel G. Proceedings of the SPIE, volum 6.493, pàg. 649.307 (2007).

14_Ariel Amir, Florian Müller, Peter Fornaro, Rudolf Gschwind, Joachim Rosenthal, Lukas Rosenthaler: *Toward a Channel Model for Microfilm*. IS&T's 2008 Archiving Conference Proceedings, Berna, juny 2008. IS&T: The Society for Imaging Science and Technology, Springfield (VA), EUA.

15_Müller, F., Fornaro, P., Rosenthaler, L., and Gschwind, R. 2010. PEVIAR: Digital originals. ACM J. Comput. Cult. Herit. 3, 1, Article 2 (juny 2010), 12 pàgines. DOI = 10.1145/1805961.1805963, http://doi.acm.org/10.1145/1805961.1805963.

16_1 MB representa 1.000.000 bytes de dades.

17_Jacob Ziv and Abraham Lempel; *Compression of Individual Sequences Via Variable-Rate Coding*, IEEE Transactions on Information Theory, setembre 1978.

18_La Transformada Discreta Cosinus, com la Transformada de Fourier, és un mètode matemàtic per descriure una funció d'una o dues dimensions en forma d'un espectre de freqüència.

RESUMEN

La longevidad de los datos digitales se puede conseguir a través de la implementación de un modelo de migración basado en las siguientes normas:

Redundancia: Se deben guardar como mínimo tres copias de dos tipos diferentes de soportes en ubicaciones geográficas diferentes.

Sumas de verificación: se deben calcular sumas de verificación que se han de archivar con los ficheros de datos. Permitirán comprobar si un fichero ha cambiado o contiene errores.

Revisión: Cada 12-24 meses hay que revisar los datos y comparar las sumas de verificación. Si se detectan errores, se debe iniciar inmediatamente una migración. Migración: Se debe hacer cada cinco años aproximadamente.

Documentación: Se debe documentar detalladamente cada paso y todos los medios se han de etiquetar correctamente.

Una vía alternativa puede ser el uso de un medio visual permanente, como es el microfilm, para registrar datos analógicos, datos basados en texto y datos binarios digitales. Si bien el coste inicial de almacenar los datos digitales en microfilms puede ser más elevado, a largo plazo puede acabar resultando menos caro que la migración.

RÉSUMÉ

La longévité des informations numériques peut être obtenue en mettant en place un modèle de migration reposant sur les normes suivantes :

Redondance. Il faut conserver au moins trois copies sur deux types de supports différents dans des lieux différents.

Sommes de vérification. Il faut calculer des sommes de vérification qui doivent être archivées avec les fichiers de données et qui permettront de vérifier si un fichier été modifié ou s'il contient des erreurs.

Révision. Tous les 12 à 24 mois, il faut examiner les données et comparer les sommes de vérification. Si des erreurs sont détectées, une migration doit être immédiatement réalisée. La migration doit être effectuée environ tous les cinq ans.

Documentation. Il est nécessaire de documenter en détail chaque étape et tous les milieux doivent être étiquetés correctement.

Une autre méthode pourrait être l'utilisation d'un milieu visuel permanent, tel que le microfilm, qui permet l'enregistrement de données analogiques, de données fondées sur du texte et de données binaires numériques. Bien qu'au départ, le stockage des données numériques sur des microfilms puisse être plus coûteux, il peut à long terme se révéler plus économique que la migration.

SUMMARY

The longevity of digital data can be best achieved by implementing a migration model based on the following rules:

Redundancy: Data have to be kept at least 3 copies on a minimum of 2 different types of storage media. Should be kept at geographically different locations.

Checksums: For all data files, checksums should be calculated and archived with the data files. This allows at any time to check if a data files has changed or contains errors due to aging.

Proofreading: Every 12-24 months, the data should be proofread and the checksums compared. If errors are detected, a migration should be launched immediately.

Migration: A bitstream migration is necessary about every 5 years.

Documentation: Every step has to be documented in detail, all media have to be labeled properly.

An alternative may be the use of a permanent visual medium such as microfilm to record analogue, text-based and digital binary data. While the initial cost of a storing digital data on microfilm may be higher, it may be less expensive than the migration method on the long-term.