

Formazione, gestione e conservazione degli archivi digitali

Il Master FGCAD dell'Università degli Studi di Macerata

a cura di Giorgetta Bonfiglio Dosio
e Stefano Pigliapoco



eum x archivistica informatica

eum x archivistica informatica

Formazione, gestione e conservazione degli archivi digitali

Il Master FGCAD

dell'Università degli Studi di Macerata

a cura di Giorgetta Bonfiglio Dosio
e Stefano Pigliapoco

eum

isbn 978-88-6056-437-5

Prima edizione: ottobre 2015

©2015 eum edizioni università di macerata

Centro direzionale, via Carducci snc – 62100 Macerata

<http://eum.unimc.it>

info.ceum@unimc.it

Indice

7	Giorgetta Bonfiglio Dosio Prefazione
13	Giorgetta Bonfiglio Dosio L'archivio digitale: specificità ed esigenze formative degli archivisti
29	Stefano Pigliapoco Formare professionisti esperti di archiviazione e conservazione digitale. Il Master FGCAD
105	Federico Valacchi «Per l'interesse della scienza e del pubblico servizio». Una Cibrario 2.0 che riconosca agli archivi «il potere degli archivi».
167	Giulio M. Salerno Documento informatico e trasparenza delle pubbliche amministrazioni
189	Pierluigi Feliciati I metadati nel ciclo di vita dell'archivio digitale e l'adozione del modello PREMIS nel contesto applicativo nazionale
209	Stefano Allegrezza La componente tecnologica nei processi conservativi digitali
241	Andrea Fradeani, Alessandro d'Eri, Francesco Campanari La codifica XBRL dei bilanci redatti secondo i principi contabili internazionali: la sperimentazione su un campione di aziende di grandi dimensioni

Stefano Allegrezza

La componente tecnologica nei processi conservativi digitali

Introduzione

La conservazione del patrimonio digitale che si è venuto formando in questi anni, sia con la produzione di documenti nativi digitali che con la massiccia attività di digitalizzazione di documentazione cartacea presente in archivi, biblioteche e musei, è al centro di un dibattito che da qualche decennio¹ coinvolge la comunità scientifica internazionale nel tentativo di trovare risposte condivise e, possibilmente, definitive al problema della *digital preservation* (conservazione digitale)². I documenti digitali, infatti, per la loro dinamicità e facilità di modifica, per la fragilità dei supporti sui quali sono memorizzati, per la dipendenza da tecnologie spesso assai complesse pongono inediti problemi di conservazione, accessibilità e affidabilità nel corso del tempo.

¹ Il problema della conservazione degli archivi digitali è tra quelli che maggiormente assilla la comunità degli archivisti, la quale si interroga su quale sarà il loro futuro fin dagli inizi degli anni '90, come ci viene ricordato dagli atti di un convegno internazionale organizzato nel 1993 dalla Fondazione IBM Italia e significativamente intitolato «L'eclisse delle memorie». Si veda Maria Guercio, *La conservazione digitale nello scenario europeo e internazionale: principi, metodi, progetti. Conferenza sul libro elettronico. Atti*, Napoli, 2003, disponibile all'indirizzo <<http://eprints.erpanet.org/64/01/e-book.pdf>>.

² Tra le varie strategie di conservazione proposte si segnalano il trasferimento su supporti analogici (*output to analogue media*); la conservazione tecnologica (*technology preservation*); il riversamento diretto (*refreshing*) e sostitutivo (*migration*); l'emulazione (*emulation*); l'archeologia digitale (*digital archaeology*). Alcune di queste verranno prese in esame nel seguito.

Nell'ambito del Master in «Formazione, Gestione e Conservazione di Archivi Digitali in Ambito Pubblico e Privato» (nel seguito: Master FGCAD) queste tematiche vengono affrontate in maniera sistematica, ed in particolare viene dato ampio spazio alla componente tecnologica per l'importanza che questa assume nei processi conservativi digitali.

Il punto di partenza da cui è opportuno partire per comprendere il motivo di tale rilevanza è costituito dalla constatazione che la maggior parte delle informazioni oggi prodotte sono digitali: per scrivere un documento di testo non si usa più carta e penna ma si utilizza – ormai da qualche decennio – un *word processor*, che produce un documento digitale; per immortalare i propri ricordi si utilizza una macchina fotografica digitale, che produce immagini digitali; per registrare una lezione di un corso di formazione si utilizza un *voice recorder*, che produce una registrazione audio digitale; per videoregistrare un evento, si utilizza la videocamera digitale, che produce anch'essa una registrazione video digitale. Nel mondo delle comunicazioni la situazione è analoga: le lettere del passato (quelle scritte su carta, imbustate, affrancate e spedite) hanno ormai ceduto il posto ai messaggi di posta elettronica (e-mail) e ai brevi messaggi di testo (SMS) o agli scambi tramite *Whatsapp*. Il sistema di telefonia cellulare è anch'esso digitale, così come il sistema televisivo che ha ormai completato la transizione verso il “digitale terrestre”. Se, poi, ci spostiamo verso altri settori della vita quotidiana, scopriamo che anche le informazioni relative al conto corrente non sono altro che dati digitali memorizzati in un *database* della banca, così come sono digitali le informazioni previdenziali conservate nell'archivio dell'istituto di previdenza, le informazioni relative alla carriera scolastica di uno studente universitario presenti nell'archivio dell'ateneo, e così via: non esiste più alcun campo in cui la “rivoluzione digitale” non abbia sortito i suoi effetti. Quelli appena visti sono solamente alcuni degli innumerevoli esempi che si potrebbero fare, ma sono sufficienti per comprendere quale sia il trend evolutivo verso il quale ci si sta dirigendo: è molto probabile (se non quasi certo) che in un futuro neanche troppo lontano la componente

digitale degli archivi sarà sempre più consistente, se non addirittura esclusiva.

Purtroppo, i documenti digitali (o, più in generale, i “contenuti” od “oggetti” digitali) sono estremamente “fragili” e rischiano di non essere più accessibili anche a distanza di pochi anni dal momento della loro formazione. Ma perché è così difficile conservare i documenti digitali rispetto a quelli analogici? Semplificando le cose in maniera molto drastica si può rispondere con una semplice osservazione: per la fruizione di un documento analogico (si pensi, ad esempio, ad una lettera tradizionale su carta) è sufficiente solamente una buona vista³; al contrario, per la fruizione di un documento digitale è necessario avere a disposizione anche un *supporto* di memorizzazione (o, con locuzione inglese, *media*) su cui tale documento è stato registrato, le informazioni circa il *formato* elettronico secondo il quale tale documento è stato codificato ed, infine, un'apposita infrastruttura tecnologica (*hardware* e *software*) che ne consenta la fruizione⁴. Ed è proprio l'evoluzione tecnologica continua ed incessante che queste componenti subiscono che rende difficile la conservazione dei documenti digitali, dal momento che ciascuna di esse va rapidamente incontro al fenomeno dell'obsolescenza. Infatti:

- 1) i *supporti* di memorizzazione vanno incontro ad una rapida obsolescenza, legata al fatto che il mercato propone via

³ In realtà, alcuni contenuti analogici necessitano anch'essi della mediazione di strumenti tecnologici per la loro fruizione – si pensi ai vecchi dischi in vinile o alle tradizionali cassette audio musicali – ma in generale si tratta di strumenti tecnologici che accompagnano l'intera vita del supporto su cui tali contenuti sono memorizzati.

⁴ Infatti, in ultima analisi, un oggetto digitale altro non è che una sequenza di bit “0” ed “1” (il cosiddetto *bitstream*), che non sarebbe assolutamente leggibile né intelligibile da un essere umano se non ci fosse la mediazione di questo “strato tecnologico”: quando si vuole accedere al contenuto “racchiuso” nell'oggetto digitale memorizzato su un determinato supporto di memorizzazione, questa sequenza deve essere letta mediante un'infrastruttura *hardware* e *software* ed interpretata sulla base della conoscenza del suo formato. Il risultato dell'interpretazione, mostrato solitamente su schermo o a stampa, è una rappresentazione del contenuto digitale.

via supporti sempre più moderni che rendono obsoleti quelli utilizzati fino a poco tempo prima⁵;

- 2) i *formati* elettronici vanno anch'essi incontro ad una rapida obsolescenza, legata alla loro rapida evoluzione e al fatto che taluni formati (ad esempio quelli proprietari) dipendono fortemente dall'applicazione con cui debbono essere interpretati e che nel tempo potrebbe non essere più disponibile;
- 3) infine, anche le infrastrutture tecnologiche vanno rapidamente incontro all'obsolescenza a causa della rapida evoluzione dell'*hardware* e del *software* e al fatto che il mercato propone incessantemente sistemi sempre più moderni, rendendo obsoleti quelli che solamente ieri costituivano l'ultima novità⁶.

Già da queste prime riflessioni si può intuire l'importanza della componente tecnologica che deve essere attentamente studiata e compresa da chi vuole acquisire le competenze necessarie per gestire i processi conservativi digitali; è per questo che nell'ambito del Master FGCAD si viene incontro a questa esigenza dando ampio risalto sia alle tematiche relative ai supporti di memorizzazione ed ai sistemi di *storage management* che a quelle relative ai formati elettronici.

⁵ Il motivo per cui i supporti di memorizzazione diventano rapidamente obsoleti dipende dal fatto che nel tempo vengono introdotte nuove tecnologie che rendono obsoleti i *drive* necessari per leggere i vecchi supporti di memorizzazione (che diventano quindi "illeggibili" perché non sono più disponibili i *drive* che ne consentono la lettura).

⁶ Da diversi decenni si assiste ad una continua evoluzione di tali infrastrutture: il computer, il tablet, lo smartphone appena acquistati dopo solamente due o tre anni sono già obsoleti e vanno sostituiti con i modelli più moderni e potenti; i sistemi operativi vanno anch'essi aggiornati alle nuove versioni e parallelamente i programmi necessari per creare e gestire i contenuti digitali devono essere anch'essi aggiornati o sostituiti con altri. Nel giro di pochi anni l'evoluzione tecnologica rende del tutto obsoleta l'infrastruttura *hardware* e *software* utilizzata fino a poco tempo prima rendendo necessaria la sua sostituzione.

1. Aspetti tecnologici legati ai supporti di memorizzazione

Uno dei principali aspetti tecnologici che deve essere preso in considerazione è quello relativo ai supporti di memorizzazione. Le risorse digitali non sono “immateriali” ma sono costituite da bit che hanno bisogno, per poter semplicemente “esistere” e per poter essere tramandati nel tempo, di un supporto sui cui risiedere⁷. Vi è una grande varietà di supporti: dalle schede e dai nastri perforati di storica memoria, che rappresentavano i bit attraverso una serie di fori, alla vasta gamma di supporti di memorizzazione oggi disponibili basati sulle tecnologie magnetiche (dai *floppy disk* – ormai obsoleti – ai dischi fissi e ai nastri magnetici), ottiche (CD, DVD, *Blu-ray Disc*) ed elettroniche (*pendrive*, *memory card*, SSD, etc.).

Nel corso di questi anni si è assistito ad una rapidissima e vorticosa evoluzione di tali supporti (si veda la figura 26): basti pensare che fino alla fine degli anni '80 si utilizzava ancora il *floppy disk* da 5 pollici e $\frac{1}{4}$, che attualmente è solo un reperto storico e risulta sostanzialmente impossibile da leggere; il suo successore, il *floppy disk* da 3" e $\frac{1}{2}$, è rimasto in auge per tutti gli anni '90, ma già agli inizi del nuovo secolo il suo utilizzo cominciava a diventare sempre meno frequente, soppiantato da altre memorie più capienti sino a quando, all'inizio del 2007, è terminata la produzione a livello mondiale di questo supporto oggi obsoleto. Nel campo dei supporti ottici le cose non sono andate diversamente: quando, agli inizi degli anni '80, vennero prodotti i primi CD (*Compact Disc*), si pensava che la loro capacità di memorizzazione fosse talmente elevata che sarebbe stata sufficiente per diversi decenni a venire. Invece, dopo solo una quindicina di anni i CD sono stati soppiantati dai DVD, i quali

⁷ Questi supporti possono essere considerati l'equivalente moderno dei supporti scrittori che hanno consentito alla memoria dell'umanità di giungere fino ai giorni nostri: si pensi alle tavolette cerate o lignee e ai fogli di papiro adoperati nella civiltà greco-romana; ai rotoli di papiro che vennero utilizzati dai greci e adottati successivamente anche dai romani, fino a quando non furono sostituiti dalla pergamena; o, infine, alla carta adoperata dal Duecento fino ad oggi.

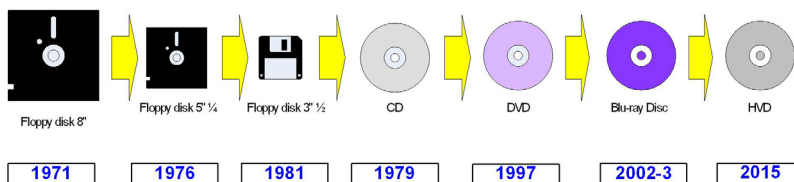


Figura 26. Alcune generazioni di supporti che si sono succeduti negli ultimi decenni

a loro volta sono stati soppiantati dai nuovi supporti con capacità ancora maggiore (quali gli HD-DVD e i *Blu-ray Disc*); ma anche questi non avranno una vita eccessivamente lunga⁸ perché saranno sostituiti dai nuovi supporti che attualmente sono in fase di sviluppo (ad esempio, i dischi olografici) e che consentiranno capacità di archiviazione finora impensabili, e così via: la regola, oggi ed in futuro, è proprio questo perenne avvicendamento di supporti in favore dei nuovi *media* che vengono via via proposti e che in breve tempo rendono obsoleti quelli utilizzati fino a poco tempo prima con la conseguenza che i documenti registrati su di essi diventano inaccessibili⁹. Accanto a questo problema occorre sempre tenere presente che nessun supporto di memorizzazione – qualunque esso sia e per quanto sia tecnologicamente e costruttivamente valido – è in grado di assicurare una capacità di riten-

⁸ Si pensi che l'HD-DVD (*High Definition DVD*), proposto intorno agli anni 2002-2003 è diventato obsoleto nel giro di pochi anni e già alla fine del 2008 non era più utilizzato.

⁹ Riprendendo l'esempio proposto poc'anzi, i dati memorizzati su un *floppy disk* da 5 pollici e 1/4 sono ormai inaccessibili se non si è provveduto a suo tempo a riversare il contenuto da quel supporto ad un altro supporto più recente. Per un elenco ragionato di memorie che sono diventate obsolete negli ultimi decenni si può fare riferimento alla "Chamber of Horror", disponibile all'interno del *Digital Preservation Management Tutorial*; le memorie sono classificate in *Punch Media* (supporti perforati), *Disks* (dischi), *Tapes* (nastri) e *Solid State Media* (supporti allo stato solido). Si veda <www.dpworkshop.org/dpm-eng/oldmedia/chamber.html>.

zione¹⁰ dei dati illimitata nel tempo: prima o poi esso perde le sue caratteristiche fino a determinare l'illeggibilità dei dati registrati su di essi.

Per quanto l'obsolescenza dei supporti di memorizzazione sia solo un aspetto del più complesso problema della conservazione digitale, l'analisi dei supporti e l'individuazione di criteri utili per la scelta dei supporti compatibili con un processo di conservazione digitale a lungo termine ha una funzione strategica, e ciò per almeno tre ordini di motivi. Innanzitutto, negli istituti di conservazione vi è ormai una quantità notevole di documentazione digitalizzata registrata su supporti di memorizzazione di varia natura (dai nastri magnetici ai CD, DVD, etc.): è prevedibile, inoltre, che nel tempo la quantità di materiale digitale non possa che aumentare, e con essa l'impiego di supporti di memorizzazione. In secondo luogo, sebbene le organizzazioni più grandi, sia pubbliche che private, si stiano orientando verso soluzioni più sofisticate e sicure che prevedono, ad esempio, la creazione di poli di conservazione digitale a cui demandare la responsabilità della conservazione, gli istituti più piccoli, gli archivi di piccole e medie dimensioni, i professionisti e le persone singole basano generalmente la conservazione della documentazione digitale prodotta – che è destinata a costituire parte integrante del futuro patrimonio culturale – sull'utilizzo *sic et simpliciter* dei supporti di memorizzazione. Infine, i sistemi di *storage management* (di cui si parlerà nel paragrafo 2), dai più semplici ai più complessi, in ultima analisi utilizzano anch'essi i supporti di memorizzazione per registrare le informazioni. Come si può intuire, risulta fondamentale analizzare le varie tipologie di supporti di memorizzazione ed indagare in merito alla loro capacità di garantire la conservazione nel tempo dei documenti digitali ed il punto di partenza è costituito dalla conoscenza dei principi di funzio-

¹⁰ Per "capacità di ritenzione" di un supporto si intende il lasso di tempo durante il quale è garantita la conservazione dei dati registrati su di esso; ad esempio, se un supporto ha una capacità di ritenzione di dieci anni, sarà possibile "leggere" e recuperare senza errori i dati dopo dieci anni dal momento della loro registrazione sul supporto.

namento di ciascun tipo di supporto, necessaria per arrivare a comprendere i suoi punti di forza e le sue criticità¹¹.

La strategia che è stata proposta per risolvere il problema dell'obsolescenza dei supporti prende il nome di *riversamento diretto* (*refreshing*) e consiste nel riversare periodicamente gli archivi digitali da un supporto di memorizzazione obsoleto ad un nuovo supporto, senza modificare la sequenza di bit di cui è costituito. Il nuovo supporto può essere dello stesso tipo (come nel caso del riversamento da un CD verso un altro CD) o di tipo diverso (ad esempio, da un CD ad un DVD). Si tratta di un processo che è sempre necessario, indipendentemente da qualunque altra strategia di conservazione che venga contestualmente adottata. Dal punto di vista tecnico è relativamente semplice, con un bassissimo rischio di perdita di dati se viene eseguito e documentato nella maniera corretta. Il riversamento diretto è un processo molto critico nei confronti del tempo: non ammette alcuna dilazione e necessita di essere eseguito prima che il supporto corrente diventi obsoleto. Se viene trascurato il riversamento per una generazione di supporti, gli archivi potrebbero essere difficili da recuperare; se viene trascurato per più generazioni, gli archivi potrebbero andare completamente persi¹². I cicli

¹¹ Ad esempio, le memorie magnetiche sono sensibili ai campi elettromagnetici; le memorie ottiche sono sensibili ai raggi ultravioletti e all'ossidazione dello strato riflettente; le memorie elettroniche hanno un numero limitato di cicli di scrittura/lettura ed una durata limitata del tempo di ritenzione dei dati memorizzati su di esse, e così via.

¹² Si tenga anche presente che il processo di riversamento da un supporto all'altro diventa sempre più difficile man mano che passano gli anni. Un esempio servirà a chiarire il concetto. Quando, alla fine degli anni '70, i *floppy disk* da 8" cominciarono a diventare obsoleti soppiantati dai *floppy disk* da 5" e ¼, era molto semplice effettuare il riversamento diretto dei dati contenuti sui supporti da 8" su quelli da 5" e ¼ perché esistevano computer che avevano contemporaneamente sia il *drive* per il primo tipo di *floppy* che quello per il secondo tipo. Successivamente, alla fine degli anni '80, quando anche il *floppy disk* da 5" e ¼ cominciò a diventare obsoleto, soppiantato dal *floppy disk* da 3" e ½, era semplice effettuare il riversamento dei dati contenuti nei *floppy disk* da 5" e ¼ sui nuovi supporti, perché erano disponibili computer che avevano entrambi i tipi di *drive*. Invece, il riversamento dai *floppy disk* da 8" verso quelli da 3" e ½ era improvvisamente diventato molto difficile, perché i *drive* per i *floppy disk* da 8" non erano più in produzione e, di conseguenza, non erano disponibili computer dotati di entrambi i *drive*. Questo esempio, che si potrebbe generalizzare a molti altri casi, fa

di *refreshing* devono essere relativamente frequenti, dal momento che pochi supporti di memorizzazione sono oggi in grado di sopravvivere più di una decina di anni senza essere sottoposti ad alcun riversamento diretto.

Il problema dell'obsolescenza dei supporti non è di facile soluzione. Tuttavia vi sono alcuni tentativi di produrre supporti di memorizzazione che abbiano maggiore affidabilità e longevità. Tra di questi merita certamente una segnalazione il *Millennium Disc* (o, più brevemente, *M-Disc*), creato dall'azienda statunitense Millenniata. Secondo le dichiarazioni rilasciate dalla casa costruttrice, l'*M-Disc* dovrebbe assicurare una capacità di ritenzione di oltre mille anni; una tale longevità, verificata anche mediante studi di laboratorio¹³, ha fatto sì che per l'*M-Disc* venisse definita una nuova categoria di appartenenza, denominata WORM (*Write Once Read Forever*)¹⁴. I primi supporti *M-Disc*, disponibili dal 2011, avevano una capacità di 4,7 GB, la stessa dei DVD

comprendere come il processo di riversamento rischi di diventare irrealizzabile se si "salta" anche solamente una generazione di supporti.

¹³ I test sono stati condotti sulla base della norma ISO/IEC 10995:2011 *Information Technology - Digitally Recorded Media for Information Interchange and Storage - Test Method for The Estimation of the Archival Lifetime of Optical Media*. Dagli esperimenti di laboratorio condotti, si è potuto concludere che l'*M-Disc* è estremamente resistente ed affidabile: può addirittura essere immerso nell'azoto liquido a una temperatura di -180 °C e poi trasferito in acqua bollente senza subire danni; si può graffiare e seppellire per cento anni, ma i dati scritti sugli *M-Disc* rimangono leggibili e non vengono persi. A conferma delle sue enormi doti di resistenza, l'*M-Disc* ha ottenuto anche una certificazione dal Ministero della Difesa statunitense. In uno studio comparato effettuato a China Lake in California, infatti, l'*M-Disc* è stato messo a confronto con alcuni tra i dischi per l'archiviazione a lungo termine più venduti al mondo (quelli prodotti da Mitsubishi, Verbatim, Taiyo Yuden, Delkin e MAM-A). Il risultato dell'esperimento, che si prefiggeva lo scopo di simulare la loro resistenza in situazioni limite, è stato che solamente l'*M-Disc*, a differenza degli altri, non ha subito alcun danno. Le conclusioni sono state pubblicate nello studio *Accelerated Life Cycle Comparison of Millenniata Archival DVD*, disponibile all'indirizzo <site.produplicator.com/downloads/Manuals/China_Lake_Full_Report.pdf>.

¹⁴ Si ricorda che sulla base della possibilità di lettura e scrittura da parte degli utenti i supporti di memorizzazione vengono tradizionalmente classificati in: ROM (*Read Only Memory*) che non possono essere scritti ma solo letti; WORM (*Write Once, Read Many*), che possono essere scritti una sola volta, ma letti più volte; WMRM (*Write Many, Read Many*), che possono essere scritti e letti più volte.

single side, single layer; dal 2013 sono stati immessi sul mercato supporti con capacità di 25 GB, la stessa dei *Blu-ray Disc*.

Nell'ambito dell'insegnamento di "Supporti di memorizzazione e sistemi di *storage management*" erogato nel Master FGCAD, vengono trattate le principali tipologie di memorie di massa; in particolare vengono approfonditi gli aspetti tecnologici alla base del funzionamento di queste memorie¹⁵, la cui comprensione è necessaria per capire quali sono le criticità ed i punti di forza di ciascuna tipologia e per valutarne l'affidabilità; vengono poi presi in esame i supporti che hanno oggi una maggiore diffusione sia in ambito professionale che *consumer*¹⁶; vengono, infine, forniti alcuni cenni in merito alle memorie del passato che rivestono particolare importanza sotto il profilo storico o scientifico e alle memorie che sono ancora in fase di studio ma che con ogni probabilità diventeranno le memorie del futuro. Vista la vasta adozione che hanno avuto, particolare risalto è riservato alle memorie ottiche che costituiscono, per le loro caratteristiche, una tipologia di supporto ampiamente utilizzato da oltre venti anni negli archivi digitali, soprattutto (ma non solo) in quelli di piccola e media dimensione.

¹⁵ Sulla base della tecnologia utilizzata per memorizzare i dati, le memorie si possono classificare in: memorie cartacee (le vecchie schede perforate ed i nastri perforati), memorie magnetiche, che utilizzano le proprietà dei campi elettromagnetici; memorie ottiche, che utilizzano un raggio laser sia in fase di lettura che di scrittura; memorie magneto-ottiche, che utilizzano sia le proprietà dei campi elettromagnetici che un raggio laser; memorie elettroniche, basate su particolari componenti elettronici (quali, ad esempio, i transistor); memorie olografiche, basate sulla tecnologia impiegata negli ologrammi.

¹⁶ Con il termine "consumer", di evidente derivazione inglese, si fa riferimento a quelle situazioni in cui le memorie digitali vengono utilizzate per esigenze di natura privata. Si pensi, per fare un esempio attuale, agli archivi digitali di persona, costituiti dalla documentazione digitale che va sedimentandosi con sempre maggior frequenza nelle varie memorie che ciascuna persona possiede (si pensi alle chiavette USB, ai *Compact Disc* e ai DVD su cui viene archiviata la propria documentazione personale, ai messaggi di posta elettronica salvati sul disco fisso del proprio computer, etc.).

2. Aspetti tecnologici legati ai sistemi di storage management

Se è vero che è importante conoscere le caratteristiche dei supporti di memorizzazione in maniera da avere dei criteri per la loro scelta, è anche vero che è importante conoscere i sistemi di *storage management*, ovvero quei sistemi tecnologici che sovrain-tendono alla gestione dell'attività di archiviazione dei dati e che si basano, appunto, sull'utilizzo di supporti di memorizzazione. Per questo motivo, nell'ambito del Master FGCAD è previsto uno specifico modulo di approfondimento sui sistemi di *storage management* e sui criteri per la loro scelta.

Per assicurare la protezione dei dati nel tempo, questi sistemi si sono tradizionalmente basati sul fondamentale principio della *ridondanza*, ovvero sulla presenza di più dispositivi utilizzati per svolgere una determinata funzione ed organizzati in maniera tale che un guasto del sistema possa verificarsi solo in conseguenza del guasto contemporaneo di *tutti* questi dispositivi. Nel caso dei sistemi di *storage management* la ridondanza ha lo scopo di aumentarne l'affidabilità e garantire la protezione dei dati; normalmente si attua mediante le tecniche di duplicazione, di triplicazione o, in generale, di replica multipla dei componenti critici di un sistema. Tra i sistemi che, in varia misura, vengono utilizzati per mettere in pratica questo principio è opportuno conoscere almeno i sistemi di *backup* e *restore* ed i sistemi RAID.

I sistemi di *backup* e *restore* (copia e ripristino) consentono di effettuare delle copie degli archivi memorizzati su un determinato supporto di memorizzazione su un altro supporto; qualora gli archivi presenti sul primo supporto non dovessero essere più accessibili – ad esempio in seguito a guasti *hardware*, errori umani od altre evenienze – è possibile recuperarli a partire da quelli salvati sul secondo supporto: questa operazione è detta *restore* dei dati. La periodica attività di *backup* è senza dubbio la prima regola da seguire nella gestione degli archivi digitali e costituisce una funzionalità fondamentale di qualsiasi sistema di *storage management* perché è l'unica che può evitare la perdita

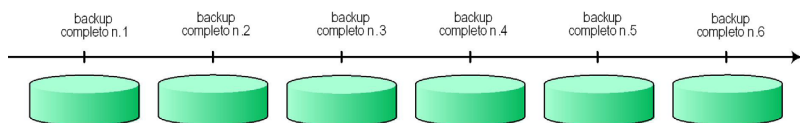


Figura 27. Il *backup* completo

degli archivi nel caso si verifichi un guasto, una manomissione, un evento calamitoso, un furto, etc.¹⁷

Nell'ambito del Master FGCAD, allo scopo di fornire agli allievi conoscenze subito spendibili sul mercato del lavoro, vengono illustrate in dettaglio le varie tipologie di *backup* oggi disponibili (asincrono e sincrono¹⁸; locale e remoto¹⁹, on-line ed off-line²⁰). Inoltre, vengono illustrate le diverse strategie di *backup* normalmente utilizzate: completo, differenziale ed incrementale. Il *backup completo*, il più semplice ma anche il meno efficiente,

¹⁷ Nei centri di conservazione digitale il tipo di *backup* e la relativa periodicità sono solitamente regolati da una apposita procedura soggetta a verifica periodica. Il responsabile della conservazione è tenuto ad annotare i controlli periodici e gli interventi sui sistemi. I supporti su cui viene effettuato il *backup* normalmente devono essere di tipo e marca approvati nella procedura ed è necessario che siano periodicamente verificati e sostituiti. Devono inoltre essere conservati in accordo con le politiche di sicurezza aziendale, per esempio, ma non solo, per questioni legate alla *privacy*.

¹⁸ Con il *backup* asincrono la copia viene eseguita in determinati istanti temporali (ad esempio, una volta al giorno); invece con il *backup* sincrono si ha una copia sempre aggiornata dei dati (come se il *backup* venisse eseguito virtualmente in ogni istante). È evidente che questa seconda tipologia richiede investimenti tecnologici maggiori rispetto ai *backup* di tipo asincrono ed è giustificata nei casi in cui non è accettabile la benché minima perdita dei dati.

¹⁹ Nel primo caso le copie di *backup* vengono collocate nello stesso locale o in un locale attiguo a quello in cui, in caso di necessità, dovrebbe essere effettuato il *restore*. Nel secondo caso le copie di *backup* vengono invece collocate a distanza rispetto al locale in cui verrà eventualmente effettuato il *restore*.

²⁰ La prima prevede che il supporto sul quale si effettua la copia dei dati sia sempre immediatamente disponibile al sistema (ad esempio: un NAS, un disco fisso dedicato). La seconda prevede che, una volta effettuata la procedura, il supporto venga archiviato a sua volta e non sia immediatamente accessibile dal sistema (ad esempio: supporti ottici come CD e DVD).

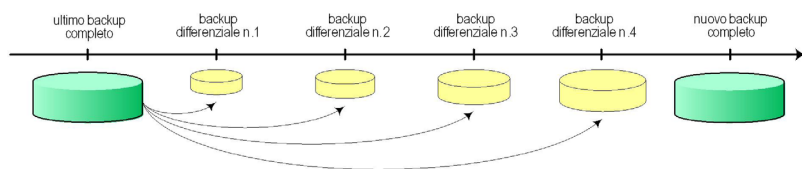


Figura 28. Il *backup* differenziale

consiste nell'effettuare periodiche copie di tutto l'archivio (si veda la figura 27). È molto semplice ma richiede evidentemente grandi capacità di memorizzazione, anche se è quello che richiede il minimo tempo per effettuare il *restore*.

Per ovviare a questo inconveniente sono state proposte altre strategie di *backup*, come quello *differenziale* che consiste nel memorizzare prima un *backup* completo dell'archivio che si vuole salvaguardare e poi successivi *backup* differenziali che contengono solo le modifiche che sono state apportate all'archivio a partire dal momento in cui è stata effettuata l'ultima copia completa (si veda la figura 28). Nel tempo è possibile creare *backup* differenziali successivi, che vengono creati molto velocemente, perché vengono salvate solo le modifiche apportate all'archivio rispetto all'ultima copia completa; anche la quantità di spazio di memorizzazione utilizzato dai *backup* differenziali è significativamente inferiore rispetto a quello dei *backup* completi. Per ripristinare completamente l'archivio sono necessari solo il *backup* completo e l'ultimo *backup* differenziale e non servono i *backup* differenziali intermedi (che potrebbero anche essere cancellati se vi sono esigenze di capacità di memorizzazione). Purtroppo, più lungo è l'intervallo di tempo che intercorre tra il *backup* completo e quello differenziale, maggiore sarà la dimensione del *backup* differenziale e il tempo necessario per crearlo²¹.

²¹ Per ridurre questo tempo, è necessario eseguire occasionalmente un *backup* completo in maniera da diminuire la dimensione delle immagini differenziali successive.

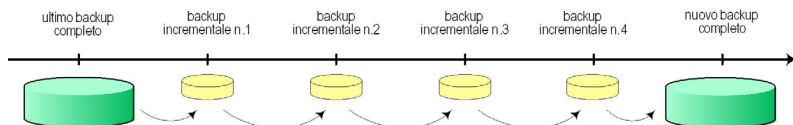


Figura 29. Il *backup* incrementale

Con il *backup incrementale*, invece, dopo aver creato il *backup* completo dell'archivio, vengono memorizzate solo le modifiche intervenute dopo l'ultimo *backup*, sia essa completo o incrementale (si veda la figura 29). Di conseguenza i *backup* incrementali richiedono occupano poco spazio e sono veloci da effettuare, soprattutto se effettuati spesso. Per ripristinare l'archivio occorrono, tuttavia, l'ultimo *backup* completo e tutti i *backup* incrementali; se dovesse mancare anche uno solo di essi non sarebbe possibile ripristinare l'archivio. Nella tabella 22 sono riepilogati i vantaggi e gli svantaggi comunemente attribuiti a ciascun tipo di *backup*.

Tabella 22. Confronto tra le tipologie di *backup*

Parametro	<i>Backup</i> completo	<i>Backup differenziale</i>	<i>Backup incrementale</i>
Spazio di archiviazione	Massimo	Medio	Minimo
Tempo di creazione	Massimo	Medio	Minimo
Tempo di ripristino	Minimo	Medio	Massimo

I sistemi di *backup* e *restore* offrono una buona protezione nei confronti degli incidenti che possono causare la perdita degli archivi, ma presentano una criticità intrinseca che non è in alcun modo superabile: se il *backup*, come spesso avviene, è effettuato ad intervalli di tempo regolari (e non è di tipo sincrono – si veda la nota 18), in caso di guasto non è possibile recuperare ciò che è stato aggiunto all'archivio successivamente all'ultimo *backup* effettuato. È per questo motivo che sono stati introdotti i sistemi RAID (*Redundant Array of Independent Disks*, cioè “insieme

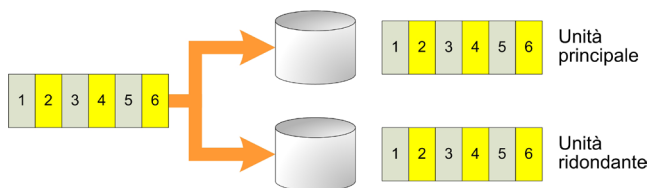


Figura 30. Schema di funzionamento della configurazione RAID 1

ridondante di dischi indipendenti”), ovvero quei sistemi di *storage management* che utilizzano un insieme (*array*) di unità di memoria (ad esempio, dischi fissi) allo scopo di aumentare l’affidabilità, la capacità di memorizzazione o le prestazioni complessive dell’intero sistema²². Vista la fondamentale importanza di queste tecnologie per i processi conservativi digitali, nell’ambito del Master FGCAD l’argomento è trattato in maniera approfondita.

Nella pratica sono state proposte diverse configurazioni RAID. La configurazione RAID 0 punta a massimizzare le prestazioni a scapito dell’affidabilità: le informazioni sono suddivise in blocchi (*strips*, cioè “strisce”) che vengono poi distribuiti parallelamente su due unità di memoria mediante una tecnica denominata *striping*. Le prestazioni migliorano, dal momento che l’accesso in lettura e scrittura avviene parallelamente su più unità e quindi aumenta la velocità di trasferimento dei dati; inoltre, la capacità è pari al doppio di quella dell’unità meno capiente. Purtroppo, però, non c’è alcuna ridondanza: in caso di rottura di un’unità tutti i dati sono perduti. Con il RAID 1 i dati vengono scritti contemporaneamente su entrambe le unità (quella “principale” e quella “ridondante”, denominata *mirror*), conservando di fatto una copia di sicurezza aggiornata in tempo reale (si veda la figura 30).

²² Nella loro implementazione originaria (nella quale l’acronimo era l’abbreviazione di *Redundant Array of Inexpensive Disks*, cioè “insieme ridondante di dischi economici”), il fattore chiave era l’abilità di combinare un certo numero di unità di memoria (ad esempio, dischi fissi) spesso obsolete e a basso costo in modo da rendere il sistema complessivamente migliore in quanto a prestazioni e/o affidabilità e/o capacità di memorizzazione rispetto ad una unità di memoria di ultima generazione.

In questo caso si ha la massima affidabilità, dal momento che in caso di guasto ad una delle due unità i dati restano comunque accessibili; la soluzione è però la più costosa, dal momento che la capacità di memorizzazione complessiva è pari a quella dell'unità con capacità di memorizzazione minore. Il RAID 5 utilizza più unità e, grazie ad un meccanismo denominato dei "blocchi di parità" riesce a sopperire al guasto di una di esse ricostruendo i dati a partire da quelli presenti nelle rimanenti unità.

Ciascuna configurazione presenta vantaggi e svantaggi e la soluzione migliore dipende dagli aspetti che si intende privilegiare: il RAID 0 ha le migliori prestazioni, ma non fornisce alcuna garanzia in termini di affidabilità; il RAID 1 è la configurazione più affidabile, poiché ogni unità può danneggiarsi senza causare la perdita dei dati, ma è anche la più costosa. Un buon compromesso è costituito dalla soluzione RAID 5, ma vi sono anche configurazioni più complesse, chiamate RAID 5e (*extended*), RAID 6 e RAID 6e (*extended*), oppure configurazioni multilivello, come RAID 1+0 o RAID 0+1, che combinano i vantaggi delle configurazioni viste in precedenza cercando di minimizzare gli svantaggi.

L'ultima parte dell'insegnamento di "Supporti di memorizzazione e sistemi di *storage management*" è dedicata alle architetture di *storage management* impiegate per aggregare singole unità di memoria e realizzare infrastrutture in cui non vi siano limiti fisici alla quantità di capacità di memorizzazione disponibile, e che, nel contempo, consentano di centralizzare tali risorse su di un unico dispositivo per ridurre i costi di gestione di un insieme complesso di sistemi differenti. Le architetture più utilizzate sono descritte dagli acronimi DAS, NAS e SAN. Quella presente da più tempo sul mercato è l'architettura DAS (*Direct Attached Storage*) che prevede il collegamento diretto del dispositivo di *storage* con la macchina che necessita di spazio di memorizzazione. L'architettura NAS (*Network Attached Storage*) è la soluzione oggi più economica per garantire risorse di *storage* elevate e condivise tra un numero di macchine che può variare dalle poche unità fino a qualche centinaio. Di fatto, un dispositivo NAS viene collegato in rete in maniera da poter essere direttamente utilizzato da tutte

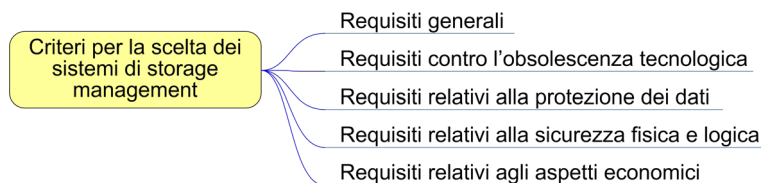


Figura 31. I criteri per la scelta di un sistema di *storage management*

le macchine della rete. Infine, l'architettura SAN (*Storage Area Network*) è in assoluto la più scalabile, ma gli apparati coinvolti sono più complessi che nelle soluzioni DAS e NAS: è costituita da diversi sistemi di *storage management* che vengono collegati a formare una rete locale, a sua volta collegata alla rete locale "principale"; in questo modo le risorse di *storage* sono disponibili per tutte le macchine di quest'ultima rete. La flessibilità dell'architettura e la scalabilità offerta sono gli elementi che fanno delle SAN le architetture di *storage* più usate in ambito aziendale.

A causa dell'inevitabile obsolescenza tecnologica, per i sistemi di *storage management* valgono le stesse considerazioni viste per i supporti di memorizzazione: anch'essi devono essere sottoposti a periodici processi di "aggiornamento tecnologico" (*refreshing*), sia per quanto riguarda i supporti di memorizzazione utilizzati che per quanto riguarda l'infrastruttura *hardware* e *software*. Ecco quindi che la scelta del sistema di *storage management* a cui affidare la conservazione del patrimonio digitale deve essere basata su dei criteri che consentano di individuare il sistema che consenta di massimizzare l'intervallo di tempo tra un'operazione di *refreshing* e l'altra (e, quindi, minimizzare il numero di tali operazioni); semplificare i processi di gestione delle varie attività e fornire le massime garanzie in termini di affidabilità.

Nell'ambito del Master FGCAD, un apposito modulo dell'insegnamento di "Supporti di memorizzazione e sistemi di *storage management*" è dedicato all'individuazione di questi criteri (si veda la figura 31). Innanzitutto devono essere presi in considerazione i seguenti requisiti generali:

- 1) il sistema deve avere una aspettativa di vita²³ ragionevole. Si considera come requisito minimo una durata di almeno 6-8 anni, mentre una longevità maggiore non è necessariamente un vantaggio, dal momento che l'obsolescenza tecnologica costringe comunque ad effettuare operazioni di aggiornamento entro periodi di tempo anche inferiori;
- 2) il sistema deve possedere una adeguata maturità sotto il profilo tecnologico; ciò significa che deve essere fondato su tecnologie mature (ma non obsolete né ancora troppo giovani), ben affermate sul mercato, disponibili presso diversi fornitori di tecnologia (e non presso un solo fornitore) e preferibilmente basate su standard aperti;
- 3) il sistema deve fornire una adeguata capacità di memorizzazione (che deve essere, ovviamente, appropriata alla quantità di dati da memorizzare) e deve essere scalabile, ovvero deve essere possibile aumentare la capacità di memorizzazione aggiungendo ulteriori moduli;
- 4) il sistema deve avere una adeguata capacità di ritenzione dei dati: esso deve prevedere l'utilizzo di supporti di memorizzazione con una capacità di ritenzione garantita nel tempo e possibilmente certificata²⁴;
- 5) il sistema deve basarsi su una componente *software* (sistemi operativi ed altro *software* di base) matura, stabile²⁵ ed ampiamente supportata.

²³ Per "aspettativa di vita" si intende il periodo di tempo per cui si stima, con ragionevole certezza, che il sistema possa essere utilizzato, in relazione alle tecnologie utilizzate e al supporto da parte del produttore o da terze parti. Ciò significa, in particolare, che le tecnologie su cui si basa non devono essere obsolete o in procinto di diventarlo, né d'altra parte, devono essere così "giovani" da presentare ancora incertezze riguardo la loro affidabilità.

²⁴ Non è importante che il sistema garantisca la massima capacità di ritenzione oggi possibile, ma che fornisca una garanzia certa sulla capacità di ritenzione, in maniera tale non si abbiano dubbi sullo stato di conservazione degli archivi digitali.

²⁵ Ciò significa, ad esempio, che devono essere presi in considerazione sistemi operativi per i quali sono stati risolti tutti i *bug* noti tramite il rilascio di *patch*, *hotfix* e *service pack*.

Accanto a questi requisiti generali, devono essere valutati anche gli ulteriori requisiti specifici che mirano a contrastare l'obsolescenza tecnologica o sono relativi alla protezione dei dati, alla sicurezza fisica e logica e agli aspetti economici. È chiaro che gli elementi che occorre prendere in considerazione sono numerosi e si estendono ben oltre i requisiti di ovvia ed immediata comprensione. Una metodologia valida per giungere ad una scelta ponderata consiste nel realizzare, per ogni possibile fornitore del sistema di *storage*, una *balanced scorecard* nella quale vengono valutati i criteri precedentemente descritti e a ciascuno di essi viene assegnato un "peso" ed un punteggio, variabile da un valore minimo (corrispondente a "non rispetta il criterio") ad un valore massimo (corrispondente a "rispetta pienamente il criterio"). Sommando i punteggi raggiunti da ogni fornitore si può ottenere una ragionevole scala di valutazione.

Come risulta evidente, non vi sono soluzioni che si possano considerare a priori migliori o peggiori ma occorre individuare quella più adatta a seconda dello scenario di utilizzo e degli obiettivi che si intende raggiungere. Questa capacità critica di analisi e di scelta è una delle abilità che il Master FGCAD intende formare nei suoi allievi.

3. *Aspetti tecnologici legati ai formati elettronici*

Un altro aspetto tecnologico che deve essere ben tenuto in considerazione da chi si occupa di conservazione di archivi digitali è quello del formato elettronico, ovvero quel complesso di informazioni che consentono ad un'applicazione *software* di riprodurre la sequenza di bit che costituisce un documento digitale restituendolo nella forma e con i contenuti stabiliti dal soggetto produttore. Esempi di formati elettronici sono il formato DOCX di Microsoft Word (dalle versione 2007 in poi), il diffusissimo formato PDF, il formato JPG per le immagini fotografiche, il formato MP3 per le registrazioni audio, e così via. Il problema dei formati elettronici, a volte sottovalutato ma in realtà da tene-

re in debita considerazione, è legato sostanzialmente al fatto che essi diventano rapidamente ed inevitabilmente obsoleti, rendendo gli oggetti digitali codificati secondo quei formati difficilmente fruibili. A titolo di esempio, basti pensare che poco più di una ventina di anni fa, nel periodo a cavallo tra la fine degli anni '80 e l'inizio degli anni '90 del secolo scorso, il *software* di elaborazione testi più utilizzato era *Wordstar*, un programma che utilizzava una interfaccia “a carattere” assai spartana (nulla a che vedere con l'interfaccia grafica dei moderni programmi di elaborazione testi) e che produceva documenti di testo in un formato proprietario identificato dall'estensione *.ws*²⁶. Oggi, chi possedesse ancora dei documenti di testo codificati in quel formato (magari archiviati sui *floppy disk* in uso all'epoca – i “famosi” *floppy disk* da 5 pollici e ¼), incontrerebbe sicuramente grosse difficoltà nel visualizzarne i contenuti, dal momento che il formato di *Wordstar* non è più gestito dagli attuali programmi di elaborazione testi (come *Word* della Microsoft oppure *Writer* di *OpenOffice*) ed è difficile trovare strumenti di conversione che consentano di trasformare quei *file* in formati oggi accessibili²⁷. Quanto visto per *Wordstar* è un discorso che può essere generalizzato a tantissimi altri formati

²⁶ In realtà, le prime versioni di *Wordstar* per DOS non utilizzavano alcuna estensione nella denominazione dei *file* (e quindi, una tesi realizzata con *Wordstar* poteva anche essere denominata semplicemente “tesi”). Solo successivamente è invalsa l'abitudine di inserire una estensione che poteva essere semplicemente *.ws* oppure *.ws* seguita dalla cifra indicante la versione del *software* (ad esempio, *.ws3* nel caso il *file* fosse stato prodotto con *Wordstar 3*) ma anche in questo caso si trattava di una usanza nata presso gli utilizzatori e non obbligatoria. Cfr. Claudio Pozzoli, *Come scrivere una tesi di laurea con il personal computer*, prefazione di Umberto Eco, Milano, Rizzoli, 1986.

²⁷ Si tenga presente il seguente esempio: alla fine degli anni '80 il *software* di videoscrittura più utilizzato era *Wordstar* e quando, agli inizi degli anni '90, venne introdotto Microsoft *Word*, le prime versioni disponevano dei “filtri” di conversione che avrebbero potuto essere utilizzati per effettuare la conversione dal formato di *Wordstar* verso il formato di *Word*. Ma già verso la fine degli anni '90 nelle nuove versioni di Microsoft *Word* questi filtri di conversione erano stati eliminati per cui in soli dieci anni era diventato molto difficile effettuare il riversamento sostitutivo dei documenti di testo creati con *Wordstar* solo pochi anni prima.

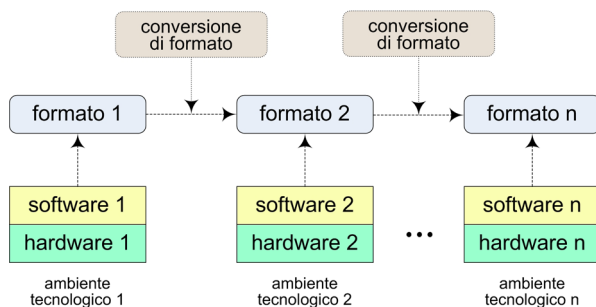


Figura 32. Schema di funzionamento della strategia di conservazione digitale basata sul riversamento sostitutivo

elettronici che nel tempo sono diventati obsoleti; basti pensare ai documenti di testo prodotti con programmi come IBM *Writing Assistant*, o ai fogli di calcolo creati con Lotus 1-2-3 o Quattro Pro, molto utilizzati fino alla fine degli anni '80 ma ormai del tutto obsoleti.

La questione dell'obsolescenza dei formati elettronici è stata ampiamente studiata e sono state individuate alcune soluzioni che sembrano fornire risposte soddisfacenti. Quella che in questo momento sembra la più efficace prende il nome di *riversamento sostitutivo* o *migrazione* (*migration*) e consiste nella conversione di un documento digitale da un formato fruibile in un determinato ambiente tecnologico ad un formato fruibile in un altro ambiente (si veda la figura 32). Normalmente ciò si realizza convertendo i documenti codificati secondo vecchi formati in documenti codificati secondo formati più attuali che possono essere utilizzati su computer più moderni. Così come il riversamento diretto, anche quello sostitutivo non ammette dilazione nel tempo e deve essere eseguito prima che il formato corrente diventi obsoleto; se questa operazione viene trascurata per una sola generazione di formati, i documenti potrebbero essere difficili da recuperare; se vengono trascurate i riversamenti per più generazioni, i documenti potreb-

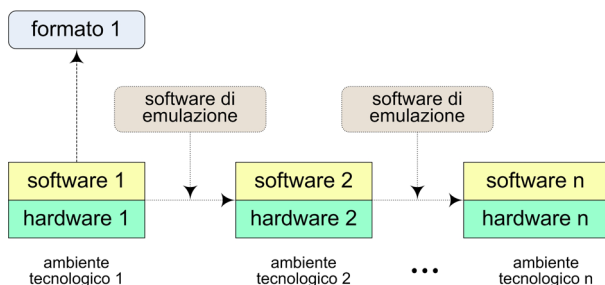


Figura 33. Schema di funzionamento della strategia di conservazione digitale basata sull'emulazione

bero andare completamente perduti²⁸. Dal momento che il riversamento sostitutivo comporta una modifica nella sequenza di bit che costituisce il documento digitale, occorre operare affinché sia mantenuto il contenuto informativo del documento di origine e vengano salvaguardate le sue proprietà significative (*significant properties*), assicurandone nel contempo l'autenticità e l'integrità.

La considerazione del fatto che a seguito dei processi di riversamento sostitutivo possono verificarsi delle perdite nel contenuto informativo – tanto più probabili quanto più il contenuto informativo è complesso – ha condotto alcuni ricercatori a suggerire la strategia di conservazione denominata *emulazione* (*emulation*), che permette di mantenere inalterata la sequenza di bit che costituisce il contenuto digitale, ma è più complessa, dal punto di vista

²⁸ Riprendendo l'esempio precedente, un documento di testo realizzato alla fine degli anni '80 con *Wordstar* potrebbe aver subito un primo riversamento sostitutivo nel formato di *WordPerfect* agli inizi degli anni '90, poi, a metà degli anni '90 altri due riversamenti sostitutivi, prima nel formato di *Microsoft Word 2.0* e poi in quello di *Microsoft Word 6.0/95*; successivamente, alla fine degli anni '90 potrebbe aver subito il riversamento sostitutivo nel formato di *Microsoft Word 97/2003*; infine, nel 2007, un ulteriore riversamento sostitutivo nel formato di *Microsoft Word 2007 (.docx)*, e così via. La serie di riversamenti sostitutivi che occorrerà eseguire su un determinato oggetto digitale per garantirne la conservazione nel tempo non avrà termine fintantoché si vorrà conservare quel determinato documento.

tecnico, rispetto al riversamento diretto e sostitutivo. La teoria che è alla base dell'emulazione si fonda sulla convinzione che l'autenticità di un documento digitale possa essere garantita solo mantenendo quel documento accessibile nella forma e nell'ambiente originali (ovvero con il computer, il sistema operativo e il programma con cui è stato creato o gestito). Tuttavia per ottenere questo scopo occorre ricreare un ambiente tecnologico obsoleto su un nuovo e più moderno ambiente utilizzando degli opportuni *software* di emulazione (si veda la figura 33). A differenza della strategia basata sul riversamento sostitutivo, in questo caso non c'è nessuna conversione di formato e quindi nessun cambiamento nella sequenza di bit che costituisce il *file*. Questo significa che il contenuto potrà essere riprodotto, in linea di principio, sempre nello stesso modo, mantenendo il "comportamento" e l'aspetto originali, indipendentemente dal tempo che è trascorso e dalla piattaforma sulla quale il *software* di emulazione sarà eseguito nel futuro.

Nel corso degli anni sono state sviluppate diverse strategie di emulazione di diversa complessità, tra cui l'emulazione *hardware*, l'emulazione *software* e l'emulazione basata sull'UVC (*Universal Virtual Computer*). Nell'emulazione *hardware* un ambiente *hardware* obsoleto viene ricreato in un computer moderno; a tale scopo viene scritto un programma che imita le strutture del vecchio computer su una nuova piattaforma. In questo modo è possibile operare nel nuovo ambiente come se ci si trovasse nel vecchio: i vecchi *file* sono riprodotti esattamente alla stessa maniera di quando sono stati creati nell'ambiente originale. Il vantaggio principale dell'emulazione *hardware* è costituito dal fatto che il *file* originale non necessita di essere convertito. Come rovescio della medaglia l'emulazione *hardware* non è assolutamente semplice da realizzare e richiede competenze informatiche di alto livello. Nell'emulazione *software* il problema è affrontato in un modo simile ma ad un più alto livello di astrazione: viene ricreato solamente l'ambiente *software* che "girava" sul vecchio computer e non l'ambiente *hardware* completo. Grazie a questo *software* che "emula" l'ambiente *software* originale è possibile

avere accesso ai contenuti digitali creati per quel determinato ambiente anche su un computer moderno²⁹. Come nell'emulazione *hardware*, anche nell'emulazione *software* il principale vantaggio è costituito dal fatto che l'oggetto digitale non necessita di alcuna conversione e rimane nel tempo esattamente identico all'originale, mantenendo anche il *look and feel*. Lo svantaggio è, anche in questo caso, costituito dal fatto che l'emulazione è una strategia di conservazione piuttosto laboriosa e quindi costosa. L'*emulazione basata sull'UVC (Universal Virtual Computer)* costituisce un'ulteriore variante: anche in questo caso deve essere scritto un emulatore *software*, ma non per un computer reale bensì per un computer virtuale, non esistente, denominato appunto *Universal Virtual Computer (UVC)*³⁰.

L'emulazione è una strada che può nel tempo diventare molto complessa, sia per la sfida tecnologica che sottintende, sia per gli elevati costi connessi con la realizzazione dei *software* di emulazione.

Il problema dell'obsolescenza dei formati elettronici è amplificato dal fatto che ne esistono diverse migliaia³¹ e, almeno in teoria, per ciascuno di essi occorrerebbe prevedere delle procedu-

²⁹ Per fare un esempio, in un computer con un sistema operativo come Linux può "girare" un programma progettato per emulare un altro sistema operativo, per esempio Windows. Le applicazioni che "girano" sotto Windows possono essere così utilizzate attraverso l'emulatore Windows anche su un computer che ha come sistema operativo Linux.

³⁰ L'UVC è un computer con una architettura ed un insieme di istruzioni talmente semplici che ogni sviluppatore di *software* nel futuro sarà in grado di scrivere un emulatore per esso. Gli oggetti digitali debbono essere convertiti attraverso un programma scritto nel linguaggio di programmazione dell'UVC, in un formato denominato *Logical Data Description (LDD)*. Il *Logical Data Description* è un formato indipendente, auto-descrittivo e chiaramente strutturato, che contiene tutte le informazioni che permetteranno di "riassemblare" l'oggetto digitale. Infatti tali informazioni sono realizzate in maniera tale che, nel futuro, sarà possibile realizzare un visualizzatore (*viewer*) che interpreterà il formato LDD e visualizzerà a schermo l'oggetto digitale originale. Come si può notare, si tratta di una strategia di conservazione che si basa sull'emulazione ma contiene anche aspetti che riconducono alla migrazione.

³¹ Secondo un studio condotto nell'ambito del progetto DELOS nel 2004, i formati elettronici allora esistenti erano oltre 15.000. È presumibile che nel momento attuale i formati elettronici esistenti od esistiti siano più di ventimila. Si veda DELOS, *File Formats Typology and Registries for Digital Preservation*, Urbino, Istituto di studi per la tutela dei beni archivistici e librari (ISTBAL), 2004.



Figura 34. Classificazione dei requisiti dei formati elettronici

re per il riversamento sostitutivo. Ecco perché è molto importante limitare al massimo il numero dei formati elettronici da gestire (e quindi, ad esempio, produrre *policy* che stabiliscano chiaramente quali sono i formati accettabili). Inoltre, è importante adottare fin dal momento della formazione di un archivio digitale, formati elettronici che forniscano le maggiori garanzie in termini di conservazione a lungo termine; ecco, quindi, che risulta fondamentale individuare i criteri che devono guidare nella scelta di un formato compatibile con un processo di conservazione digitale. Uno studio approfondito su queste tematiche è stato condotto presso il Laboratorio di Informatica Documentale dell'Università degli Studi di Macerata ed i risultati di questo studio sono confluiti nell'insegnamento di "Formati elettronici, linguaggio XML e modulistica elettronica" del Master FGCAD. Nell'ambito di questo studio sono stati identificati i requisiti che un formato deve possedere per essere considerato compatibile con un processo di conservazione digitale³². Tali requisiti (si veda la figura 34)

³² I risultati di questo studio sono stati pubblicati nel volume Stefano Pigliapoco, Stefano Allegrezza, *Produzione e conservazione del documento digitale. Requisiti e standard per i formati elettronici*, Macerata, eum, 2008, vol. I.

sono stati classificati in *generali*, ovvero applicabili a qualsiasi tipologia di formato, e *specifici*, ovvero relativi ad una particolare categoria di formati (ad esempio, i formati immagine, i formati video, i formati audio, etc.). I requisiti generali a loro volta sono classificabili in requisiti di *primo livello* (che consentono una prima selezione dei formati sulla base delle caratteristiche di non proprietà, apertura, standardizzazione e trasparenza) e requisiti di *secondo livello* (che vanno ad analizzare altre caratteristiche non meno importanti, quali la robustezza, la stabilità, l'auto-contenimento, etc.).

Non è certamente questa la sede adatta per una trattazione completa ed approfondita dei vari requisiti; è importante, tuttavia, fornire alcuni criteri operativi che possano indirizzare verso una corretta scelta dei formati compatibili con un processo di conservazione digitale. Uno dei principali requisiti generali è quello della *non proprietà*: per garantire le migliori condizioni ai fini della conservazione nel tempo è preferibile utilizzare formati che siano non proprietari, ovvero non legati all'esistenza di una specifica azienda che ne detiene i diritti di proprietà intellettuale e che potrebbe in qualsiasi momento decidere di modificarne le specifiche, renderle inaccessibili, o imporre restrizioni sul loro utilizzo. Ad esempio, sono non proprietari i formati dei documenti prodotti con le *suite* d'automazione d'ufficio Microsoft Office dalla versione 2007 in poi, mentre sono proprietari (di proprietà Microsoft) i ben noti formati DOC, XLS e PPT della *suite* Microsoft Office fino alla versione 2003.

Altro requisito fondamentale è quello dell'*apertura*: un formato adatto per la produzione di documenti informatici compatibili con un processo di conservazione digitale deve essere aperto, ovvero le sue specifiche devono essere pubbliche e liberamente accessibili. Tuttavia la sola apertura non è un criterio sufficiente perché è necessario anche quello della *completa documentazione*: occorre, cioè, che le specifiche documentino in maniera completa ed esaustiva il formato. Un terzo requisito da tenere in considerazione è quello della *standardizzazione*. Un formato è standard quando le sue specifiche sono definite o approvate da un organi-

smo di standardizzazione e quindi ha ottenuto un riconoscimento ufficiale (in questo caso si parla di standard *de jure*); oppure quando le sue specifiche non sono state ratificate da nessun organismo di normazione, ma è diventato, *di fatto*, uno standard grazie alla sua ampia diffusione (in questo caso si quando di standard *de facto*)³³. Ai fini della conservazione digitale è importante scegliere formati standard e possibilmente *de jure* in quanto meno soggetti ad obsolescenza. Infine, è molto importante il requisito della *trasparenza*: un *file* è codificato secondo un formato trasparente se ne è possibile la lettura diretta adoperando semplici strumenti *software* di base³⁴. Più un formato è trasparente e maggiori sono le garanzie fornite in merito alla sua leggibilità nel futuro.

Una prima risposta concreta al problema dell'obsolescenza dei formati è costituita dalla standardizzazione del PDF/A³⁵, un formato appositamente creato per la conservazione a lungo termine dei documenti con prevalente contenuto testuale. Si tratta di una versione "limitata" del formato PDF³⁶ e di questo utiliz-

³³ Esempi di standard *de facto* sono il DOC, l'XLS, l'RTF; esempi di standard *de jure* sono l'HTML (W3C Recommendation e ISO 15445), l'XML (W3C Recommendation), l'ODF (ISO 26300), il TIFF (ISO 12639), il JPG (ISO 10918-1), il JPEG2000 (ISO 15444), il PNG (ISO 15948), il PDF (ISO 32000) e alcuni suoi *subset* quali il PDF/A, il PDF/E e il PDF/X.

³⁴ In questo senso, ad esempio, i documenti di testo prodotti nel formato TXT sono più trasparenti di quelli prodotti nel formato DOC, dato che i primi sono fruibili utilizzando un semplice *text editor* (come Blocco Note di Windows), mentre i secondi necessitano di uno strumento più complesso quale un *word processor*.

³⁵ Il PDF/A è un formato non proprietario, aperto e standard *de jure*, essendo stato riconosciuto standard ISO il 28 settembre 2005 con la denominazione ISO 19005-1:2005 *Document Management - Electronic Document File Format for Long-Term Preservation - Part 1: Use of PDF (PDF/A1)*. La "/A" contenuta nel nome del formato, sebbene non venga formalmente definita in alcun documento tecnico, fa evidentemente riferimento ai termini inglesi "archiving" o "archive".

³⁶ Il formato PDF, universalmente adottato e divenuto ormai lo standard *de facto* per la presentazione e divulgazione dei contenuti digitali a prevalente contenuto testuale, non è adatto, nella sua forma standard, alla conservazione dei documenti digitali poiché non è in grado di garantire la riproducibilità a lungo termine né la conservazione dell'aspetto visivo. Infatti, i *file* PDF standard non sono necessariamente auto-contenuti (possono presentare dipendenze dai font utilizzati o da oggetti esterni al *file* stesso). Di conseguenza, se si cerca di visualizzare questi *file* su sistemi diversi da quelli in cui sono stati creati, non è garantito che vengano riprodotti correttamente.

za solamente quelle caratteristiche che permettono di produrre *file* in grado di conservare nel tempo la loro rappresentazione (sia su schermo che a stampa o attraverso altri sistemi di *output*) indipendentemente dagli strumenti tecnologici utilizzati. Ad esempio, il PDF/A richiede che i *font* utilizzati nel *file* siano “incorporati” al suo interno per avere la certezza della loro disponibilità in fase di rappresentazione; inoltre il *file* deve essere auto-contenuto e per questo motivo è vietato l’uso di *link* esterni. Per le sue caratteristiche deve essere preso seriamente in considerazione ai fini della sua adozione come formato per la conservazione dei documenti digitali. Data la sua rilevanza, nell’ambito del Master FGCAD al formato PDF/A è dedicato un apposito modulo che tratta approfonditamente le tre versioni del formato che sono state fino ad oggi riconosciute standard ISO, il PDF/A-1, il PDF/A-2 e il PDF/A-3.

Un’altra strada che si sta sperimentando e che sembra piuttosto promettente, è quella del formato “universale”, ossia un unico formato elettronico capace di veicolare qualsiasi tipo di contenuto, dai documenti di testo, alle immagini, agli elaborati tecnici, fino alle registrazioni audio e video. È evidente che se fosse disponibile un simile formato, le operazioni di migrazione che nel tempo si renderebbero necessarie sarebbero molto semplificate perché si tratterebbe di operare solamente su un unico formato elettronico anziché su migliaia di formati. Su questo versante vi sono già alcune proposte che sembrano essere in grado di fornire soluzioni interessanti, ma la situazione non è ancora sufficientemente definita per poter dare un giudizio sulla loro validità; occorre, pertanto, attendere ulteriori sviluppi.

4. Aspetti tecnologici nella digitalizzazione dei documenti analogici

Nell’ambito del Master FGCAD viene dato ampio spazio alle tematiche inerenti la digitalizzazione dei documenti analogici, dal momento che molti archivi digitali sono il risultato di operazio-

ni di digitalizzazione effettuate, spesso su larga scala, su archivi analogici. Poiché vi sono prospettive lavorative anche molto interessanti in questo settore, queste tematiche vengono trattate in un apposito insegnamento di “Tecniche per la digitalizzazione dei documenti analogici”, con un approccio non solo teorico ma anche pratico: vengono, infatti, proposti esempi concreti di attività di digitalizzazione sia di archivi cartacei che di archivi sonori ed audiovisivi.

Anche in questo caso la componente tecnologica è fondamentale: con il termine “digitalizzazione” si intende, infatti, il processo che, mediante opportuni apparati tecnologici, consente la trasformazione di un documento, di una fotografia, di una registrazione sonora o audiovisiva o di qualsiasi altro oggetto analogico in un oggetto digitale – ovvero una sequenza di “0” e “1” – interpretabile da un computer. È chiaro, quindi, che la componente tecnologica riveste un ruolo di primo piano in queste attività, consentendo di ottenere risultati più o meno validi a seconda della qualità degli apparati di digitalizzazione messi in campo.

Le motivazioni alla base di un’attività di digitalizzazione possono essere molteplici. Quella che storicamente ha avuto forse maggior peso e che ha spinto gli archivi, le biblioteche, le cineteche, i musei ed altre istituzioni (culturali e non: si pensi, ad esempio, al progetto Google Books) ad avviare numerosi progetti di digitalizzazione del patrimonio analogico è la finalità di *conservazione*, che consente di conseguire almeno due risultati. Innanzitutto, la digitalizzazione sottrae gli originali al deterioramento causato dalla consultazione da parte degli utenti; inoltre, la messa a disposizione dell’utenza di copie digitali consente di mantenere permanentemente gli originali (ad esempio, cartacei e pergamenei) in ambienti che ne consentono le migliori condizioni di conservazione in termini di temperatura, umidità relativa, irraggiamento, etc. In secondo luogo, alcune tipologie documentarie vanno incontro inesorabilmente ad un degrado anche molto veloce (si pensi alle registrazioni audio su cassetta o alle registrazioni video su VHS, i cui nastri sono soggetti ad un degrado naturale

che aumenta con il trascorrere del tempo fino a diventare “illeggibili”); in questi casi la digitalizzazione consente la produzione di una copia che mantiene inalterato il valore informativo dell’originale e che può essere conservata in sostituzione degli originali³⁷. Una seconda finalità è quella di *valorizzazione*: la digitalizzazione degli originali analogici permette una maggiore diffusione dei materiali, tramite, ad esempio, le possibilità offerte dalla messa a disposizione on line degli archivi digitalizzati; in questo modo è possibile raggiungere un pubblico più vasto e fornire l’accesso a quantità elevate di documenti, anche rari o inavvicinabili; gli utenti, da parte loro, hanno la possibilità di raggiungere risorse presenti in luoghi diversi ed anche fisicamente lontane. Vi sono, infine, motivazioni di carattere economico: la gestione di archivi digitali può comportare un risparmio in termini economici, ad esempio per la riduzione dei costi necessari per la gestione degli archivi cartacei (si pensi al mantenimento di condizioni climatiche ed ambientali controllate, agli adempimenti necessari per la messa a norma dei locali, per l’installazione e manutenzione di impianti di spegnimento incendi, per l’ottenimento del Certificato Prevenzione Incendi, etc.), alla riduzione degli spazi necessari per lo stoccaggio dei materiali, alla diminuzione del tempo necessario per la ricerca dei documenti, e così via.

Nella convinzione che sia necessario fornire agli allievi gli strumenti idonei per poter operare in questo settore, nell’ambito dell’insegnamento già citato vengono proposte esercitazioni pratiche sulla gestione di progetti di digitalizzazione simulando situazioni concrete che possono verificarsi nella vita lavorativa. Ad esempio, vengono proposti progetti di digitalizzazione di volumi antichi, di fogli di mappa catastali, di microfiche contenenti schede tecniche particolareggiate, di diapositive e negativi fotografici, di audiocassette (le cosiddette “compact cassette”) contenenti le registrazioni audio di sedute del consiglio comuna-

³⁷ A questo proposito è opportuno sottolineare che la convinzione che “il digitale si conserva meglio dell’analogico”, che fino alla fine degli anni ’80 del secolo scorso veniva data come un’certezza, oggi è stata ampiamente messa in discussione, soprattutto in considerazione delle difficoltà insite nella conservazione dei documenti digitali.

le, di bobine contenenti le registrazioni audio di concerti di musica classica, di registrazioni audio su dischi in vinile, di registrazioni video su videocassette VHS e di pellicole cinematografiche in formato Super 8 mm. Per ciascun progetto vengono affrontati vari aspetti: oltre alle motivazioni in base alle quali si propone l'avvio dell'attività di digitalizzazione, occorre prendere in considerazione tutta una serie di aspetti tecnologici, tra cui le tecnologie utilizzate per la conversione da analogico a digitale, i formati elettronici scelti con l'indicazione dei criteri adottati per la scelta, i parametri tecnici (come la risoluzione e la profondità di colore nel caso delle immagini), la stima della capacità di memorizzazione necessaria, la descrizione dei supporti di memorizzazione e dei formati obsoleti attualmente utilizzati, i supporti di memorizzazione scelti per il riversamento e le motivazioni per la scelta, i metadati che si intende utilizzare per la descrizione degli oggetti digitali, la stima dei tempi necessari per l'esecuzione del lavoro e dei relativi costi. Alla conclusione dell'insegnamento, gli studenti hanno acquisito le conoscenze teoriche e le abilità necessarie per gestire progetti di digitalizzazione nelle varie situazioni che possono presentarsi nella vita lavorativa.

Conclusioni

Le problematiche connesse con la conservazione degli archivi digitali coinvolgono ambiti scientifici e settori di ricerca diversi e sono estremamente complesse da risolvere; è, ormai, ampiamente condivisa la convinzione che i contenuti digitali non si conservano “da soli” e che, per evitare la loro perdita, è necessario individuare al più presto delle soluzioni concretamente realizzabili³⁸. Allo stato attuale non si è ancora raggiunta una unanimità di vedute su quali, tra le varie strategie di conservazione digitale, siano quelle più affidabili; tuttavia è possibile affermare che quel-

³⁸ «Digital resources will not survive or remain accessible by accident». Cfr. Bernard Smith, *Preserving Tomorrow's Memory: Preserving Digital Content for Future Generations*, «Information Services and Use», 22, 2-3, 2002, pp. 133-139.

le che sembrano fornire, ad oggi, le maggiori garanzie – il riversamento, sia diretto che sostitutivo, e l'emulazione – la componente tecnologica gioca un ruolo determinante. Infatti, come si è avuto modo di mettere in rilievo nel corso della trattazione, solo una conoscenza approfondita degli aspetti tecnologici può guidare la ricerca delle soluzioni e suggerire le decisioni da prendere per assicurare un futuro alla memoria digitale che si sta formando in questi anni³⁹.

³⁹ Ecco perché a livello internazionale diventa sempre più richiesta una nuova figura professionale, quella del *digital curator*, il “conservatore digitale”, una figura che sia in grado di coniugare competenze archivistiche, informatiche, giuridiche, economiche, organizzative per affrontare, in una logica di gruppo, la complessa sfida della conservazione digitale. Si veda Chiara Cirinnà, *Formare le professionalità del futuro prossimo: arrivano i Digital Curators*, «La Rivista di FormareNetwork», Fondazione Rinascimento Digitale, <<http://e-learning.dti.unimi.it/Portale/rivista/?p=642>>.

Formazione, gestione e conservazione degli archivi digitali

Il volume affronta il tema della formazione di figure professionali dotate delle conoscenze, competenze e abilità necessarie per la tenuta degli archivi digitali, presentando il Master di I livello in «Formazione, gestione e conservazione di archivi digitali in ambito pubblico e privato (FGCAD)» dell'Università degli Studi di Macerata come un'esperienza concreta, pluriennale, che sta facendo registrare una partecipazione e un consenso sempre crescenti.

Le considerazioni di natura teorica-concettuale sugli elementi che differenziano gli archivi digitali da quelli tradizionali cartacei, presentate dagli autori nei diversi capitoli del libro, dimostrano la necessità di una preparazione basata su un percorso di studi di tipo multi-disciplinare come quello del Master FGCAD, il cui piano didattico comprende oltre agli insegnamenti dell'archivistica e della diplomatica del documento contemporaneo anche quelli del diritto, dell'informatica, della gestione documentale, dell'archiviazione e conservazione digitale.



eum edizioni università di macerata

Stefano Pigliapoco è professore di Informatica documentale e multimediale (M-STO/o8) presso il Dipartimento di Studi Umanistici dell'Università degli Studi di Macerata e Direttore del Master di I livello in «Formazione, gestione e conservazione di archivi digitali in ambito pubblico e privato (FGCAD)». La sua attività di studio e ricerca riguarda principalmente l'archivistica informatica, il *records management*, i sistemi ERMS, l'archiviazione e la conservazione digitale.

Giorgetta Bonfiglio Dosio, già professore ordinario di Archivistica (M-STO/o8) nella Facoltà di lettere e filosofia dell'Università degli Studi di Padova, ha al suo attivo più di 250 pubblicazioni tra monografie e articoli riguardanti anche studi e ricerche su metodologie per la organizzazione di servizi archivistici complessi e la reingegnerizzazione di sistemi di gestione documentale in ambiente digitale e ibrido.

ISBN 978-88-6056-437-5



9 788860 564375

€ 16,00