HANNO COLLABORATO A QUESTO NUMERO:

Stefano ALLEGREZZA, Università degli studi di Macerata
Stephan HEUSCHER, Bedag Informatik AG, Svizzera
Stefano PALAGIANO, Università degli studi di Urbino
Gianfranco PONTEVOLPE, Cnipa
Silvio SALZA, Università degli studi di Roma

La redazione invita tutti coloro che hanno informazioni, opinioni, domande da porre su temi che riguardano l'automazione degli archivi a segnalarle alla Segreteria di Archivi & Computer, presso l'Archivio Storico Comunale di San Miniato (Loggiati di San Domenico, 3 – 56027 San Miniato – Pisa; tel. 0571 418381; telefax 0571 406233; e-mail archilab@comune.san-miniato.pi.it).

ISSN 1121-2462

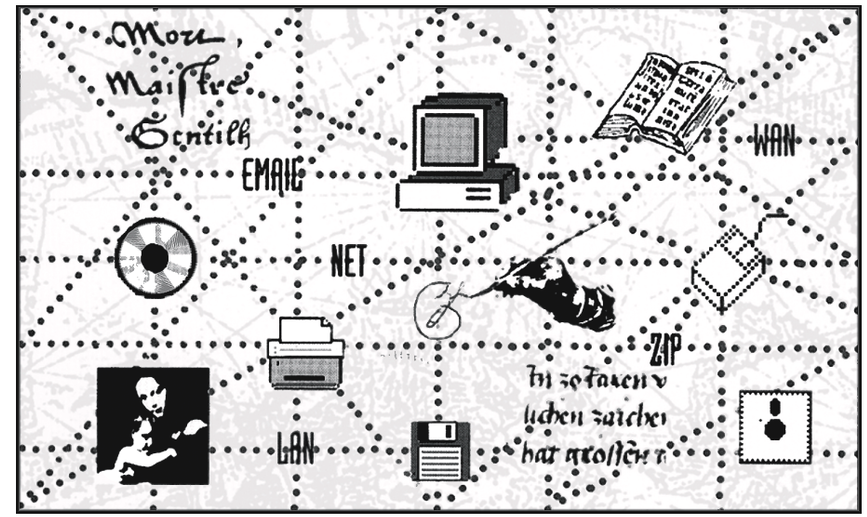
€ 15,00

ARCHIVI & COMPUTER

Semestrale – Spedizione in A.P. – Filiale di Pisa – comma 34, art. 2, L. 549/95 – Pisa

1/2010

ARCHIVI & COMPUTER
AUTOMAZIONE E BENI CULTURALI



Anno XX

Titivillus

Fascicolo 1/010

Stefano ALLEGREZZA

Il formato PDF/A per la conservazione a lungo termine dei documenti

Abstract: *The article analyses the PDF/A format with the specific aim of identifying its strenght and its weakness for digital preservation. The author described the PDF/A standardization process, the levels of compliance, the file creation and validation and its relevance for the requirements in the field of digital preservation.*

1.1. Introduzione

Del formato PDF/A si sente parlare sempre più spesso, soprattutto nel campo della gestione documentale e della conservazione digitale. Le pubbliche amministrazioni richiedono sempre più spesso che i documenti ricevuti siano codificati secondo questo formato.

Ad esempio, in Italia il decreto del Presidente del Consiglio dei ministri del 10 dicembre 2008[1] ha stabilito che tutti gli atti da iscrivere nel Registro delle Imprese debbano essere obbligatoriamente codificati nel formato PDF/A-1.

In Francia, all'inizio del 2009 la Direction générale de la modernisation de l'Etat, Ministère du budget[2] ha pubblicato la raccomandazione *Référentiel général d'interopérabilité* (RGI)[3] per la gestione dei dati elettronici nella quale si consiglia di utilizzare lo standard PDF/A per l'archiviazione dei documenti amministrativi con contenuti statici e inalterabili[4].

In Germania, il Koordinierungs und Beratungsstelle der Bundesregierung für

[1] Decreto del Presidente del Consiglio dei ministri del 10 dicembre 2008 riguardante "Specifiche tecniche del formato elettronico elaborabile (XBRL) per la presentazione dei bilanci di esercizio e consolidati e di altri atti al registro delle imprese", pubblicato sulla «Gazzetta Ufficiale» n. 304 del 31/12/2008 ed entrato in vigore a partire dal 15/01/2009 ed immediatamente applicabile per quanto riguarda l'art. 6, comma 3 che recita: "Nelle more della definizione delle specifiche di cui al comma 1, l'interessato allega all'istanza di cui all'art. 4 un documento informatico in formato PDF/A con il contenuto dell'atto, anche senza immagini ottenute dalla scansione di documenti cartacei."

[2] Il sito web è disponibile all'indirizzo <http://www.modernisation.gouv.fr>.

[3] La versione 1.0 della raccomandazione è disponibile all'indirizzo <http://www.references.modernisation.gouv.fr/rgi-interoperabilite>. Le informazioni sul PDF/A si trovano a p. 63.

[4] Si veda all'indirizzo <http://references.modernisation.gouv.fr/rgi-interoperabilite>.

Informationstechnik in der Bundesverwaltung (KBSt)[5] ha recentemente pubblicato la versione aggiornata del documento *Standards und Architekturen für E-Government-Anwendungen* (SAGA)[6], che descrive gli standard, le tecnologie e i metodi per l'utilizzo dell'informatica da parte delle autorità federali e fornisce le raccomandazioni per l'e-government nel settore pubblico. La nuova versione pone una maggiore enfasi sul PDF/A-1 e consiglia espressamente questo standard.

In Austria, la Bundeskammer der Architekten und Ingenieurkonsulenten (BAIK)[7] richiede che i documenti che vengono resi pubblicamente disponibili siano conformi allo standard PDF/A-1b. Inoltre, utilizza il PDF/A come formato standard per la raccolta di documenti in relazione ai registri immobiliari, richiedendo l'utilizzo di una firma digitale qualificata per garantire l'autenticità di tali documenti elettronici.

In Danimarca, una decisione del parlamento danese ha stabilito che, a partire dal 1° aprile 2011, le autorità governative dovranno utilizzare il formato standard aperto ODF[8] mentre per i documenti non modificabili è previsto l'utilizzo del formato PDF/A-1.

In Norvegia, le amministrazioni pubbliche sono ormai obbligate ad utilizzare i formati aperti, tra cui il PDF/A. Infatti, il governo norvegese, con un regolamento entrato in vigore il 1° gennaio 2009, ha stabilito che tutte le informazioni pubblicate sui siti *web* statali debbano essere codificate secondo formati aperti come l'HTML, il PDF, il PDF/A e l'ODF. In particolare, i formati PDF e PDF/A sono consigliati per tutti i documenti per i quali è importante la salvaguardia dell'aspetto originale[9]. Inoltre, dal 1° gennaio 2010, anche le amministrazioni comu-

[5] L'Ente di coordinamento e consulenza informatica dell'amministrazione federale tedesca; si tratta dell'equivalente, in Italia, dell'ex Centro nazionale per l'informatica nella pubblica amministrazione (CNIPA), che ha assunto ora la denominazione "DigitPA" ai sensi del Decreto legislativo 1° dicembre 2009, n. 177, recante "Riorganizzazione del Centro nazionale per l'informatica nella pubblica amministrazione, a norma dell'art. 24 della legge 18 giugno 2009, n. 69".

[6] "Standard ed architettura per le applicazioni di e-Government". Il documento, nella versione 4.0, è disponibile in lingua tedesca all'indirizzo <http://http://www.cio.bund.de/DE/Standards/ SAGA/saga_node.html>.

[7] La Camera federale degli architetti e consulenti in ingegneria. Ulteriori informazioni sono disponibili sul sito web <https://www.baik-archiv.at>.

[8] L'OpenDocument Format (ODF), abbreviazione di *Open Document Format for Office Applications*, è un formato per la produzione e lo scambio dei *file* prodotti con le *suite* di *office automation* (documenti di testo, fogli elettronici, presentazioni, grafici, etc.). Si tratta di un formato non proprietario, aperto, standard *de jure*, di tipo non binario, basato sull'XML, indipendente dalla piattaforma e dall'applicazione. Per le sue caratteristiche può essere, a ragione, considerato uno standard aperto (*open standard*). Per ulteriori informazioni si rimanda al sito della *OpenDocument Format Alliance* disponibile all'indirizzo <http://www.odfalliance.org>.

[9] Ulteriori informazioni si trovano nell'annuncio del Ministero dell'amministrazione governativa e delle riforme del 21 dicembre 2007, disponibile all'indirizzo <http://www.regjeringen.no/en/dep/fad/pressesenter/pressemeldinger/2007/ Open-document-standards-to-be-obligatory.html?id=494810>.

nali hanno l'obbligo di utilizzare i formati aperti per la pubblicazione di informazioni sul *web*[10].

In Svizzera, il Consiglio federale ha stabilito[11] che le comunicazioni elettroniche tra la pubblica amministrazione e i cittadini debbano avvenire utilizzando il formato PDF/A per la sua idoneità alla conservazione nel lungo periodo. La direttiva è entrata in vigore il 1° gennaio 2008 e la sua completa implementazione dovrebbe avvenire entro dieci anni.

Nel settore archivistico e biblioteconomico, sono già diversi i Paesi che hanno ufficialmente riconosciuto la validità del formato. Solo per fare alcuni esempi, la Biblioteca nazionale tedesca ha indicato il PDF/A come formato per la trasmissione di contenuti in formato digitale, preferendolo ad altri formati, come il PDF "tradizionale", l'HTML e i formati basati sull'XML (si veda la Figura 1)[12]. Sempre in Germania, l'Università di Potsdam utilizza il formato PDF/A per la conservazione dei documenti; in una comunicazione presente sul suo sito *web*[13] si legge che l'obiettivo dell'Università è di avere, nel prossimo futuro, tutti i documenti pubblicati sui suoi server in formato PDF/A. Anche la Biblioteca nazionale austriaca preferisce il PDF/A: nell'apposita sezione del suo sito *web* dedicata ai formato accettati[14] viene precisato che il formato preferito per ricevere *file* è il PDF/A (si veda la Figura 2).

[10] Ulteriori informazioni sono disponibili, in lingua norvegese, all'indirizzo <http://avisenagder.no/Nyheter/tabid/250/ Default.aspx?ModuleId=47196&articleView=true)>.

[11] Si veda il regolamento sulla trasmissione elettronica negli atti amministrativi *Verordnung über die elektronische Übermittlung im Rahmen eines Verwaltungsverfahrens* disponibile all'indirizzo <http://www.admin.ch/ch/d/as/2007/ 5093. pdf>.

[12] Si veda l'elenco completo dei formati accettati per la presentazione dei documenti, disponibile sul sito *web* della Biblioteca nazionale tedesca all'indirizzo <http://www.d-nb.de/eng/netzpub/ablief/np_dateiformate.htm>.

[13] Si veda all'indirizzo <http://opus.kobv.de/ubp/doku/formate.php>. Sempre sul sito è possibile trovare consigli sulle modalità per produrre documenti PDF/A dai formati originali, per esempio Word, e su come effettuare la conversione da altri formati molti utilizzati in ambito scientifico (come LaTeX) verso il PDF/A.

[14] Si veda la pagina web *Informationen für AnbieterInnen und Bibliotheken* (Informazioni per i fornitori e le biblioteche) all'indirizzo <http://www.onb.ac.at/bibliothek/digitale_medien_informationen.htm>.

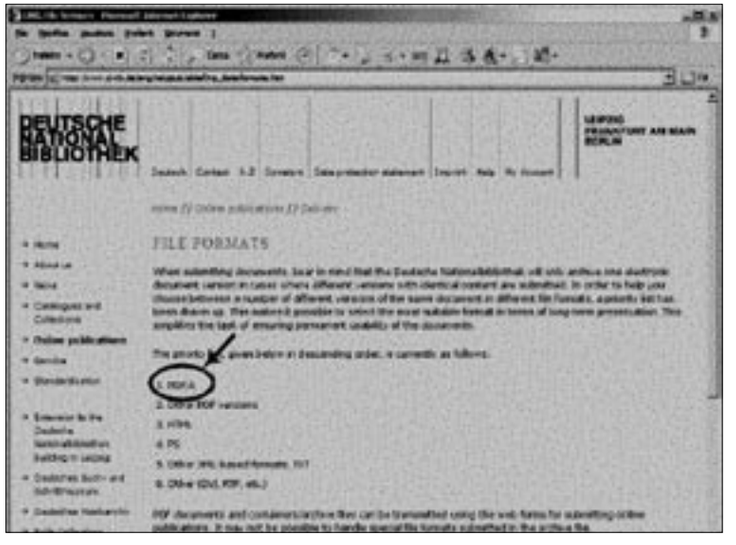


Figura 1. La sezione del sito della Biblioteca nazionale tedesca in cui sono specificati i formati di *file* accettati

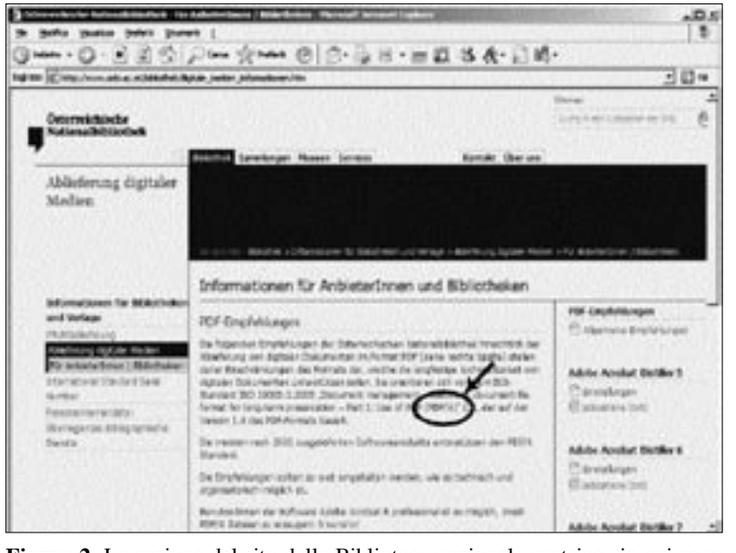


Figura 2. La sezione del sito della Biblioteca nazionale austriaca in cui sono specificati i formati di *file* accettati

Tra i numerosi altri esempi che si potrebbero citare vi sono quelli relativi al United States National Archives and Records Administration (NARA), agli Swedish National Archives, a DeepBlue (il centro di conservazione istituzionale della University of Michigan[15]), allo stesso International Organization for Standardization (ISO), al Florida Digital Archive[16]. Anche MoReq2, l'ultima versione delle specifiche europee MoReq[17], include il PDF/A nell'elenco dei formati elettronici consigliati, ad esempio, per la conservazione dei documenti acquisiti mediante scansione e per l'archiviazione a lungo termine.

Tutti questi esempi rivelano, indiscutibilmente, un interesse verso questo formato che è cresciuto in maniera esponenziale negli ultimi anni, in special modo dal 2005 in poi[18].

Ma che cosa è esattamente questo formato PDF/A? Che relazione c'è tra il formato PDF/A e il formato PDF? Perché è riconosciuto come formato idoneo per la conservazione nel lungo periodo dei documenti elettronici? Quali sono le proprietà che garantiscono che il documento sia visualizzabile sempre allo stesso modo, anche a distanza di tempo e indipendentemente dal programma utilizzato per visualizzare i documenti codificati secondo tale formato? Per fornire una risposta a tutte queste domande è necessario entrare nel dettaglio delle caratteristiche tecniche del formato ed è bene partire dal suo "progenitore", il formato PDF.

1.2. PDF e PDF/A

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 e neanche la conservazione dell'aspetto visivo[19]. Questo è in sostanza da porre in relazione con la natura estremamente ricca di funzionalità del PDF, che consente ai *file* codificati secondo questo formato una variabilità molto ampia nella struttura interna, oltre alla possibilità di essere "composti" in maniera dinamica, nel momento in cui

[15] Si veda il documento *Best practices for producing high quality PDF files* a cura del Formats Group, Deep Blue, disponibile all'indirizzo <http://hdl.handle.net/2027.42/58005>.

[16] Si veda il documento *Guidelines for Creating Archival Quality PDF Files*, disponibile all'indirizzo <http://www.fcla.edu/digitalArchive/pdfs/PDFGuideline.pdf>.

[17] MoReq2, abbreviazione di "Model Requirements for the Management of Electronic Documents and Records" è una specifica europea per la gestione dei documenti e degli archivi elettronici. MoReq, la prima versione, era stato originariamente sviluppato per lo scambio standardizzato di documenti tra la Commissione europea e i governi degli stati membri. Si veda all'indirizzo <http://www.moreq2.eu>.

[18] Come si vedrà più avanti, il formato PDF/A è stato riconosciuto come standard proprio nel 2005.

[19] Ciò dipende da molteplici motivi come, ad esempio, il fatto che 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.

vengono visualizzati, consentendo l'utilizzo di *font* esterni alquanto diversificati. Tutto ciò conduce, come rovescio della medaglia, all'impossibilità di assicurare una loro riproduzione sempre uguale, indipendentemente dall'ambiente tecnologico in cui essa avviene.

Per risolvere queste ed altre difficoltà, è stato necessario giungere alla definizione di una versione "limitata" del formato PDF, esplicitamente ideata per la conservazione a lungo termine dei documenti elettronici e denominata PDF/A (*PDF for Archiving*). 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". Questo nuovo formato nasce quindi per rispondere al bisogno crescente di conservazione a lungo termine dei documenti elettronici.

1.3. Classificazione del formato

Il PDF/A è un formato non proprietario, aperto e standard *de jure*. È basato sul formato PDF di Adobe Systems e di questo utilizza 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 non permette alcuna funzionalità basata su tecnologie proprietarie; inoltre, richiede che i *font* utilizzati nel *file* siano "incorporati" al suo interno per avere la certezza della loro disponibilità in fase di rappresentazione[20].

1.4. Il processo di standardizzazione

Le specifiche del formato sono state riconosciute 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*)[21]. La norma si basa sul *PDF Reference, Third Edition, Version 1.4*[22], implementato nella versione 5 di Adobe Acrobat, ed è stata sviluppata dal gruppo di lavoro ISO/TC171/SC2 – *Document management applications – Application Issues*, per

[20] Alla base di ciò c'è un'implicita fiducia nel fatto che Adobe continuerà a fornire il software per la rappresentazione di *file* PDF/A o che qualche altra organizzazione produrrà un software analogo.

[21] La norma, composta da circa 290 pagine, è disponibile per l'acquisto presso diversi siti, tra cui: <http://www.iso.org>; <http://www.ansi.org>; <http://www.npes.org/standards/ orderform.html>; <http://www.aiim.org/ bookstore>. Una versione *draft* della norma è stato a lungo disponibile all'indirizzo <http://www.aiim.org/documents/standards/ ISO_19005-1_(E).doc>. Nel seguito faremo riferimento a questa norma con la denominazione abbreviata di "ISO 19005-1:2005".

[22] Il *PDF Reference* è il documento che contiene le specifiche del formato. Il *PDF Reference 1.4* contiene le specifiche del formato PDF versione 1.4, utilizzato in Adobe Acrobat 5. Analogamente il *PDF Reference 1.5* contiene le specifiche del formato PDF versione 1.5, utilizzato in Adobe Acrobat 6, e così via. I *PDF Reference* sono disponibili gratuitamente sul sito di Adobe Systems. Ad esempio, l'ultima versione, la 1.7, è disponibile all'indirizzo <http://partners.adobe.com/public/developer/en/pdf/ PDFReference.pdf>.

il quale l'AIIM agisce in qualità di Segretariato. Potrebbe sorprendere il fatto che il comitato tecnico dell'ISO che ha sviluppato le specifiche abbia scelto di utilizzare come base per il nuovo standard il PDF Reference 1.4, che risale a cinque anni prima (2000). Il motivo va ricercato proprio nel fatto che, per essere accettato dalla comunità internazionale, il formato PDF/A aveva la necessità di basarsi su una versione già largamente diffusa del formato PDF e non sulle funzioni delle ultime versioni, non ancora sufficientemente adottate.

Lo standard ISO 19005-1:2005 è la prima parte di una norma pensata in maniera modulare (*multi-parte*), ovvero costituita da più parti che potranno via via essere sviluppate per seguire l'evoluzione nel tempo del formato PDF. Al momento, è stata pubblicata la prima parte dello standard, la ISO 19005-1:2005, che definisce il formato PDF/A-1 ed è basata, come già detto, sul *PDF Reference Version 1.4*. Una nuova versione, denominata PDF/A-2 (ISO/WD 19005-2, *Document management – Electronic document file format for long-term preservation – Part 2: Use of PDF 1.7* (*PDF/A-2*) è in via di sviluppo presso lo stesso comitato tecnico dell'ISO che ha sviluppato il PDF/A-1 e si basa sul *PDF Reference, Version 1.6*, implementato nella versione 7.0 di Adobe Acrobat. Questo nuovo standard disporrà di alcune funzioni aggiuntive, tra cui il supporto al formato di compressione per immagini JPEG 2000; l'inclusione di caratteri *OpenType*, di grafica 3D e di contenuti audio e video; un più sofisticato supporto per la firma digitale; la coerenza con i formati PDF/X, PDF/E, PDF/UA. Nel tempo saranno approvate le altre parti dello standard che faranno riferimento alle successive versioni del formato PDF (si veda la Figura 3). In particolare, la versione PDF/A-3, sarà specifica per i documenti elettronici dinamici[23].

[23] È auspicabile che venga sempre garantita la compatibilità con le versioni successive del formato (*forward compatibility*): in questo modo i *file* che sono conformi alle specifiche PDF/A-1 risulteranno conformi anche alle successive specifiche PDF/A-2, PDF/A-3, e così via. Ovviamente non sarà possibile garantire il contrario. Ad esempio, i *file* conformi allo standard PDF/A-2 non saranno necessariamente conformi allo standard PDF/A-1, dal momento che il PDF/A-2 prevede la possibilità di includere oggetti (quali i contenuti audio e video) che non sono previsti dal PDF/A-1.

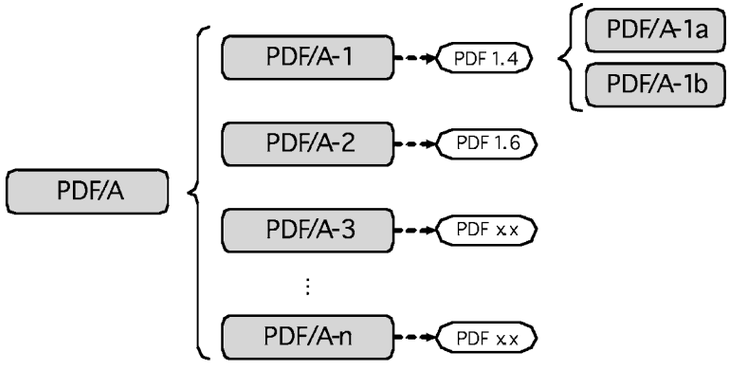


Figura 3. Le varie parti del formato PDF/A.

Lo sviluppo della prima parte dello standard ha coinvolto diverse comunità internazionali in uno sforzo collettivo volto a fornire una risposta all'esigenza, assolutamente improcrastinabile, di disporre di un formato standard per la conservazione a lungo termine dei documenti elettronici. L'iniziativa era stata lanciata nel maggio 2002 negli Stati Uniti dall'AIIM[24], dal NPES[25], dall'Administrative Office of the United States Courts, dalla Harvard University Library ed infine dal United States National Archives and Records Administration (NARA), insieme con esperti di Adobe Systems. Lo sviluppo del formato è stato il risultato di oltre tre anni di incontri, discussioni e revisioni condotte da organizzazioni e compagnie di tutto il mondo. Il primo incontro si tenne nell'ottobre 2002 e vide la creazione di un gruppo di lavoro congiunto AIIM/NPES a cui parteciparono le maggiori organizzazioni, i principali fornitori di *know-how* nel campo dei formati ed i maggiori fruitori di *file* in formato PDF[26]. Nell'aprile 2003 il gruppo di lavoro preparò un *working draft* (WD) iniziale. Nell'agosto 2003 venne predisposto un *new work item proposal* (NP) e, nell'ambito del comitato tecnico ISO/TC171 – Document Management applications, venne costituito il gruppo di lavoro congiunto denominato TC171/SC2/WG5 (noto come PDF/A Joint Working Group) con rappresentanti dei Governi, del mondo accademico e dell'industria (tra cui un ruolo molto attivo è stato ricoperto da Adobe Systems) provenienti dai comitati ISO/TC42 – Photography, ISO/TC130 – Graphics Technology e dal sottocomitato ISO/TC46/SC11 – Information and documentation – Archives/records management. L'ISO assegnò il progetto al sottocomitato tecnico ISO/TC171/SC2 – Document management

[24] AIIM, Association for Information and Image Management.

[25] NPES, Association for Suppliers of Printing, Publishing and Converting Technologies.

[26] Tra cui *The Library of Congress*, Surety Inc., Quality Associates Inc., Appligent, Merck, EMC, PDF Sages, NARA ed, ovviamente, Adobe Systems. Successivamente sono entrati a far parte del gruppo di lavoro altri membri, quali Xerox, Honeywell, EDS, Glaxo Smith Kline.

applications – Application issues. Nel dicembre 2003 venne approvato il primo *Committee Draft* (CD) e, nel settembre 2004, a distanza di quasi un anno, venne approvato un secondo *Committee Draft*. Infine, nel giugno 2005, venne approvato all'unanimità un *Draft International Standard* (DIS). Lo standard definitivo è stato approvato, ancora all'unanimità, da tutti i Paesi partecipanti all'ISO il 14 settembre 2005 ed è stato infine pubblicato il 28 dello stesso mese. A distanza di un anno e mezzo dalla pubblicazione dello standard, il 28 marzo 2007, è stato pubblicato un *Technical Corrigendum* che ha apportato alcune correzioni[27].

1.5. Identificazione del formato

L'estensione di un *file* codificato nel formato PDF/A[28] è *.pdf*, esattamente come quella di un *file* PDF, dal momento che lo standard ISO 19005-1 non richiede che debba essere utilizzata una differente estensione per distinguere il formato PDF/A dal formato PDF[29]. Lo standard specifica anche che la versione PDF/A ed il livello di conformità debbono essere dichiarati mediante un meccanismo di *identificazione interna,* utilizzando lo schema di identificazione indicato nella specifica. Questo schema ha due elementi obbligatori:

pdfaid:part (integer)
pdfaid:conformance (closed list of text values).

Ad esempio, per un *file* PDF/A-1b il valore del parametro **part** deve essere "**1**", mentre il valore del parametro **conformance** deve essere "**B**", come nel seguente esempio:

<pdfaid:part>1</pdfaid:part>
<pdfaid:conformance>B</pdfaid:conformance>

[27] Il *PDF/A Joint Working Group* (TC171/SC2/WG5) ha prontamente identificato gli errori presenti nella prima versione della norma e proposto le dovute correzioni pubblicando l'ISO 19005-1:2005/Cor 1:2007 – *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1) – Technical Corrigendum 1*. Le correzioni riguardano sostanzialmente l'identificazione del formato tramite i *namespace* URI all'interno dei metadati XMP.

[28] Da qui in poi si utilizzerà, per semplicità e laddove non diversamente specificato, la sigla PDF/A, intendendo con ciò la versione PDF/A-1, l'unica disponibile nel momento in cui si scrive.

[29] Sarà compito del programma di visualizzazione dei *file* PDF/A (*PDF/A reader*) individuare il formato corretto. Si noti che, anche per altri formati, non è possibile distinguere le varie versioni tramite l'estensione del nome del *file*. Per fare un esempio, le diverse versioni del software di videoscrittura Microsoft Word producono *file* che hanno tutti la ben nota estensione *.doc* ma che possono essere codificati secondo i differenti formati che si sono avvicendati nel tempo. È compito del software comprendere di quale versione si tratti. Allo stesso modo, le varie versioni del formato PDF (dalla 1.0 alla 1.7 attuale) producono *file* che sono identificati tutti dalla stessa estensione *.pdf*.

Dal momento che nel nome di un *file* in formato PDF/A non è presente alcun elemento che possa far identificare il formato, sta cominciando ad essere molto diffusa la consuetudine di aggiungere al nome del *file*, prima dell'estensione, la desinenza "_A1a" o "_A1b" per indicare che si tratta di un *file* nel formato PDF/A e la sua versione. In questo modo è possibile distinguere subito, a partire dal nome del *file*, se si tratta di un "semplice" PDF o se, invece, si tratta di un *file* in formato PDF/A[30]. Ovviamente occorre poi effettuare una verifica formale del *file*, dal momento che non è sufficiente la semplice modifica del suo nome per far sì che il *file* sia effettivamente codificato in quel formato.

1.6. Analisi del formato

Il PDF/A è un formato non proprietario, aperto e standard *de jure*. È basato sul formato PDF di Adobe Systems e di questo utilizza 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 non consente alcuna funzione basata su tecnologie proprietarie; inoltre, richiede che i *font* utilizzati nel *file* siano inclusi al suo interno per avere la certezza della loro disponibilità in fase di rappresentazione[31].

Nella versione PDF/A-1, il formato è basato sul *PDF Reference, Third Edition, Version 1.4*[32] e rappresenta un sottoinsieme (*subset*)[33] del formato PDF 1.4, una sua versione "limitata" che è stata privata di tutte quelle caratteristiche che non sono adatte per la conservazione a lungo termine di documenti informatici (si veda la Figura 4).

[30] Ad esempio, il nome di *file* "Deliberazione_A1a.pdf" potrebbe indicare un file in formato PDF/A-1a, così come il nome di file "Deliberazione_A1b.pdf" potrebbe indicare un file in formato PDF/A-1b.

[31] Alla base di ciò c'è un'implicita fiducia sul fatto che Adobe continuerà a fornire il software per la rappresentazione di *file* PDF/A o che qualche altra organizzazione produrrà un software analogo.

[32] I *PDF Reference* sono i documenti che contengono le specifiche del formato. Sono disponibili pubblicamente e gratuitamente sul sito *web* di *Adobe Systems*.

[33] In letteratura si utilizza anche il termine "profilo": il PDF/A costituisce un particolare *profilo* del PDF. Il termine è utilizzato, ad esempio, anche per il PDF/X, un *profilo* del PDF utilizzato nel settore della stampa e delle arti grafiche.

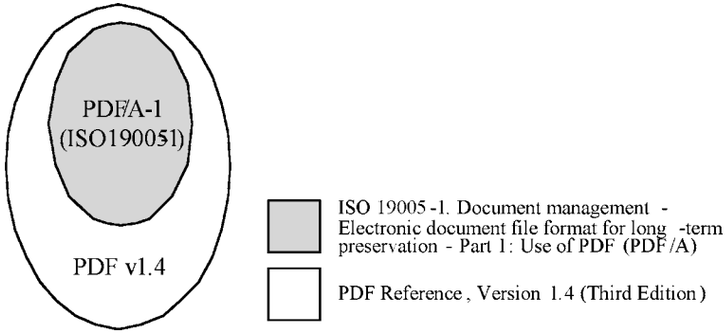


Figura 4. Il PDF/A è un *subset* del PDF.

Come già detto, l'obiettivo del PDF/A è la conservazione a lungo termine dei documenti elettronici, e quindi, rispetto al formato PDF, esso cerca di massimizzare l'indipendenza dal dispositivo (*device independence*), l'auto-contenimento (*self-containment*), e l'auto-documentazione (*self-documentation*). Nelle specifiche vengono chiaramente definite quali caratteristiche di tale formato sono *obbligatorie* (*required*), quali *raccomandate* (*recommended*), quali *limitate* (*restricted*) nel loro impiego e quali infine *proibite* (*prohibited*) (si veda la Figura 5).

Ad esempio, tra le caratteristiche del PDF 1.4 espressamente *proibite* troviamo[34] l'utilizzo di contenuti multimediali audio e video; l'utilizzo di codice eseguibile (ad es. Javascript); la crittografia del *file*; l'uso della trasparenza; l'uso di collegamenti (*link*) esterni; l'utilizzo di *file* incorporati e la compressione con l'algoritmo proprietario LZW. Altre caratteristiche sono invece obbligatoriamente *richieste*: lo spazio dei colori deve essere specificato in una maniera indipendente dal dispositivo su cui il *file* verrà rappresentato; è obbligatorio l'uso di metadati basati su standard (metadati XMP); tutti i *font* debbono essere incorporati nel *file* e devono essere, inoltre, legalmente incorporabili per garantire una rappresentazione universale e illimitata.

[34] Per una trattazione più approfondita si veda il *White Paper: PDF/A – The Basics,* disponibile presso PDF *Tool*s AG all'indirizzo <http://www.pdf-*tool*s.com/public/downloads/whitepapers/ white-paper-pdfa.pdf>.

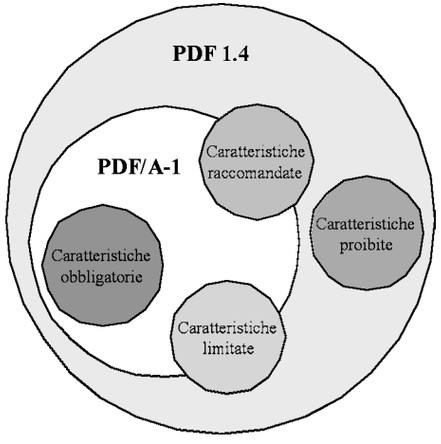


Figura 5. Relazione tra le caratteristiche del formato PDF/A e quelle del formato PDF (adattamento da S. ABRAMS, B. FANNING, D. HELANDER, S. SULLIVAN, PDF/A, *The Development of a Digital Preservation Standard,* SAA 69th Annual Meeting, New Orleans, 14-21 Agosto 2005).

General

Required	Recommended	Restricted	Prohibited
– Conformance to 1.4 requirements	– Linearization hints should be ignored	– Document information dictionary must be consistent with XMP metadata	– Encryption – LZW compression – Embedded files – Optional content – Sound and movie media types

Graphics

Required	Recommended	Restricted	Prohibited
– Device independent color – Embedded color spaces		– Image dictionaries – Separation and DeviceN color spaces – Form XObjects – Extended graphics state – Rendering intents	– Reference XObjects – PostScript XObjects – Non-PDF 1.4 defined operators – Transparency

Fonts

Required	Recommended	Restricted	Prohibited
– Fonts legally embeddable for unlimited, universal rendering – Embedded font programs – Embedded CMap	– Font subsets	– Character encodings	

– Consistent font metrics – Unicode character mapor Level A conformance only			

Annotations

Required	Recommended	Restricted	Prohibited
– Reader mechanism to expose the annotation dictionary Contents key		– Annotation dictionaries	– Non-PDF 1.4 defined types – File Attachment Stound, and Movie types

Actions

Required	Recommended	Restricted	Prohibited
– Behavior for NextPage, PrevPage, FirstPage, and LastPage actions as defined in PDF 1.4 – Reader mechanism to expose GoToR dictionary F and D keys,URI action dictionary, URI key and SubmitForm action dictionary F key			– Launch, Sound, Movie, ResetForm, ImportData, and JavaScript actions – Deprecated state and n-o op actions – Named actions other than the 4 page navigation actions – Widget annotation or Field dictionary AA key

Metadata

Required	Recommended	Restricted	Prohibited
– Requires use of Extensible Metadata Plat (XMP) – Proprietary, but open format – Used for metadata creation, processing and interchange – Based on Resource Description Framework (RDF) – Open World Wide Web Consortium (W3C) standard – Cornerstone of Semantic Web – Pre-defined schemas – Base, DC, DRM, DAM, Workflow, EXIF, PDF, PSD – Defined extension mechanism – Embedding rules – TIFF, JPEG, JPEG2000, HTML, AI, PSD, PDF, – Document level XMP metadata – Equivalent XMP metadata for all appropriate Document	– File identifier – File provenance – Font metadata		– XMP packet header bytes and encoding attributes

Tabella 1. Elenco delle principali caratteristiche del formato PDF/A (fonte: S.ABRAMS, B.FANNING, D.HELANDER, S.SULLIVAN, *PDF/A, The Development of a Digital Preservation Standard*, SAA 69th Annual Meeting, New Orleans, 14-21 Agosto 2005).

Per fare un esempio, un *file* conforme alle specifiche del formato PDF/A può contenere solo testo, immagini *raster* ed oggetti vettoriali; deve inoltre incorporare tutti i *font* utilizzati[35]; non può contenere codice eseguibile e non può essere cifrato. La richiesta relativa ai *font* garantisce che esso verrà riprodotto con i *font* desiderati[36]. Di conseguenza, un'applicazione che produce *file* nel formato PDF/A-1 (il PDF/A-1 *conforming writer*) deve includere nel *file* i *font* utilizzati e, analogamente, un lettore conforme allo standard PDF/A-1 (il PDF/A-1 *conforming reader*)[37] deve utilizzare i *font* inclusi del *file* e non quelli installati nella macchina locale. È vietato l'utilizzo di *file* incorporati, dal momento che, per poterli rappresentare, è necessario disporre dei relativi programmi e può accadere che, nel futuro, alcuni di essi non siano più disponibili rendendo irrecuperabili quei contenuti.

La Tabella 1 elenca nel dettaglio le principali caratteristiche obbligatorie, raccomandate, limitate e proibite, suddividendole nelle categorie *General*, *Graphics*, *Fonts*, *Annotations*, *Actions* e *Metadata*.

1.7. I livelli di conformità

Lo standard ISO 19005-1:2005 prevede due possibili livelli di conformità, il PDF/A-1a (*ISO 19005-1 Level A Conformance in Part 1*) e il PDF/A-1b (*ISO 19005-1 Level B Conformance in Part 1*). Essi differiscono per il grado di conformità allo standard ed in particolare per la presenza di informazioni sulla struttura del *file* e per la possibilità di interpretare semanticamente il testo. Più precisamente:

– il *PDF/A-1a* è il livello di conformità completa, che assicura la rispondenza a tutti i requisiti dell'ISO 19005-1, compresi quelli relativi all'utilizzo dei *tag*. Questi *tag*, forniscono informazioni sulla struttura del *file* e sulla semantica del testo, in maniera da rendere possibile la conservazione della sua struttura logica e del normale ordine di lettura. In sostanza, il PDF/A-1a non solo assicura che il *file* venga riprodotto sempre alla stessa maniera, ma anche che il suo contenuto possa essere interpretato in maniera semanticamente corretta, che sia accessibile alle persone con deficit sensoriali[38] e che possa essere ristrutturato (*reflowed*) per poter essere fruito, ad esempio, in dispositivi palmari o altri apparati.

– il *PDF/A-1b* è il livello di conformità minima, che assicura la rispondenza solamente ai requisiti minimi dell'ISO 19005-1, ovvero quei requisiti che sono suf-

[35] Per questo motivo, le dimensioni di un file in formato PDF aumentano leggermente rispetto allo stesso contenuto in formato PDF senza *font* incorporati.

[36] In altre parole con i *file* PDF/A non si può verificare l'errore di *missing fonts*, che si presenta quando i *font* necessari per rappresentare un contenuto non sono stati incorporati all'interno del relativo *file* né sono installati nel sistema operativo dell'elaboratore dove viene rappresentato. In questo caso tali *font* vengono sostituiti da altri "simili", con la conseguenza che si ottiene una rappresentazione non conforme all'originale.

[37] In base allo standard, il software di produzione deve "avvisare" l'utente nel caso in cui non tutti i *font* possano essere legalmente incorporati nel *file*.

[38] Come richiesto dalla sezione 508 del *U.S. Rehabilitation Act* per quanto riguarda l'accessibilità.

ficienti per garantire solamente che l'aspetto visivo di un *file* PDF/A venga conservato nel tempo[39]. In altre parole, questa versione assicura che il testo e i contenuti aggiuntivi (ad esempio, le immagini) vengano visualizzati correttamente, ma non garantisce, ad esempio, che il testo estratto sia leggibile o comprensibile o che sia possibile interpretare semanticamente il suo contenuto[40]. La Tabella 2 riepiloga le principali differenze tra i due livelli di conformità.

	Iso 19005-1:2005: Pdf/A-1a		**Iso 19005-1:2005: Pdf/A-1b**
Conformità	Conformità PDF/A completa		Conformità PDF/A ristretta
Scopo	Produrre documenti PDF che, oltre a consentire una visualizzazione corretta, consentano anche l'accesso ai contenuti		Produrre documenti PDF che garantiscano la sola visualizzazione corretta
Versione PDF	PDF 1.4		
Metadati	Le specifiche quali: autore, titolo del documento, data di creazione e programma sorgente devono essere congrue con XMP		
Struttura logica	La struttura e l'accessibilità devono essere realizzate mediante l'utilizzo dei tag, la descrizione delle immagini e la specifica del linguaggio utilizzato		Non occorre alcuna struttura logica esplicita
Crittografia	È proibito impostare criteri di sicurezza. Deve essere possibile aprire/elaborare il file PDF in oggetto senza inserire alcuna password		
Colori	I colori devono essere indipendenti dal dispositivo. Gli spazi di colore dipendenti dal dispositivo devono essere identificati mediante il comando *output intent*		
Trasparenze	Non permesse		
Layer	Non permessi		
Compressione	Non è consentita la compressione LZW e nemmeno Jpeg 2000		
Font	Tutti i font devono essere incorporati *(embedded)* all'interno del *file*		
Linguaggi programmazione	Non è consentito l'utilizzo di Javascript		
Moduli	Permessi, ma con alcune restrizioni		

Tabella 2. Confronto tra i due livelli di conformità.

[39] Ad esempio, un documento su carta, acquisito, mediante il processo di scansione, in formato immagine e successivamente convertito in formato PDF, rispetta i requisiti del livello di conformità PDF/A-1b.

[40] Al contrario del PDF/A-1a, il formato PDF/A-1b non garantisce la conformità a quanto richiesto dalla sezione 508 del *U.S. Rehabilitation Act*.

Trattandosi di condizioni più stringenti, un *file* conforme alle specifiche del PDF/A-1a è conforme anche a quelle del PDF/A-1b. Non vale ovviamente il viceversa: un *file* conforme alle specifiche del PDF/A-1b non necessariamente è conforme alle specifiche del PDF/A-1a (si veda la Figura 6, che rappresenta graficamente le relazioni tra il PDF/A1-a, il PDF/A1-b e il PDF 1.4).

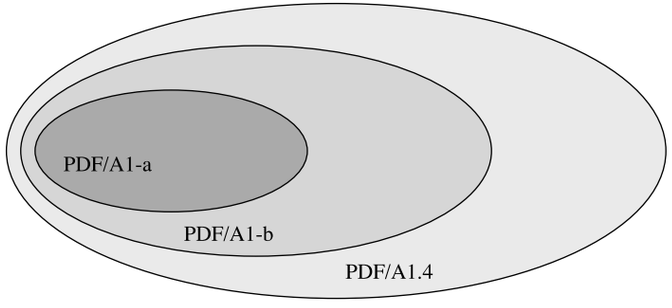


Figura 6. La relazione tra le varie versioni del PDF/A-1.

1.8. Fruizione, produzione e validazione di file nel formato PDF/A

Quando si parla di *file* nel formato PDF/A occorre distinguere tre classi di applicazioni: quelle che ne consentono la sola fruizione (solitamente la visualizzazione a schermo, sia esso quello di un computer o di un altro dispositivo), quelle che ne consentono la fruizione e la produzione (compresa la conversione da un altro formato verso il PDF) ed infine quelle che ne consentono la fruizione, la produzione e la modifica (*editing*).

1.8.1. La fruizione di file PDF/A

Essendo il PDF/A un *subset* del formato PDF, tutti i *viewer* in grado di visualizzare *file* PDF sono anche in grado di visualizzare *file* PDF/A. Il lettore di *default* per i *file* PDF/A è perciò rappresentato da Adobe Reader, ora giunto alla versione 9. Si tratta senza dubbio del *viewer* più completo, con la capacità di sfruttare al massimo tutte le funzionalità del formato. Tuttavia, esistono anche altri lettori, a volte dotati di caratteristiche che li rendono una valida alternativa al pur gratuito Adobe Reader[41].

1.8.2. La produzione di file nel formato PDF/A

La produzione di *file* nel formato PDF/A può avvenire in diversi modi, ma di

[41] A titolo di esempio segnaliamo 3-Heights™ *PDF Viewer*, prodotto da PDF *Tool*s AG <http://www.pdf-*tool*s.com>, un *viewer* che soddisfa tutte le specifiche dello standard ISO per poter visualizzare correttamente *file* PDF/A (oltre a diversi formati immagine: TIFF, JPEG, etc.).

solito un *file* PDF/A non viene prodotto nativamente in questo formato. Infatti, il PDF/A, così come il PDF, è un formato non modificabile[42], per cui raramente si utilizzano gli strumenti messi a disposizione da *tool* di *authoring* (come Adobe Acrobat) per realizzare, da zero, *file* in formato PDF/A. Normalmente il processo che si segue consiste nel produrre *file* codificati in altri formati (ad esempio DOC, XLS, PPT, RTF, HTML) e convertirli nel formato PDF/A, oppure convertire, sempre nel formato PDF/A, documenti analogici (su carta) già esistenti. Vediamo nel dettaglio le varie possibilità (si veda la Figura 7).

a) Conversione da altri formati nel formato PDF/A.

Per produrre *file* nel formato PDF/A, è possibile fare ricorso ad apposite applicazioni software che consentono di trasformare *file* prodotti in altri formati (PDF, DOC, RTF, etc.) nel formato PDF/A. Il software analizza le caratteristiche del *file* di partenza, rimuove gli eventuali elementi non compatibili con lo standard PDF/A e produce, infine, il *file* nel formato PDF/A. Anche in questo caso lo strumento di elezione è sempre Adobe Acrobat. Chi dispone della versione 7.0 di Adobe Acrobat può salvare nel formato "*PDF/A: Draft"*, non necessariamente conforme allo standard PDF/A, dal momento che Adobe Acrobat 7.0 è stato rilasciato prima che tale standard fosse definitivamente approvato. Dalla versione 7.0.7 e con le successive versioni 8 e 9, è invece possibile salvare direttamente nel formato PDF/A secondo lo standard ISO 19005-1. Oltre ad Adobe Acrobat, sono già molti i software di terze parti che permettono di produrre *file* nel formato PDF/A (in entrambi i livelli di conformità, PDF/A-1a e PDF/A-1b). Si tratta sia di prodotti *stand-alone* che di *plug-in,* che consentono di creare *file* PDF/A a partire dai *file* prodotti dai comuni applicativi di *office automation* (come quelli creati con Microsoft Word, Excel, etc.). Ad esempio, la *suite* Microsoft Office 2007 consente di creare *file* conformi allo standard PDF/A direttamente dagli applicativi utilizzando gli *add-in* "*Save as PDF*" oppure "*Save as PDF and XPS*"[43]. OpenOffice.org, dalla versione 2.4, supporta anch'esso il nuovo formato[44]. Esistono anche aziende molto attive nella produzione di software di conversione nel formato PDF/A, che implementano nei loro prodotti funzionalità interessanti e, a volte, non disponibili in Adobe Acrobat[45].

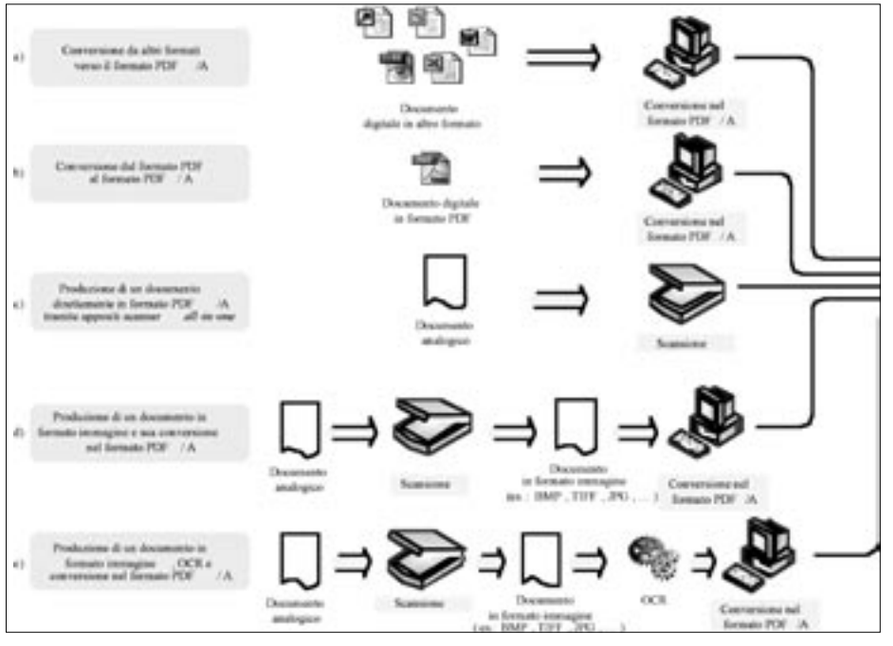


Figura 7. Le modalità di produzione di *file* nel formato PDF/A.

b) Conversione dal formato PDF al formato PDF/A.

Molte organizzazioni dispongono di ampie collezioni di *file* in formato PDF. In tal caso, un'ulteriore modalità di produzione consiste nella conversione dal formato PDF al formato PDF/A mediante procedure automatizzate di tipo *batch* che permettono di agire non sul singolo *file* ma su intere cartelle contenenti *file* PDF. Il programma effettua preliminarmente un'analisi delle caratteristiche dei *file* d'ori-

[42] Un formato si considera *non modificabile* (o *non editabile* o *statico*) quando il programma che lo gestisce non ne consente la modifica in maniera semplice o, comunque, ne consente modifiche solo parziali, con riferimento alla particolare categoria di file. Ad esempio, tra i formati testuali, il PDF è un formato non modificabile dal momento che il programma di *default* con cui viene visualizzato, Adobe Reader, non consente di modificarlo, mentre il programma con cui viene creato, Adobe Acrobat, consente sì delle modifiche, ma non in maniera così agevole come avviene operando su un documento di testo mediante un *word processor*. Al contrario, il DOC e l'ODT sono formati modificabili, dal momento che i programmi di *default* con cui vengono gestiti (Microsoft Word e OpenOffice.org Writer rispettivamente) ne consentono assai agevolmente qualsiasi modifica tra quelle tipiche per la categoria dei documenti di testo. I formati non modificabili sono, quindi, facili da visualizzare, stampare e distribuire ma difficili da modificare. Nella letteratura anglosassone vengono indicati anche come "*not revisable*".

[43] Disponibili entrambi per il *download* all'indirizzo <http://www.microsoft.com/downloads>. La possibilità di salvare nel formato PDF/A è prevista per Word 2007, Excel 2007, PowerPoint 2007, Access 2007, InfoPath 2007, OneNote 2007, Publisher 2007 e Visio 2007.

[44] Cfr. OPENOFFICE.ORG FOO, *Exporting PDF/A-1a for long-term archiving*, disponibile all'indirizzo <http://www.oooninja.com/2008/01/generating-pdfa-for-long-term-archiving.html>.

[45] A titolo di esempio segnaliamo: *3-Heights™ PDF Producer* della *software house* PDF Tools AG, un driver di stampa che genera *file* conformi alle specifiche PDF/A praticamente da qualsiasi applicazione in ambiente Windows; la famiglia di prodotti *PDFlib 7* della *software house* PDFLib GmbH <http://www.pdflib.com>, che permette di produrre *file* secondo le specifiche PDF/A. PDFLib GmBH è stata la prima azienda a produrre software per la creazione di *file* conformi non solo al PDF/A-1b ma anche al PDF/A-1a. Per ulteriori informazioni si veda il *White Paper: Creating PDF/A with PDFlib*, disponibile sul sito dell'azienda.

gine e ne verifica la compatibilità con il livello di conformità desiderato (PDF/A-1a o PDF/A-1b), poi ne effettua la conversione. Ovviamente, se il *file* di partenza è già conforme alle specifiche del formato PDF/A (*PDF/A compliant*), il processo non intodurrà alcuna modifica e il *file* PDF/A che si ottiene è uguale all'originale; se invece il *file* di partenza non è conforme alle specifiche del formato PDF/A, il *file* PDF/A che si ottiene al termine del processo è diverso da quella di partenza, dal momento che vengono rimosse le caratteristiche che non lo rendevano conforme.

È di notevole interesse la possibilità, offerta da alcuni software, di produrre *file* in formato PDF/A a partire da *file* PDF che non sono conformi alle specifiche di questo formato (si veda la Figura 8).

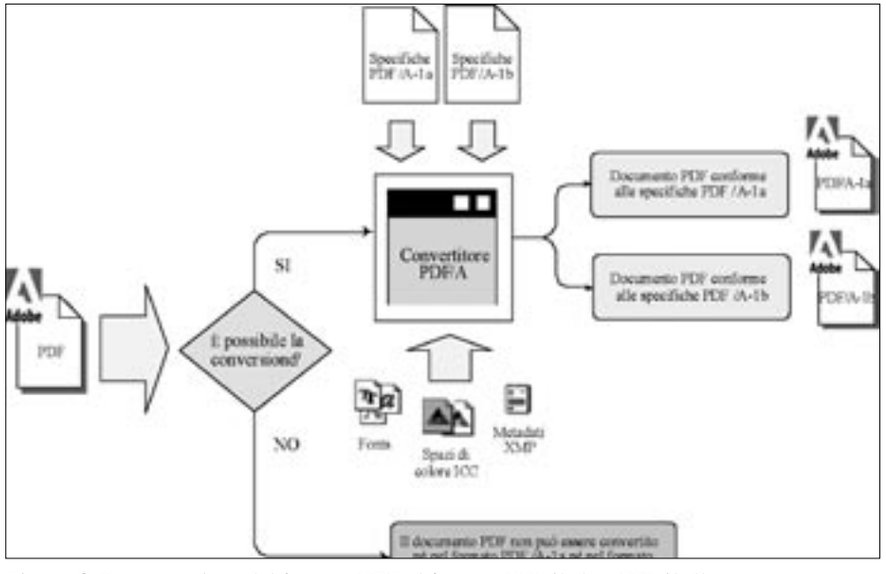


Figura 8. La conversione dal formato PDF al formato PDF/A-1a o PDF/A-1b.

La prima fase del processo prevede l'analisi del *file* PDF in input per verificarne la rispondenza al livello di conformità richiesto (A, B). Qualora ne venga rilevata la non conformità, il convertitore stesso provvede ad introdurre le necessarie modifiche affinché venga risolta: vengono incorporati eventuali oggetti mancanti (come i *font*); gli spazi dei colori dipendenti dal dispositivo sono sostituiti dai profili di colore ICC predefiniti; vengono eliminati eventuali contenuti vietati (JavaScript, etc.) e non essenziali, mentre vengono aggiunti i contenuti obbligatori (metadati XMP, etc.) e può anche essere ottimizzata la dimensione del *file* attraverso il *font*

subsetting[46]. Il *file* è riformattato durante il processo di conversione e vengono effettuate tutte le riparazioni di cui è possibile garantire il successo. Nella pratica si osserva, spesso, un aumento delle dimensioni del *file* in formato PDF/A rispetto a quelle del *file* in formato PDF, dal momento che, ad esempio, in un *file* PDF/A i *font* sono obbligatoriamente incorporati, mentre in un *file* PDF possono non esserlo.

c) d) e) Produzione di file in formato PDF/A a partire da documenti analogici.

Il formato PDF/A trova un'immediata e notevole applicazione quando si ha necessità di digitalizzare documenti analogici (su carta). In questo caso sono possibili tre modalità di produzione. La prima prevede l'utilizzo di appositi scanner in grado di produrre direttamente *file* nel formato PDF/A a partire dal documento analogico. Le altre due prevedono l'acquisizione dell'immagine del documento analogico (in uno dei vari formati immagine quali TIFF, JPEG, PNG, etc.) mediante il processo di scansione; una volta ottenuto il *file* immagine, è possibile produrre il *file* PDF/A direttamente mediante una conversione software oppure facendo precedere la fase di conversione da un'elaborazione attraverso un software di OCR (*Optical Character Recognition*), che permette di ottenere un *file* PDF/A con capacità di estrazione del contenuto testuale e di ricerca *full-text*.

Sul mercato esistono diversi *tool* in grado di convertire *file* in formato immagine in *file* nel formato PDF/A, sia direttamente che effettuando preliminarmente un'elaborazione attraverso un modulo di OCR.

1.8.3. La validazione di documenti PDF/A

Un *validatore* è un'applicazione software che consente di verificare se un *file* in un determinato formato è conforme ad una specifica versione di quel formato. Nel caso del PDF/A, un validatore PDF/A consente di verificare la conformità di un *file*

[46] Per capire cosa sia il *font subsetting*, occorre chiarire prima la distinzione tra *font*, *tipo di carattere* e *carattere*, tre termini che vengono spesso considerati sinonimi ma hanno significati diversi. Un *font* è un insieme completo di caratteri e comprende le lettere, sia maiuscole che minuscole, i numeri, i caratteri speciali, i segni diacritici e così via. Un *tipo di carattere* (*typeface*) contiene una serie di *font*. Ad esempio, Arial è un tipo di carattere e comprende i *font* Arial, Arial Bold, Arial Italic, Arial Bold Italic. Altri esempi di tipi di carattere sono Times New Roman, Garamond, Tahoma, Verdana, Book Antiqua, etc. Quando si genera un *file* PDF, è possibile includere solo i caratteri di un *font* che sono stati utilizzati in quel *file*. Ad esempio, se non vengono utilizzati i caratteri accentati, è possibile evitare di includerli. Questo insieme parziale di caratteri viene chiamato *font subset* e la sua inclusione all'interno del *file* viene definita *font subsetting*. Il principale vantaggio del *font subsetting* è costituito dal fatto che riduce la dimensione del *file* rispetto all'inclusione del *font* completo. Ovviamente, sia nel caso di inclusione del *font* completo che di quello parziale, è assicurata la corretta rappresentazione del *file* anche in assenza di quel particolare *font* nel sistema in cui viene rappresentato. Invece, nel caso in cui il *font* non venga incluso, affinché il *file* venga rappresentato correttamente è necessario che sul sistema sia installato quel *font*; in caso contrario, esso viene sostituito con uno simile e questa sostituzione, specialmente con alcuni tipi di carattere, può dar luogo a rappresentazioni anche molto differenti rispetto all'originale.

al formato PDF/A-1a o PDF/A-1b. È interessante analizzare il suo meccanismo di funzionamento secondo lo schema rappresentato nella Figura 9.

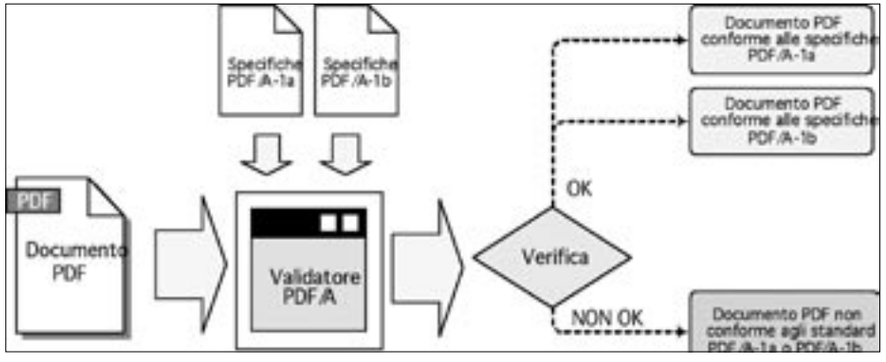


Figura 9. Esempio di validatore PDF/A.

Il validatore riceve in input un *file* PDF qualunque (anche di versioni successive alla 1.4) ed effettua una verifica della sua conformità alle specifiche dei formati PDF/A-1a o PDF/A-1b. L'esito di tale verifica può produrre tre possibili risultati:

– conformità alle specifiche PDF/A-1a (e, di conseguenza, anche a quelle del PDF/A-1b);
– conformità alle sole specifiche PDF/A-1b;
– non conformità alle specifiche PDF/A-1a o PDF/A-1b.

La validazione (o verifica) di *file* in formato PDF si può effettuare utilizzando Adobe Acrobat[47] (versioni 8 e 9), dal momento che tali versioni includono un validatore sia nei confronti del formato PDF/A-1a che PDF/A-1b; esistono, inoltre, anche validatori di terze parti, sia *open source* che commerciali, che svolgono in maniera analoga queste funzioni[48].

[47] Con Adobe Reader 9 è anche possibile verificare se un file PDF *non* è in formato PDF/A. È sufficiente aprire Acrobat Reader 9 e dal menu "Modifica" selezionare la voce "Preferenze" quindi verificare/impostare l'opzione "visualizza i documenti in modalità PDF/A" con il valore "Solo per documenti PDF/A" (in genere è l'opzione predefinita). A questo punto Adobe Reader evidenzierà in automatico ad inizio pagina l'informazione "Il documento viene visualizzato in modalità PDF/A". L'uso di Adobe Reader 9 risulta quindi utile per verificare se un documento non è in formato PDF/A; infatti, se non appare la dicitura sopra indicata il file non è in tale formato.

[48] Si veda il sito del *PDF/A Competence Center*, già citato, per un elenco aggiornato di prodotti.

1.9. Valutazione del formato in relazione ai requisiti per la conservazione digitale

Il PDF/A viene da più parti suggerito come un formato adatto alla conservazione a lungo termine dei contenuti digitali costituiti principalmente da testo ed immagini e nei quali assume particolare importanza la conservazione dell'aspetto visivo. Vediamo nel dettaglio quali sono i requisiti che il formato PDF/A-1 soddisfa[49]:

– *Apertura*. Il PDF/A è un formato aperto, essendo basato su una specifica pubblicamente disponibile. Adobe Systems ha concesso all'AIIM e al NPES i diritti di pubblicazione di questa specifica sui rispettivi siti Internet senza restrizioni di tempo. Inoltre, ha concesso una licenza *royalty-free* per l'utilizzo della specifica, per cui chiunque può creare applicazioni che gestiscono i *file* PDF/A.
– *Completa documentazione*. Il formato PDF/A-1 è pienamente documentato nello standard ISO 19005-1:2005. Questo standard va utilizzato assieme al *PDF Reference, Third Edition, Version 1.4*, che è considerato come norma di riferimento ed è anch'esso pienamente documentato e liberamente disponibile.
– *Non proprietà*. Il formato PDF/A è standard ISO e, come tale, non proprietario. Il mantenimento delle specifiche è a carico dell'ISO.
– *Standardizzazione*. Il PDF/A è un formato *standard*, approvato nel maggio 2005 e pubblicato dall'ISO nel settembre 2005. Si noti che il PDF/A è uno *standard aperto* (*open standard*), essendo un formato che possiede entrambe le proprietà di standardizzazione ed apertura.
– *Ampia adozione*. Da quando, alla fine del 2005, lo standard è stato pubblicato, diverse aziende hanno cominciato a produrre *tool* per la creazione, la conversione e la validazione di *file* PDF/A. Già Adobe Acrobat Professional nella versione 7.0 consentiva la creazione di *file* in un formato compatibile con la versione DIS (*Draft International Standard*) del PDF/A, non essendo questo ancora pubblicato come standard. Le successive versioni, Adobe Acrobat 8 e 9, supportano pienamente il formato nella versione definitiva. Dal 2006 ad oggi, diverse società commerciali hanno prodotto soluzioni per la gestione di *file* PDF/A.
– *Trasparenza*. I *file* PDF/A sono di tipo binario, di conseguenza non sono *human-readable* e la loro fruizione può avvenire solo mediante appositi visualizzatori.
– *Robustezza*. Essendo un formato binario, i *file* PDF/A non sono particolarmente robusti e la corruzione anche di pochi *byte* può portare alla perdita dell'intero contenuto del *file*.

[49] Per un'analisi dettagliata dei requisiti che un formato deve possedere per poter essere considerato compatibile con un processo di conservazione digitale si rimanda al volume S. Pigliapoco, S. Allegrezza, *Produzione e conservazione del documento digitale. Requisiti e standard per i formati elettronici. Volume I*, Edizioni EUM, Macerata, 2008.

- *Auto-contenimento.* Tutte le informazioni necessarie per poter rappresentare correttamente (a video o a stampa) un *file* PDF/A, sono incluse al suo interno[50].
- *Auto-documentazione*. Un *file* in formato PDF/A è auto-documentato. La specifica del formato richiede che venga utilizzata la tecnologia Adobe *eXtensible Metadata Platform* (XMP)[51] per includere i metadati all'interno del *file;* tuttavia viene consentito l'utilizzo di atri metadati standard.
- *Indipendenza dal dispositivo.* Un *file* PDF/A può essere visualizzato, stampato o comunque riprodotto in maniera affidabile e coerente, sempre nello stesso modo indipendentemente dalla piattaforma hardware e software utilizzata.
- *Assenza di meccanismi tecnici di protezione*. Il formato PDF/A non consente meccanismi di protezione tecnica. Esso vieta, ad esempio, la crittografia, così come l'accesso mediante *username* e/o *password*[52].
- *Assenza di limitazioni sull'utilizzo.* Il formato PDF/A non presenta limitazioni sul suo utilizzo[53].
- *Accessibilità*. La struttura logica del *file* PDF/A è disponibile solo se durante il processo di creazione vengono fornite le necessarie informazioni mediante l'uso corretto dei *tag*[54], come raccomandato dalla specifica.
- *Stabilità*. Il PDF/A-1 si basa sul formato PDF versione 1.4. È già prevista la successiva versione che si baserà sulla versione 1.6 del PDF, ma non è stata ancora rilasciata, per cui non è possibile fare ipotesi sulla stabilità del formato[55].
- *Non modificabilità*. Come il PDF, anche il PDF/A appartiene alla categoria dei formati non modificabili.
- *Sicurezza*. Come il formato PDF da cui deriva, anche il PDF/A, allo stato attuale delle conoscenze, non può contenere virus o altre forme di malware.
- *Efficienza*. Ereditando le caratteristiche del PDF, anche il PDF/A è un formato che, a parità di contenuti, consente dimensioni dei *file* che sono sensibilmente inferiori rispetto a quelle di altri formati, come il DOC o l'ODT.

[50] In particolare, i testi, le immagini, i *font* utilizzati, le informazioni relative al colore e tutte le altre informazioni necessarie per rappresentare un documento devono essere contenute nel *file*. Inoltre, un *file* PDF/A non può contenere informazioni che siano raggiungibili da fonti esterne (ad esempio, tramite *link*).

[51] XMP (*eXtensible Matadata Platform*) è una tecnologia, creata da Adobe Systems Inc., per l'elaborazione e la conservazione di metadati standard e proprietari.

[52] Nel caso sia necessario un controllo dell'accesso, esso può essere realizzato mediante meccanismi esterni: ad esempio, è possibile comprimere in formato ZIP il *file* PDF/A e applicare al *file* compresso la protezione mediante *username* e *password*.

[53] Sebbene la specifica includa frasi che sembrerebbero indicare la possibilità che alcuni degli elementi del formato possano essere soggetti a licenze.

[54] Ovvero, se si utilizzano i PDF strutturati o, meglio ancora, i *tagged* PDF.

[55] Anche se più stabile rispetto al PDF, dal momento che, ad esempio, non sono previste le versioni corrispondenti ai formati PDF 1.5 e 1.6.

Gli altri formati proposti per la conservazione nel tempo dei documenti non sempre soddisfano un numero così elevato di requisiti. Ad esempio, il formato immagine TIFF (*Tagged Image File Format*), utilizzato per molti anni sia da enti pubblici che da aziende private per conservare negli archivi digitali ogni genere di documentazione (fatture, corrispondenza, contratti, etc.), non è dotato degli stessi requisiti del PDF/A, pur rimanendo comunque un buon formato. Infatti, il TIFF è un formato immagine di tipo *raster* e, come tale, presenta vantaggi e svantaggi. I formati *raster* immagazzinano l'aspetto delle pagine pixel per pixel. Non vi sono problemi relativi ai *font* mancanti, perché il formato immagazzina tutti gli elementi della pagina, compresi i testi, come immagini. Tuttavia, proprio a causa del fatto che si tratta di un formato immagine, il TIFF non consente il riconoscimento del testo o la sua ricercabilità, a meno di non sottoporre i *file* codificati in questo formato ad un processo di riconoscimento ottico dei caratteri (OCR)[56]. Inoltre, sebbene a causa della sua grande diffusione possa essere considerato un formato standard *de facto*, le specifiche del TIFF non sono state ufficialmente riconosciute e non è, quindi, uno standard *de jure*.

Anche il PDF, di cui, come si è visto, il PDF/A costituisce un *profilo*[57], gode di proprietà interessanti ma non "obbliga" l'utente a produrre *file* che forniscano valide garanzie ai fini della conservazione digitale. Ad esempio, nel formato PDF non è obbligatoria l'inclusione dei *font* o l'utilizzo di spazi di colore indipendenti dal dispositivo e questo fatto può dar luogo a rappresentazioni a video o a stampa che non sono del tutto conformi all'originale. Il PDF è stato, inoltre, riconosciuto standard *de jure* a seguito della pubblicazione della norma ISO 32000[58], ma questa norma non definisce un formato specificatamente pensato per la conservazione digitale. Analoghe riflessioni si potrebbero fare per altri formati che vengono via via proposti si fini della conservazione digitale, come l'XPS (*XML Paper Specification*), il formato sviluppato alla fine del 2006 da Microsoft e che costituisce un potenziale "concorrente" del formato PDF di Adobe Systems.

La Figura 10 mostra un confronto schematico fra le caratteristiche dei formati PDF/A, PDF e TIFF; si può osservare come il primo formato sia quello che soddisfa il maggior numero di requisiti nella prospettiva della conservazione dei documenti digitali.

[56] I sistemi di OCR (*Optical Character Recognition*) sono programmi che consentono di riconoscere il testo (ed, eventualmente, la sua formattazione) contenuto in un'immagine e di ottenere, quindi, del testo "elaborabile".

[57] Si veda la nota 33.

[58] ISO 32000-1:2008 – *Document management — Portable document format — PDF 1.7*. A seguito del riconoscimento come standard, il formato PDF non solo è divenuto standard *de jure*, ma ha perso anche la caratteristica di proprietà, dal momento che, per conseguire la standardizzazione, Adobe Systems ha dovuto rilasciare la specifica del formato all'ISO, a cui spetta, d'ora in poi, il mantenimento.

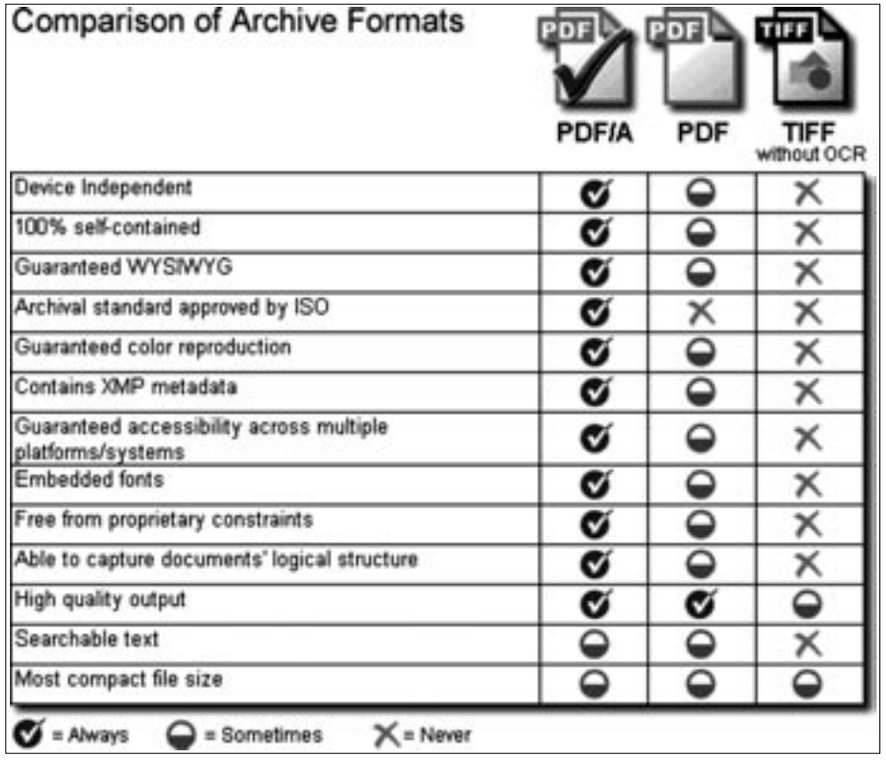


Figura 10. Confronto tra i formati PDF/A, PDF e TIFF (fonte: http://www.soliddocuments.com).

1.10. Conclusioni

Come si è visto, il PDF/A è un formato non proprietario, standard *de jure* (ed in via di diventare anche standard *de facto*), indipendente dalla piattaforma, aperto, completamente documentato, auto-contenuto, auto-documentato, privo di meccanismi tecnici di protezione e di limitazioni sul suo utilizzo, accessibile (se viene fatto un uso corretto dei *tag*), non modificabile ed efficiente. Le limitazioni imposte dallo standard nei confronti del PDF da cui deriva (come l'obbligo di inclusione di tutti gli oggetti necessari per una fedele rappresentazione – ad esempio, i *font* – e il divieto di inserire altri elementi che potrebbero rendere difficile l'accesso – ad esempio i contenuti multimediali), fanno si che sia possibile prevedere, con assoluta precisione, il risultato che si otterrà quando il *file* verrà visualizzato a schermo o stampato su carta, anche a distanza di anni.

Inoltre, il formato presenta ulteriori caratteristiche estremamente interessanti,

come la possibilità di essere sottoscritto con firma elettronica e la possibilità di inserire, nei *file* ottenuti da scansioni di documenti di testo, due livelli informativi: uno contenente l'immagine scansionata (garantendo, quindi, la conservazione dell'aspetto visuale) e l'altro, "invisibile" e integrato "sotto l'immagine", contenente il testo estratto mediante OCR (garantendo, quindi, la ricercabilità e la consultabilità al pari di un "normale" documento di testo in formato elettronico)[59].

Il PDF/A deve, quindi, essere preso seriamente in considerazione ai fini della sua adozione come formato per la produzione di documenti elettronici conservabili.

In Italia esso è ormai ampiamente diffuso e quotidianamente utilizzato da aziende e pubbliche amministrazioni. L'interesse verso questo formato nel nostro Paese è tale che agli inizi del 2010 è stato istituito il comitato italiano del *PDF/A Competence Center*, l'associazione internazionale nata nel 2006 in Germania con l'obiettivo di promuovere lo scambio di informazioni ed esperienze nel campo della conservazione a lungo termine in accordo alla norma ISO 19005. La prima riunione del comitato ha avuto luogo lunedì 8 febbraio 2010, ed ha visto la partecipazione di più di dieci società che hanno già adottato il PDF/A come standard per la conservazione dei documenti a lungo termine. Inoltre, a conferma dell'interesse che questo formato sta suscitando, proprio l'Italia è stata scelta per ospitare la prossima Conferenza internazionale sul PDF/A (*4th International PDF/A Conference*), in programma a Roma dal 29 settembre al 1 ottobre 2010[60].

[59] Al contrario delle tradizionali scansioni in formato immagine che immagazzinano i contenuti sotto forma di pixel e quindi rendono il testo in essi presenti del tutto inutilizzabile.

[60] Per i dettagli dell'evento si rimanda al sito del *PDF/A Competence Center*, già citato.