Elemento di Informatica

Obiettivo della lezione

- I documenti digitali
- Formati dei documenti digitali
- Applicazioni per elaborare documenti
  - Editor di testi
  - Gestore di foglio elettronico
  - Editor di presentazione
  - Applicazione grafica bitmap
  - Applicazione grafica vettoriale

Documento

Documento:

Contenitore (o *supporto*) di dati (testo, numeri, figure, ecc.) strutturati per essere usati come informazioni

Esempi

- Documento Word
- Foglio elettronico Excel
- Presentazione Powerpoint
- Documento PDF
- Documento HTML
- …

Documento digitale

Documento digitale:

Documento rappresentato in forma **binaria** e **memorizzato** in un sistema informatico

Ha un **nome**, un **contenuto**, una **struttura**, alcuni **comportamenti**, alcune **relazioni** con altri documenti

Il contenuto è di solito **codificato** (anche più volte, da codici sovrapposti)

Codici sovrapposti


The Imagery of Chess Revisited

Documento visualizzato

Versione HTML

↔

010011101
010010111
000111110
111000111
011001101
001010101
010011110
101010001
010110101
010011010

Codice binario

Versione PDF

↔

010011101
010010111
000111110
111000111
011001101
001010101
010011110
101010001
010110101
010011010

Codice binario

Dimensioni del codice

Questo documento (che stai leggendo) salvato in formati diversi ha dimensioni diverse, perché nei tre casi si usano codici diversi:

- Elinf5.ppt: 3.200 KB
- Elinf5.pdf: 2.200 KB
- Elinf versione pagina Web: 8.000 KB

Codici sovrapposti

- I documenti digitali che contengono testo sono rappresentati internamente ad un sistema informatico:
  - da un ***codice alfanumerico***, ad esempio ASCII o Unicode, che definisce la **rappresentazione del testo** del documento in forma di bit
  - da un ***codice di markup***, ad esempio HTML, che definisce la ***struttura*** del documento
  - un ***formato di fruizione***, ad esempio PDF, che definisce una rappresentazione *intellegibile* del documento in **forma *grafica***, per es. mediante caratteri di stampa a video

Nomi di documenti

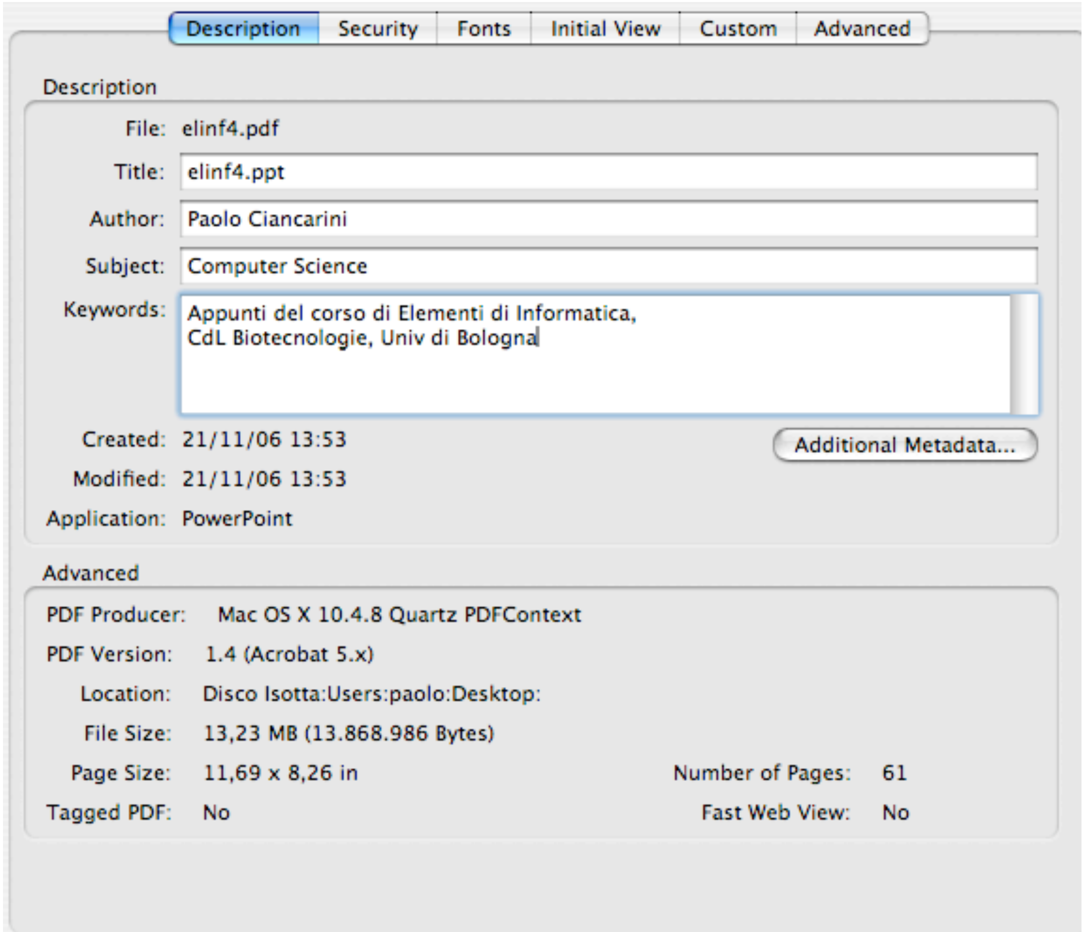
- Nomi all’interno di un computer
  - Es. /Doc/didattica/elinf/lezione1.ppt
- URL: nome di una risorsa su Web
  - Es.: http://www.cs.unibo.it/ciancarini/didattica/elinf/lezione1.pdf
- URN: nome in uno spazio di nomi (namespace)
  - Es.: *urn:isbn:0-395-36341-1*
- Metadati

Metadati

- I metadati, o proprietà, descrivono alcuni attributi del documento
- Vanno riempiti a cura dell'autore o del proprietario del documento

Metadati in .doc

Metadati in .pdf



Struttura di un documento

Titoletto primo livello

Sommario

Titoletto secondo livello

Elenco numerato

7 Struttura globale di un documento HTML

Sommario

7.1 Introduzione alla struttura di un documento HTML

Un documento HTML si compone di tre parti:

1. una riga contenente informazioni sulla versione di HTML,
2. una sezione esplicativa di intestazione (delimitata dall'elemento `HEAD`),
3. un corpo, che contiene il contenuto effettivo del documento. Il corpo può essere implementato per mezzo dell'elemento `BODY` o dell'elemento `FRAMESET`.

Dello spazio bianco (spazi, a capo, tabulazioni e commenti) può comparire prima o dopo ciascuna sezione. Le sezioni 2 e 3 dovrebbero essere delimitate dall'elemento `HTML`.

Ecco qui un esempio di un semplice documento HTML:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
   "http://www.w3.org/TR/html4/strict.dtd">
<HTML>
   <HEAD>
      <TITLE>Il mio primo documento HTML</TITLE>
```

Struttura di un documento

- Descrizione **implicita**
- Descrizione **esplicita**

- Una descrizione implicita della struttura si ottiene definendo un **tipo** di documento (Es. *libro* oppure *articolo* oppure *lettera* oppure *spartito* oppure…)
- Una descrizione esplicita della struttura si ottiene con un *linguaggio di markup* (es. XML)
- In ogni caso esiste una **grammatica formale** del documento, che contiene le regole che dicono se una struttura è corretta oppure no

Esempio

- Titolo
- Autore
- Sommario
- Introduzione
- Paragrafo
  - Sottoparagrafo
  - Sottoparagrafo
- Paragrafo

Struttura corretta

- Titolo
  - Sottoparagrafo
  - Autore
- Introduzione
- Paragrafo
- Sommario
  - Sottoparagrafo
- Paragrafo

Struttura scorretta

Strumenti di fruizione

- Uno **strumento di fruizione** è un programma capace di visualizzare documenti digitali codificati in un formato particolare
- Esempi:
  - Web Browser per documenti HTML
  - Microsoft Word per documenti .doc
  - Adobe Reader per documenti PDF
  - Microsoft Reader per documenti .lit

Ciclo di vita dei documenti digitali

1. **Authoring**: fase (e relativi strumenti di *editing*) in cui vengono creati i **contenuti** di un documento digitale
2. **Transformation**: fase (e relativi strumenti di *presentazione*) in cui vengono elaborati i **formati di fruizione** di un documento digitale
3. **Delivery**: fase (e relativi strumenti di *publishing*) in cui un documento digitale viene **trasmesso** e fruito da un utente mediante qualche **dispositivo**

Cosa crea la fase di authoring

Gli strumenti di authoring (Es. Word) si basano sulla metafora della "pagina vuota": aiutano l'autore a riempirla di *contenuti*

I documenti digitali contengono principalmente

- Testo (lettere e numeri)
- Grafica vettoriale
- Grafica bitmap

Componenti dei documenti

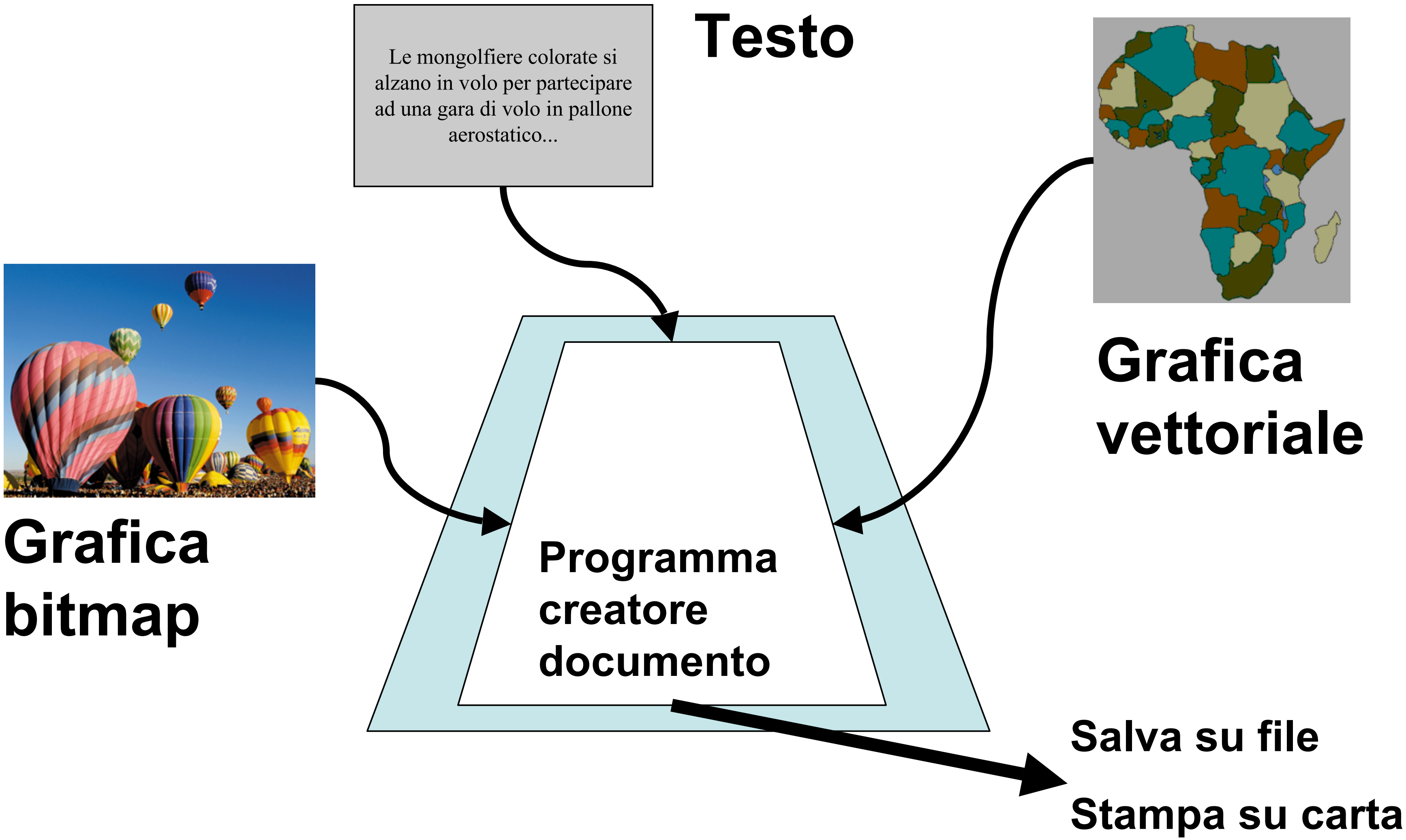
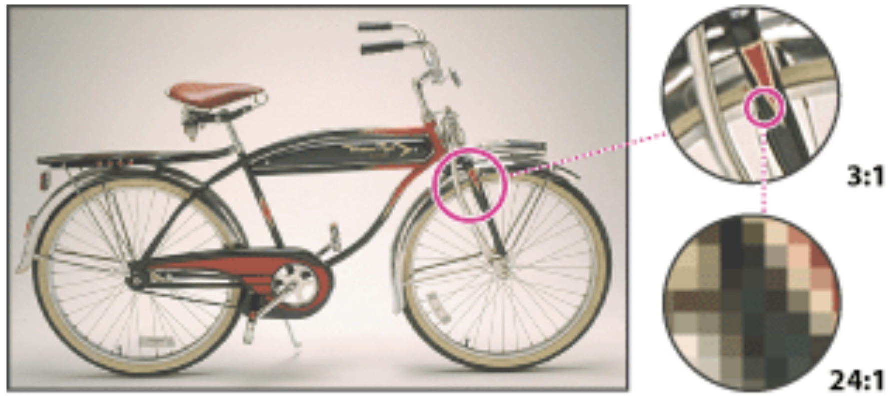
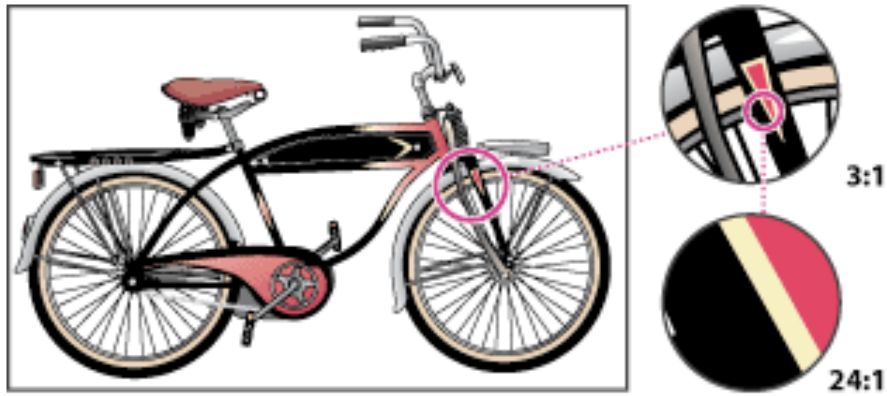


Figure bitmap vs vettoriali

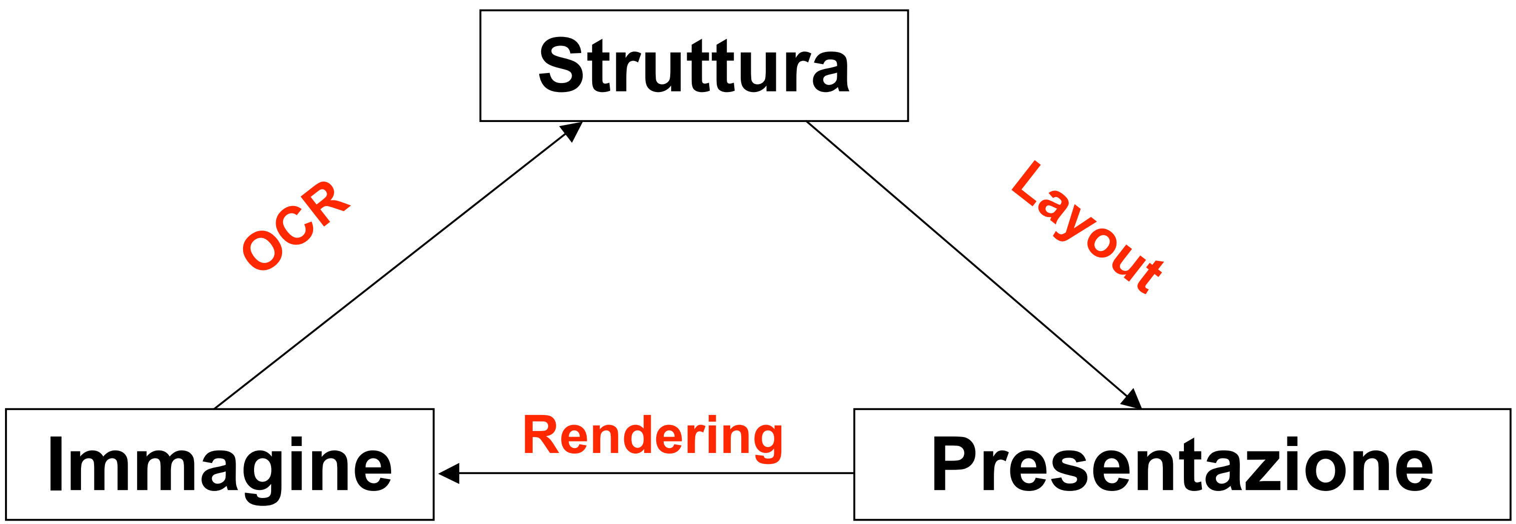
- Grafica bitmap



- Grafica vettoriale



Conversione di documenti digitali



Transformation e Delivery

Mentre gli strumenti di authoring aiutano a "riempire pagine vuote",

- gli strumenti di *transformation* ridefiniscono il layout o la struttura di documenti esistenti (es.: browser HTML)
- gli strumenti di *delivery* "conservano le pagine" e le fanno arrivare al lettore così come l'autore vuole che arrivino (es.: formato PDF)

Formato di fruizione

- **Formato di fruizione** (o *delivery format*): tipo del file che riceve l'utente che accede un documento digitale
- E' rilevante non solo per la miglior conservazione del contenuto e della sua forma, ma ad esempio anche per i motori di ricerca
- Nessun formato di fruizione esistente è superiore agli altri in ogni situazione
- Aspetti critici:
  - Formati aperti o chiusi
  - Usabilità e accessibilità
  - Aspetti sensibili al dominio applicativo dei documenti. Esempio: documenti da archiviare

Formati per documenti digitali

- Formati aperti: XML e HTML
- Formati proprietari: Word, Excel, ppt
- Formato di interscambio: RTF
- Immagini: TIFF, JPEG, PNG, SVG
- Formati ibridi (immagini e testo): PDF, DjVu
- Animazioni: Flash

Immagini pure: TIFF, GIF, JPEG

- TIFF miglior scelta per archivio immagini da scanner
- GIF scelta popolare per grafica su Web (ma non foto)
- JPEG formato compresso per immagini digitali

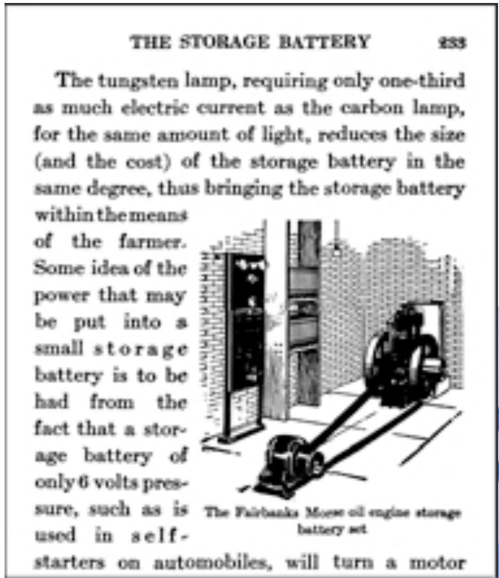
Pro

- Visualizzazione nativa nei web browser
- Formati aperti

Contro

- TIFF: Dimensione gigante dei file
- Testo disponibile solo via OCR (se il tasso di errore arriva al 20%, questo influenza i motori di ricerca)
- Supporto debole per documenti multipagina
- Quando un JPEG è fortemente compresso, la visualizzazione del testo è pessima
- Metadati solo sul file fisico, non sui contenuti

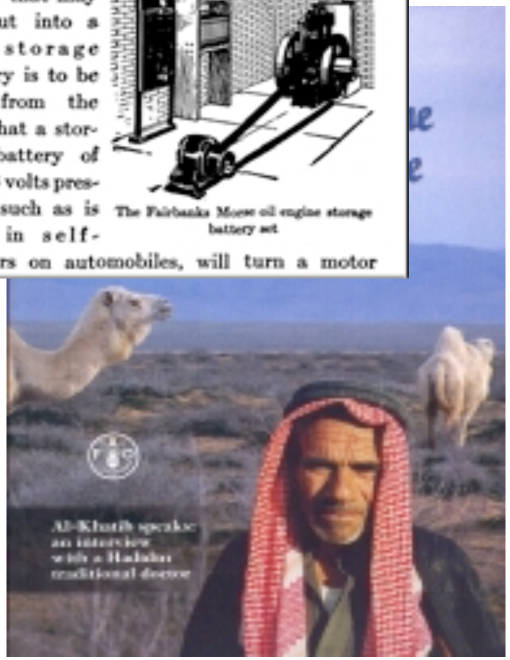
TIFF (Tagged Image File Format)



THE STORAGE BATTERY 233

The tungsten lamp, requiring only one-third as much electric current as the carbon lamp, for the same amount of light, reduces the size (and the cost) of the storage battery in the same degree, thus bringing the storage battery within the means of the farmer. Some idea of the power that may be put into a small storage battery is to be had from the fact that a storage battery of only 6 volts pressure, such as is used in self-starters on automobiles, will turn a motor

The Fairbanks Morse oil engine storage battery set



TIFF progettato da Aldus (oggi Adobe) nel 1987, l'ultima versione è del 1992. E' oggi il più comune formato bitmap, adatto per ogni profondità cromatica.

Buona scelta per l'archiviazione di documenti,ma non per publishing su Web.

partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

TIFF

- Può memorizzare molti tipi diversi di immagini (monocrome, grigie, 8-bit & 24-bit RGB)
- Progettato per interoperabilità di applicazioni che manipolano immagini
- Diffuso per desktop publishing, scanning, e elaborazione delle immagini

GIF **G**raphics **I**nterchange **F**ormat

- Standard del 1987, rivisto 1989
- Immagini a colori su 8-bit di profondità cromatica
- Dunque un'immagine GIF ha 256 colori (o 256 punti di grigio)
- Supporta trasparenza e animazione
- Molto diffuso a causa della sua efficienza
- Tutti i browser lo supportano efficacemete senza bisogno di plug-in

`www.mwasoftware.co.uk/`**`gif`**`/`**`gif`**`f89a.pdf`

JPEG **J**oint **P**hotographics **E**xperts **G**roup

- **JPEG/JPG** (Joint Photographic Experts Group) non è un formato, quanto un metodo di codifica e compressione e si usa spesso insieme a TIFF
- Il formato prodotto dal metodo si chiama JPEG FIF (File Interchange Format) ed è standard dal 1992
- Progettato per comprimere immagini fotografiche a colori o B&W
- Ottimo per foto su Web; supporta milioni di colori con ottima comprimibilità

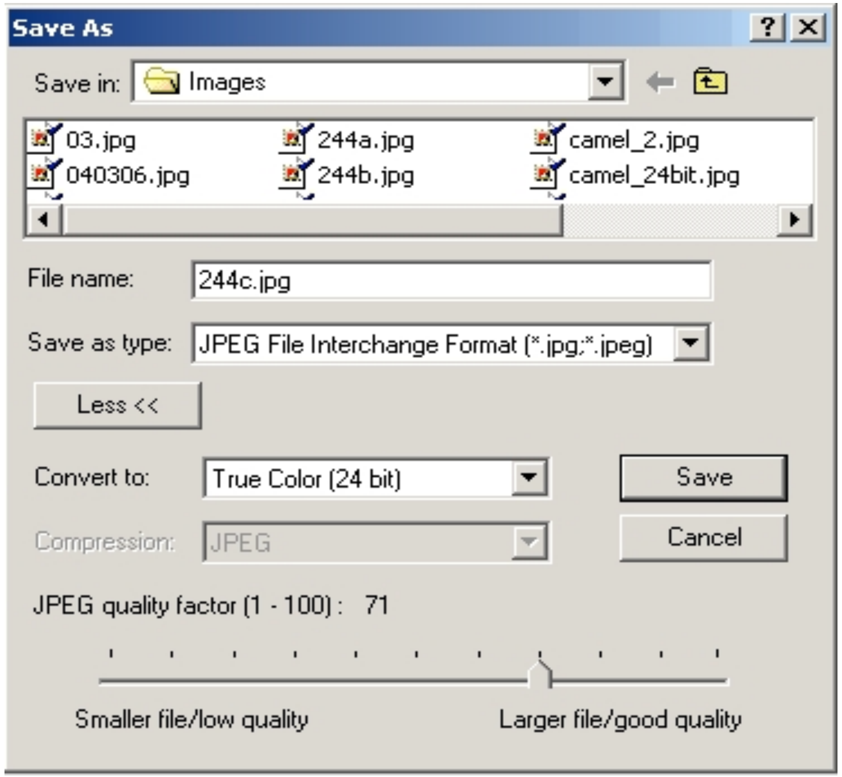
`www.w3.org/Graphics/JPEG/`

JPEG

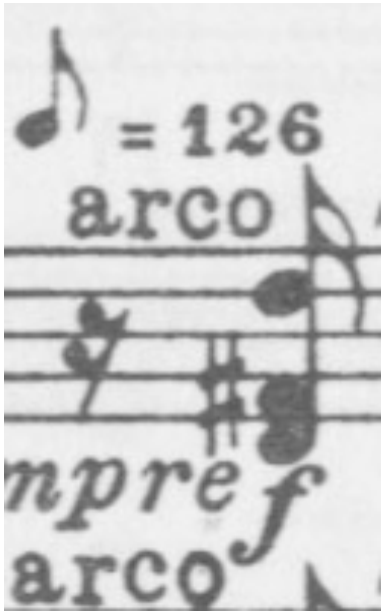
JPEG offre un metodo di compressione per immagini fotografiche con una profondità da 6 a 24 bit. La compressione primaria è **lossy**. E' possibile scegliere quanta compressione applicare, ma maggiore la compressione maggiore l'informazione che viene persa.

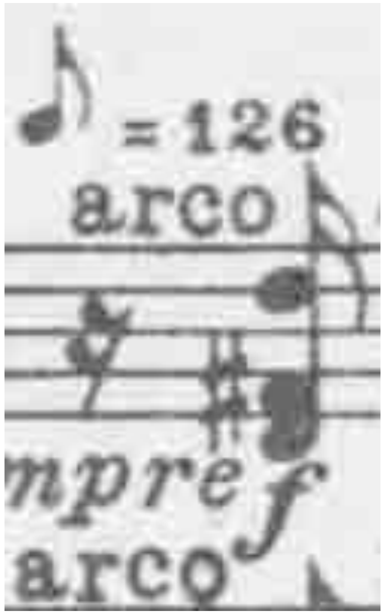
Alcune forme di compressione JPEG sono considerate **visualmente lossless**. In genere, un file JPEG comprime una foto da 2 a 3 volte in più rispetto a GIF.

La compressione lossy rende JPG una **cattiva scelta** per scopi di archiviazione o comunque se in seguito occorre **piena qualità dell'immagine**.



Compressione in JPEG



Bassa compressione

Alta compressione

PNG (Portable Network Graphics)

- PNG è uno standard W3C per rimpiazzare GIF; profondità 16 (BW) o 48 (colore) bit
- Vantaggi su GIF: trasparenza variabile, controllo della luminosità, correzione del colore e visualizzazione progressiva
- Compressione lossless

SVG (Scalable Vector Graphics)

- SVG: standard W3C basato su XML
- Descrive grafica vettoriale a due dimensioni
- Occorre plugin (Adobe)

Tabella dei formati

	Progettato per…	Uso sul Web
TIFF	Immagini alta risoluzione da stampare o archiviare.	Non adatto perché un TIFF può essere gigantesco
GIF	Immagini con grosse aree omogenee di colore (e.g. logo, diagrammi, grafici)	Adatto, supportato da tutti i Web browsers
JPEG	Immagini con più di 256 colori (e.g. foto)	Adatto, supportato da tutti i Web browsers
PNG	Rimpiazzare GIF e TIFF	Adatto, supportato da molti Web browser
WMF EMF	Interscambio per MS Office di immagini vettoriali	Non adatto per publishing Web
EPS	Importazione di immagini PostScript	Non adatto per publishing Web
SVG	Immagini vettoriali, linguaggio basato su XML	Non supportato pienamente dai browser, occorre plugin

Formati aperti o chiusi

- Chi controlla il formato del formato dei documenti?
  - Un comitato di standardizzazione? Il formato si dice *aperto*
  - Un'azienda specifica? Il formato si dice *chiuso*, o *proprietario*
- Vantaggi dei formati aperti (cioè standard)
  - Supporto nel futuro
  - Interoperabilità
  - Buona integrazione con software open source
  - Buona diffusione tra utenti e web designers

Formati proprietari

- MS Word, WordPerfect

Pro:

- Formati molto diffusi per authoring
- Integrazione con varie applicazioni (es. MS Office)

Contro:

- Formati chiusi
- Strumenti di solito costosi
- Non adatti a fruizione su Web
- Mal supportati su piattaforme multiple
- Indicizzazione problematica con strumenti open source

RTF **R**ich **T**ext **F**ormat

- RTF formato pubblicato da Microsoft che specifica informazioni di layout su documenti testuali
- Versione 1.6 del 1999
- markup procedurale
- Utile come formato di interscambio anche al di fuori di applicazioni Microsoft
- http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrtfspec/html/rtfspec.asp

Formati aperti: XML e HTML

- Si creano in tre modi:
  - Documenti digitali "nativi"
  - Documenti creati a mano
  - Da scanner e OCR, eventualmente con correzioni

Pro

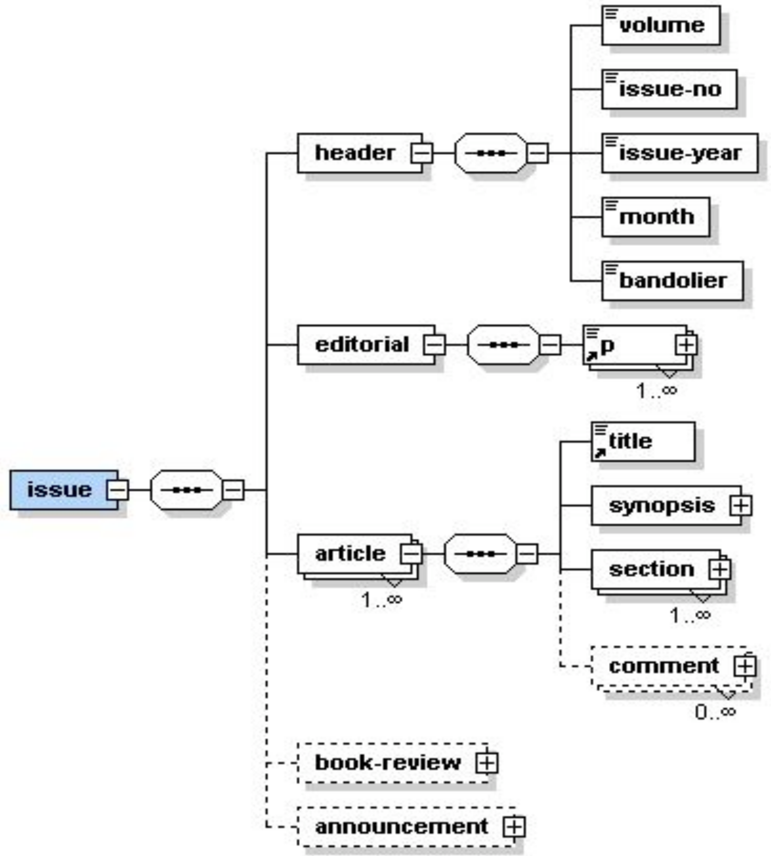
- Ricerca facilitata (sono facilmente indicizzati dai motori di ricerca)
- Standard internazionali (sono raccomandazioni W3C)
- Accessibili per tutti i browser
- Metadati facilmente aggiungibili
- Editabili con qualsiasi editor (disponibilità di strumenti gratuiti)

Contro

- Costosi da creare a mano o correggere da OCR
- Pienamente sfruttabili solo da programmatori
- Non preservano il layout

XML EXtensible Markup Language

- Mark-up descrittivo
- Standard W3C
- Descrive la struttura logica del documento
- Lo stile di presentazione viene definito da uno stylesheet



Esempio: book.xml

```
<book>
	<title>Libro di Esempio</title>
	<titleabbrev>Esempio</titleabbrev>
	<bookinfo>
<author>
	<firstname>Paolo</firstname>
	<surname>Ciancarini</surname>
</author>
</bookinfo>

<preface>
	<title>Premessa</title>
<para>C’è sempre una premessa.</para>
</preface>

<chapter><title>Capitolo</title>
	<para>In un libro c’è sempre almeno un capitolo.</para>
</chapter>

<appendix><title>Appendice opzionale</title>
<para>Le appendici possono non esserci.</para>
</appendix>
</book>
```

Differenze tra XML e HTML

- XML non rimpiazza HTML
- XML serve a **descrivere** la struttura dei dati/documenti
- HTML serve a **visualizzare** i dati/documenti
- La visualizzazione di un documento XML ha bisogno di un “foglio di stile”

Formati ibridi: PDF, DjVu

- Tecnologie proprietarie molto diffuse nelle comunità Open Source
- Si definiscono "ibridi" perché contengono entrambi un "layer" per il testo e un'immagine (*thumbnail*) per ciascuna pagina
- Platform neutral, disponibili nei browser via plugin
- Entrambi preservano l'aspetto dei documenti stampati
- E' facile convertire in tali formati un documento nato digitale

DjVu

- Tecnologia scanner-web
- `http://www.djvuzone.org/`

Pro:

- Compressione ottimale per il Web
- Documenti lunghi scarcabili velocemente
- Plug-in DjVu integrabile in applicazioni sw
- Server Any2DjVu disponibile gratuitamente

Contro:

- Poco diffuso in confronto a PDF
- Nessuno standard per i metadati nei documenti

Che cos’è un file PDF?

PDF sta per **P**ortable **D**ocument **F**ormat.

- Formato binario, universalmente diffuso
- Mantiene stabile l'aspetto grafico (layout)
- Meccanismi di sicurezza (password, ecc.)
- Compatibile su più piattaforme (Windows, Mac, ecc.)
- Quando usare PDF?
  - Per conservare il layout
  - Per la fruibilità multi-piattaforma
  - Per difendere il contenuto
  - Per documenti lunghi

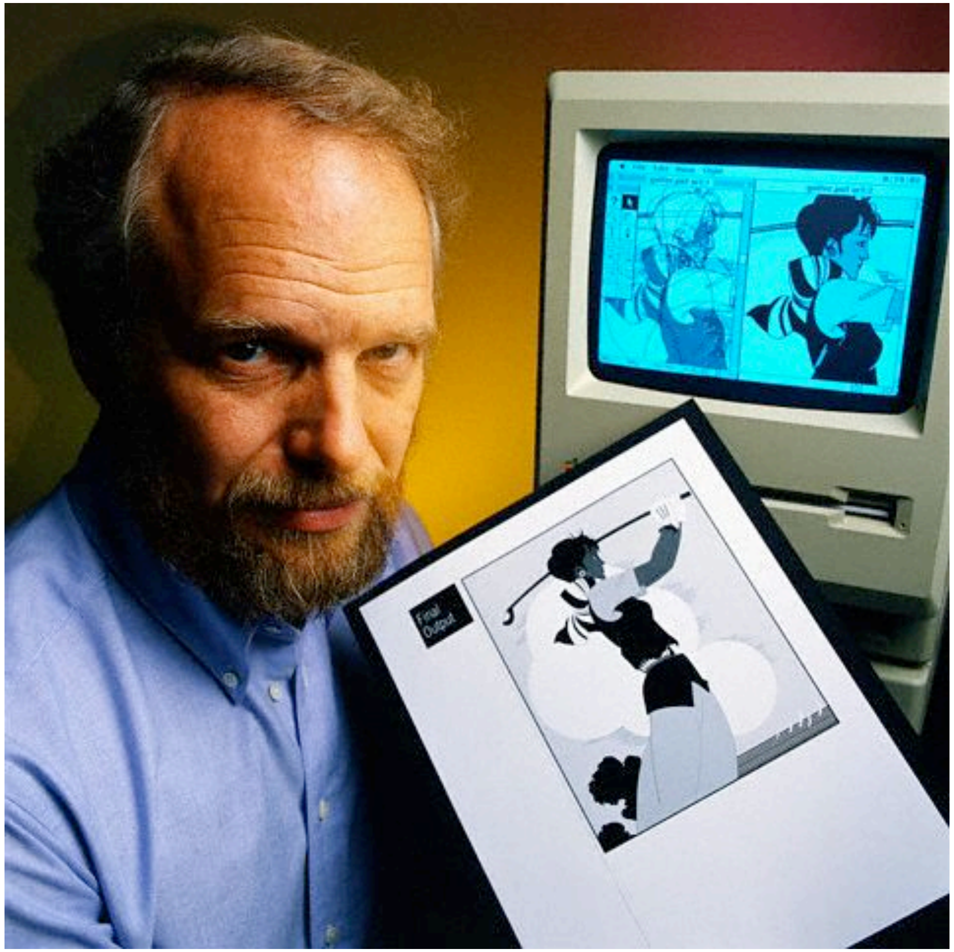
PDF

- Standard per publishing di documenti digitali
- Formato che preserva fonti, layout e colore del documento sorgente
- Ideale per documenti scientifici con simboli particolari o per documenti multilingua
- Formato compresso, adatto alla trasmissione su rete
- Molti software (alcuni a pagamento) sono capaci di produrlo;
- Software di lettura gratuito (Adobe Reader) presente su tutte le piattaforme

PDF: l'idea iniziale

Our vision for Camelot (=Acrobat) *is to provide a collection of utilities, applications, and system software so that a corporation can effectively capture documents from any application, send electronic versions of these documents anywhere, and view and print these documents on any machines.*

J.Warnock, 1991



PDF Portable Document Format

Pro:

- Molto diffuso, perché molto versatile e ricco di funzioni
- Base di PDF-Archive, uno standard di archiviazione di AIIM (Association for Information and Image Management)
- Manuale e software di sviluppo disponibili gratis da Adobe
- Offre un metodo standard per metadati: XMP Standard (Extensible Metadata Platform), compatibile con Semantic Web

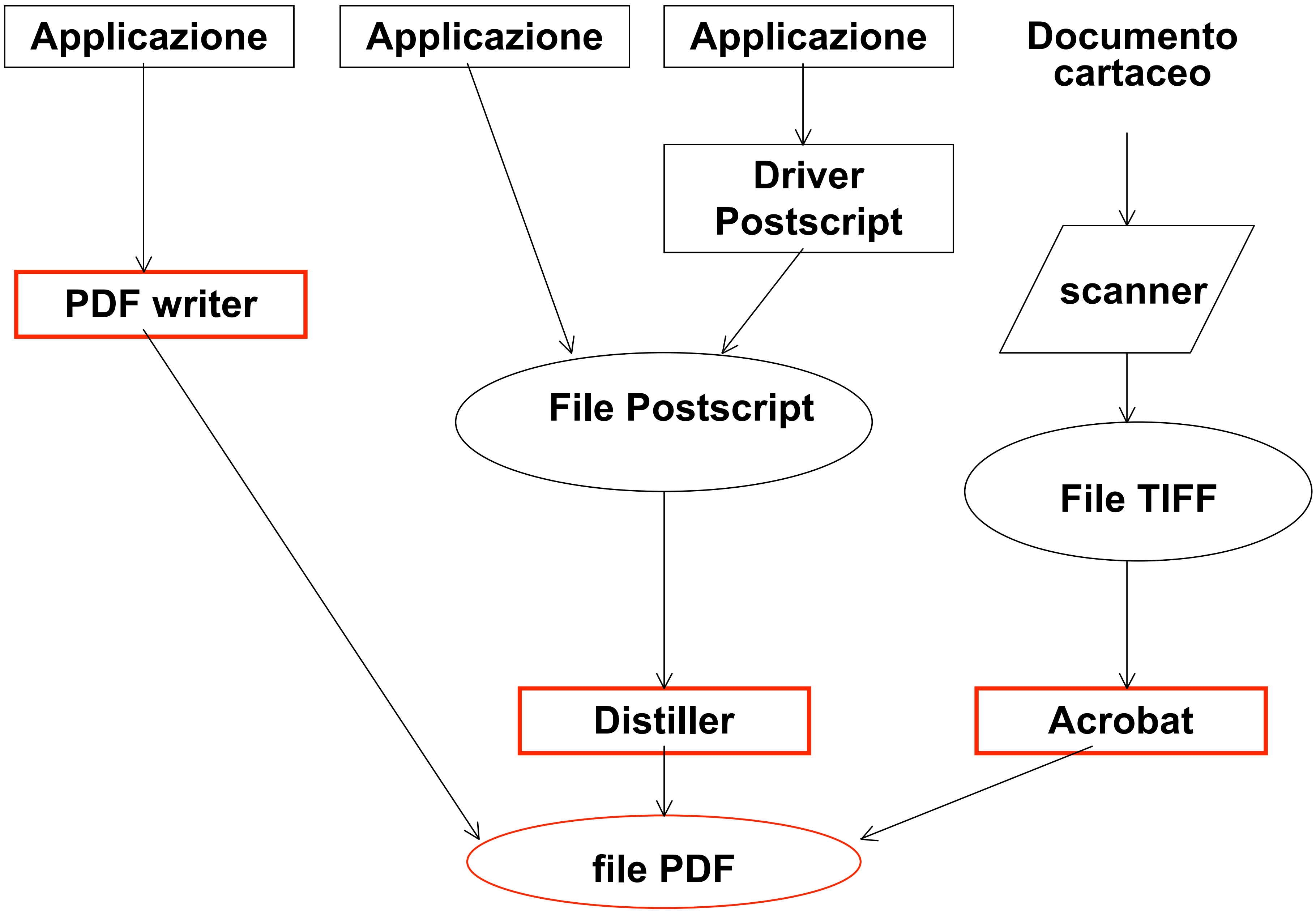
Contro:

- Molte versioni dei reader, con funzioni diverse: gli utenti si possono confondere
- Prestazioni Web basate su plug-in lente per documenti lunghi
- Un file PDF fatto di immagini da scanner può essere molto grande, anche per documenti piccoli
- Acrobat è uno strumento costoso

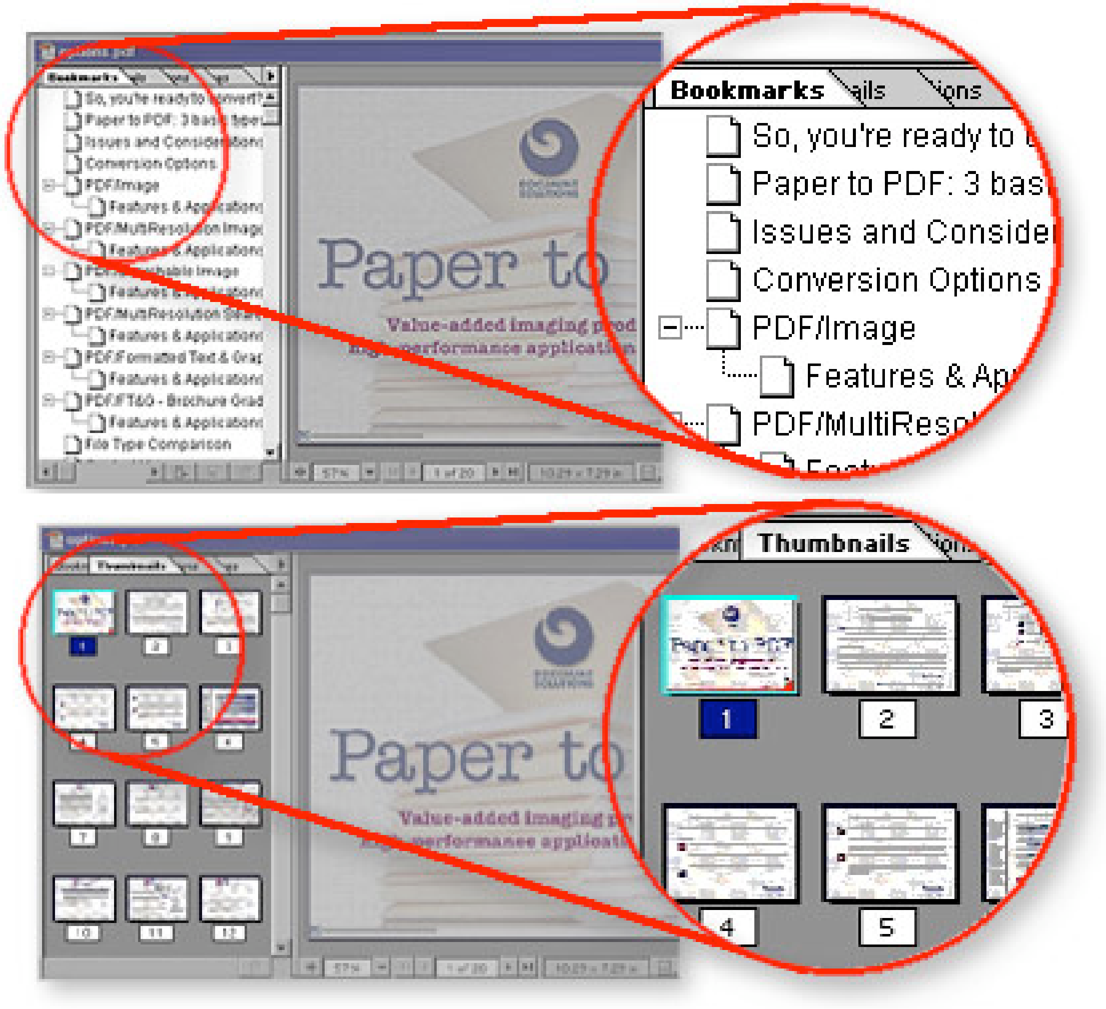
Come si crea un PDF

Quattro modi:

- Esportazione diretta da applicazione
- Distiller, via Postscript
- PDFWriter (driver stampa virtuale)
- Da scanner, con Acrobat


Applicazione
Applicazione
Applicazione
Documento cartaceo
Driver Postscript
PDF writer
scanner
File Postscript
File TIFF
Distiller
Acrobat
file PDF

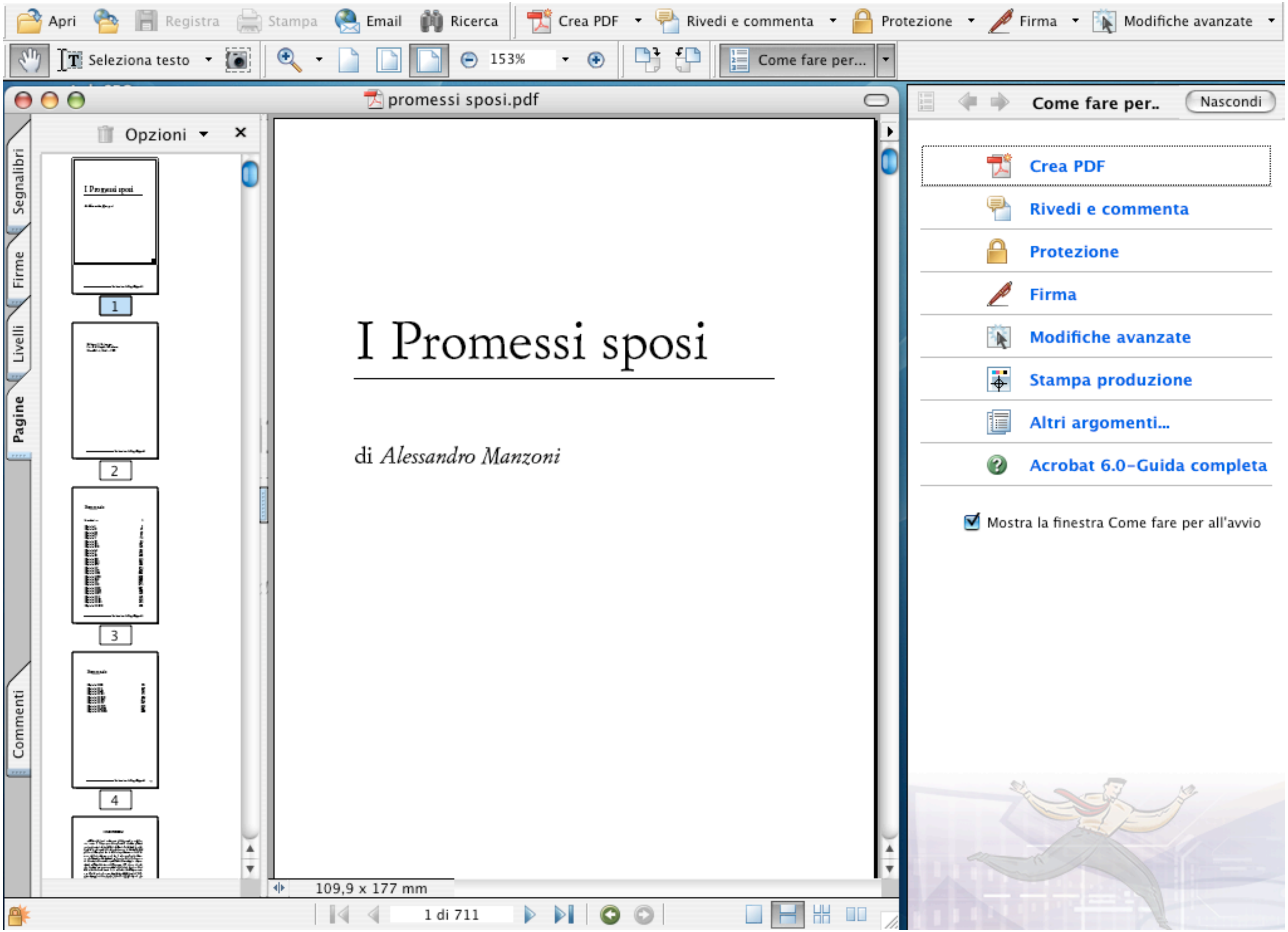
Accessori del file PDF



Storia di PDF e Acrobat

- 1991: Adobe presenta Interchange Postscript
- 1992: Adobe pubblica lo standard PDF1.0
- 1993: Acrobat 1.0
- 1994: Acrobat 2.0 e standard PDF1.1
- 1996: Acrobat 3.0 e standard PDF1.2
- 1998: Standard PDF/X (prepress data interchange)
- 1999: Acrobat 4.0 e standard PDF1.3
- 2000: Illustrator e standard PDF1.4
- 2001: Acrobat 5.0
- 2003: Acrobat 6.0 e standard PDF1.5
- 2004: Acrobat 7.0 e standard PDF1.6
- 2005: PDF/A (PDF1.4) diventa standard ISO (per archiviazione)
- 2006: Acrobat 8.0 e standard PDF1.7

Acrobat


Apri
Registra
Stampa
Email
Ricerca
Crea PDF
Rivedi e commenta
Protezione
Firma
Modifiche avanzate
Seleziona testo
153%
Come fare per...
promessi sposi.pdf
Opzioni
Segnalibri
Firme
Livelli
Pagine
Commenti
I Promessi sposi
di Alessandro Manzoni
109,9 x 177 mm
1 di 711
Come fare per..
Nascondi
Crea PDF
Rivedi e commenta
Protezione
Firma
Modifiche avanzate
Stampa produzione
Altri argomenti...
Acrobat 6.0-Guida completa
Mostra la finestra Come fare per all'avvio

Acrobat

- Adobe Acrobat è l'applicazione di riferimento per manipolare file PDF
- Esistono al momento (2006) due versioni principali
  - Acrobat Reader 7.0 (gratuita, solo lettura)
  - Acrobat 8.0 (a pagamento), nelle versioni
    - Acrobat Elements
    - Acrobat Standard
    - Acrobat Professional
    - Acrobat 3D

Esempio di domanda del Web test

- Un documento digitale…
  - È un file di Word
  - È un documento in codice ASCII
  - È un contenitore di testo e grafica
  - È un documento rappresentato in un codice binario

Esempio di domanda del Web test

- Quali di questi sono formati proprietari?
  - HTML
  - Doc
  - PDF
  - ASCII

Esempio di domanda del Web test

- Quali di questi formati sono più adatti per immagini fotografiche?
  - TIFF
  - GIF
  - JPEG
  - ASCII

Domande?