L'informazione: codifica ed elaborazione

Appunti di

Antonio BERNARDO

Corso di Informatica di base

Informatica

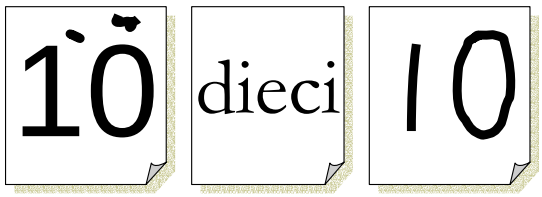
- Studio sistematico degli algoritmi che descrivono e trasformano l'informazione: la loro teoria, analisi, progetto, efficienza, realizzazione e applicazione

ACM (Association for Computing Machinery)

- Scienza della rappresentazione e dell'elaborazione dell'informazione

La tecnologia dei calcolatori occupa un ruolo marginale rispetto a quello delle tematiche dell'elaborazione delle informazioni.

Supporti e codifica



A. Bernardo, Informatica di base 3

I fogli portano l'informazione ma non sono l'informazione.

Fogli e macchie costituiscono il supporto fisico per ciò che ci interessa.

L'informazione è disponibile solo in quanto è accessibile il supporto su cui essa è mantenuta.

Vi sono supporti che sono in grado di memorizzare l'informazione, altri supporti (aria) sono particolarmente adatti a trasportare informazione.

L'informazione può essere codificata in modi differenti e su supporti differenti.

La codifica è l'operazione con cui l'informazione viene trascritta su un supporto.

La decodifica è l'operazione con cui l'informazione viene letta da un supporto.

I livelli dell'informazione

- Livello sintattico
- Livello semantico
- Livello pragmatico

Livello sintattico quando ci si chiede se e come un certo supporto fisico è in grado, in base alle configurazioni che può assumere può costituire la base per scrivere un messaggio. Ci si pone il problema di studiare possibili configurazioni del sistema: macchie di inchiostro su un foglio, lampadine accese o spente in un'insegna, segnali elettrici, elettromagnetici, stato magnetico, …

Al **livello semantico** ci si pone il problema del significato da attribuire a una certa configurazione del sistema fisico che fa da supporto per la memorizzazione o la trasmissione dell'informazione. Ci si pone quindi un problema di relazioni tra segni e significati.

A **livello pragmatico** ci si pone il problema tra significati e valori.

Se un messaggio viene scritto in una lingua che non è nota al destinatario, la quantità di informazione semantica e quella pragmatica è nulla, non è nulla invece la quantità di informazione sintattica.

L’informazione sintattica

- 1928 R. Hartley – 1948 C. Shannon
- Adeguatezza del supporto
- Accuratezza della trasmissione

A partire dal 1920 è nata la teoria dell’informazione, nel senso del livello sintattico di cui si è detto. Un certo supporto può essere utilizzato per memorizzare una certa quantità di informazione? Con quale velocità un certo supporto è in grado di trasferire informazione? Con quale grado di accuratezza un messaggio viene trasmesso? Per tutte queste tematiche è sufficiente restare al livello sintattico della comunicazione.

Claude Shannon, in un famoso articolo del 1948, ha proposto di utilizzare il concetto di scelta (o decisione) per misurare la quantità di informazione contenuta in un messaggio.

Per semplificare questa misura, Shannon suggerisce di ridurre ogni scelta a una successione di *scelte binarie*.

Per esempio, il semaforo ha tre lampadine, ciascuna delle quali può essere accesa o spenta. La nostra scelta è dunque fra otto alternative: 1. tutte e tre le lampadine spente; 2. verde acceso; 3. giallo acceso; 4. rosso acceso; 5. verde e giallo accesi; 6. giallo e rosso accesi; 7. verde e rosso accesi; 8. verde, giallo e rosso tutti accesi. Il semaforo, in un certo senso, *rappresenta* una scelta fra otto alternative in termini di tre scelte binarie, cioè di tre scelte (ognuna rappresentata da una lampadina) fra due sole alternative (acceso o spento). La scelta fra *n* alternative diverse (nel nostro esempio 8) corrisponde a *log(2) n* (nel nostro esempio, log(2)(8)=3) scelte binarie.

Per capire meglio questo meccanismo pensate a un gioco: supponiamo che dobbiate indovinare il numero (intero) pensato da un amico avendo a disposizione domande alle quali il vostro amico può rispondere solo 'sì' o 'no'.

La codifica binaria

- L’alfabeto più semplice che si può adottare è costituito da due soli simboli solitamente indicati con

  0 1

- Dispositivi bistabili – bit (binary digit)
- Bit – unità elementare di informazione

I dispositivi bistabili sono in grado di assumere una configurazione a scelta tra due.

Informazione analogica / digitale

- Nella forma analogica una grandezza è rappresentata in modo continuo.
- Nella forma digitale una grandezza è rappresentata in modo discreto.



Le grandezze percepite dai nostri sensi sono tipicamente di natura analogica.

Un segnale analogico è una grandezza che varia in modo continuo nel tempo o nello spazio o in entrambi i domini.

Sia il valore che misuriamo, sia gli istanti (nello spazio o nel tempo) in cui possiamo eseguire le misurazioni variano in modo infinitesimale.

Suono: variazione continua, nel tempo, della pressione dell'aria.

Foto in bianco e nero: variazione continua, nello spazio, della luminosità.

Filmato in bianco e nero: variazione continua, nello spazio e nel tempo, della luminosità.

Standard IEC per i prefissi binari

- Con k bit si possono rappresentare 2^k informazioni
- 8 bit → 1 byte
- I multipli del byte sono

nome	simbolo	IEC (RAM)	SI (Hard Disk)
Chilo	KB	$2^{10} = 1024$	10^3
Mega	MB	$(2^{10})^2$	$(10^3)^2$
Giga	GB	$(2^{10})^3$	$(10^3)^3$
Tera	TB	$(2^{10})^4$	$(10^3)^4$

IEC International Electrotechnical Commission

1KB corrisponde a circa 1.000 byte

1MB corrisponde a circa 1.000.000 byte

1GB corrisponde a circa 1.000.000.000 byte

Codifica binaria di dati numerici

Sistema decimale	Sistema binario
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001

10	1010
11	1011
12	1100
13	1101
14	1110
15	1111
16	10000
17	10001
18	10010
19	10011

20	10100
21	10101
…	
…	
…	
99	
100	
101	
102	
…	

Trasformazione di base

- $2435 = 5*10^0+3*10^1+4*10^2+2*10^3$ base 10
- $11011 = 1*2^0+1*2^1+0*2^2+1*2^3+1*2^4$ base 2

$= 31_{(10)}$

Come si trasforma 53 nel sistema binario:

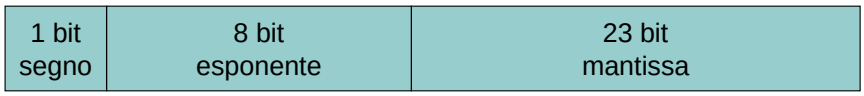
- 53:2=26 resto 1
- 26:2=13 resto 0
- 13:2=6 resto 1
- 6:2=3 resto 0
- 3:2=1 resto 1
- 1:2=0 resto 1
- 53 =110101

Operazioni nel sistema binario

	0	1	0	1	1	+						1	0	1	*		
	1	0	0	1	1	=							1	1	=		
	1	1	1	1	0							1	0	1			
											1	0	1				
											1	1	1	1			
	1	1	0	1	1	0	1	0	+								
	1	1	1	0	0	1	1	1	=								
1	1	1	0	0	0	0	0	1									
o	**v**	**e**	**r**	**f**	**l**	**o**	**w**										
		r	i	s	u	l	t	a	t	o		e	r	r	a	t	o

Standard per la codifica dei numeri

- Numeri interi relativi
  - rappresentazione per segno e modulo
  - complemento a 2, utile per semplificare il calcolo
- Numeri razionali - standard IEEE754
  - A virgola mobile in notazione esponenziale $R=M*10^e$



Precisione singola

Precisione doppia

Per rappresentare un numero con segno e modulo occorre 1 bit per il segno e un certo numero di bit per il numero.

Per rappresentare un numero con segno nella notazione con complemento a 2 si complementano a 2 i bit della rappresentazione binaria del numero (cioè si modifica 1 in 0 e viceversa 0 in 1) e si aggiunge 1 al risultato ottenuto:

Utilizzando per esempio 3 bit

Intero in base 10	con segno e modulo	complemento a due
- 3	111	101
+3	011	011

Standard per la codifica binaria del testo – ASCII 7 bit

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
010	Space	!	“	#	$											
011	0	1	2	3	4	5	6	7	8	9	:	;				
100	@	A	B	C	D	E	F	G								
101	P	Q	R													
110	‘	a	b	c	d	e	f	g	h	i	j	k	l	m	n	O
111	p	q	r	s												Canc

a = 1100001

A. Bernardo, Informatica di base 13

ASCII è l'acronimo di **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange (ovvero *Codice Standard Americano per lo Scambio di Informazioni*).

Lo standard ASCII utilizza 7 bit, oggi vengono utilizzati standard a 8 bit e quindi 1 byte per rappresentare un carattere di testo, spazio e punteggiatura inclusa.

8 bit non sono sufficiente nemmeno a rappresentare i caratteri utilizzati in Europa (alfabeto cirillico, greco, …)

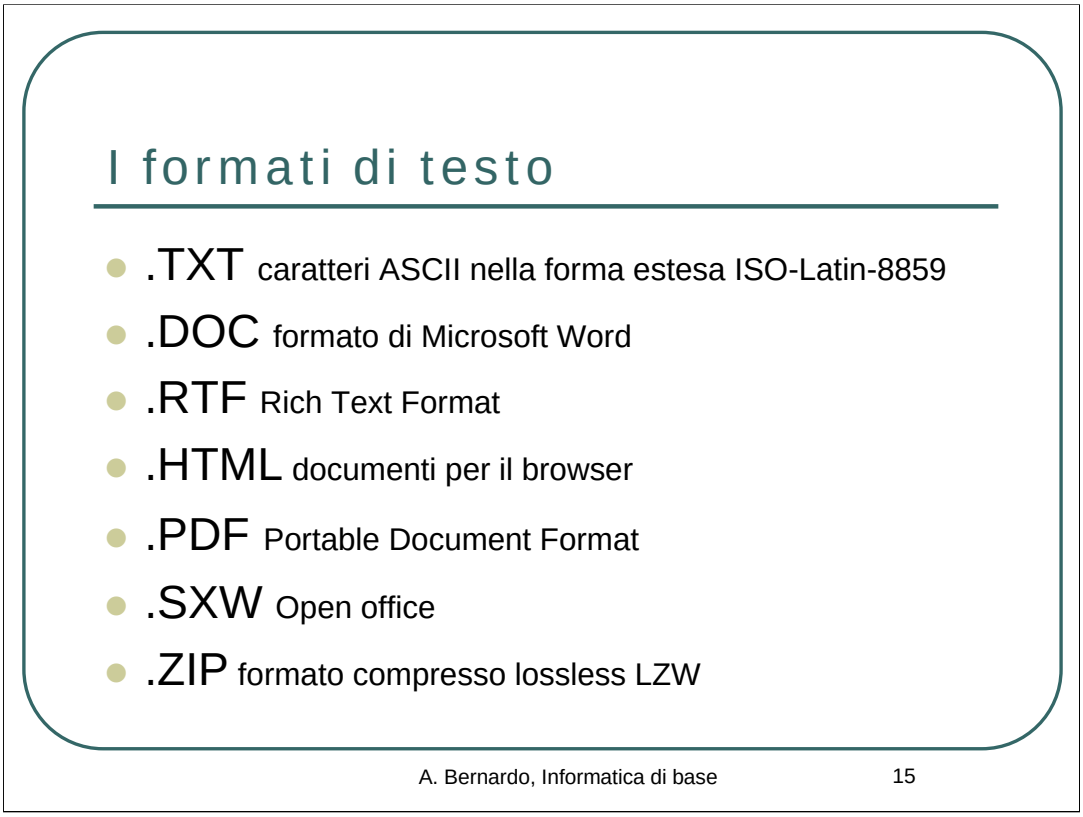
Lo standard UNICODE prevede 16 bit per rappresentare tutti i caratteri utilizzati nel mondo.

Il problema di definire gli standard per la codifica dei caratteri, attualmente i caratteri in uso sono oltre 200.000, è praticamente irresolubile, perché mentre nelle lingue a simboli alfabetici le nuove parole sono costituite dagli stessi simboli dell’alfabeto, nelle lingue come il cinese e il giapponese, i nuovi termini hanno nuovi simboli.

I caratteri vengono classificati in tre categorie: caratteri di comando descrivono codici di trasmissione o di controllo delle stampe; caratteri alfanumerici descrivo le lettere e le cifre; simboli includono i separatori per la punteggiatura.

Quanta memoria occupa il testo di un libro?

- Per un libro con 500 pagine
- Con 50 righe per pagina
- 80 caratteri per riga
- Occorrono 2.000.000 byte ossia circa 2MB



Il formato **.txt** si basa semplicemente sull'uso dei caratteri ASCII, usualmente nella forma estesa ISO-Latin-8859 che consente la rappresentazione delle lettere accentate. Con questo formato si ha il massimo della portabilità ma anche la minima capacità espressiva poiché non è possibile impiegare arricchimenti grafici né includere immagini.

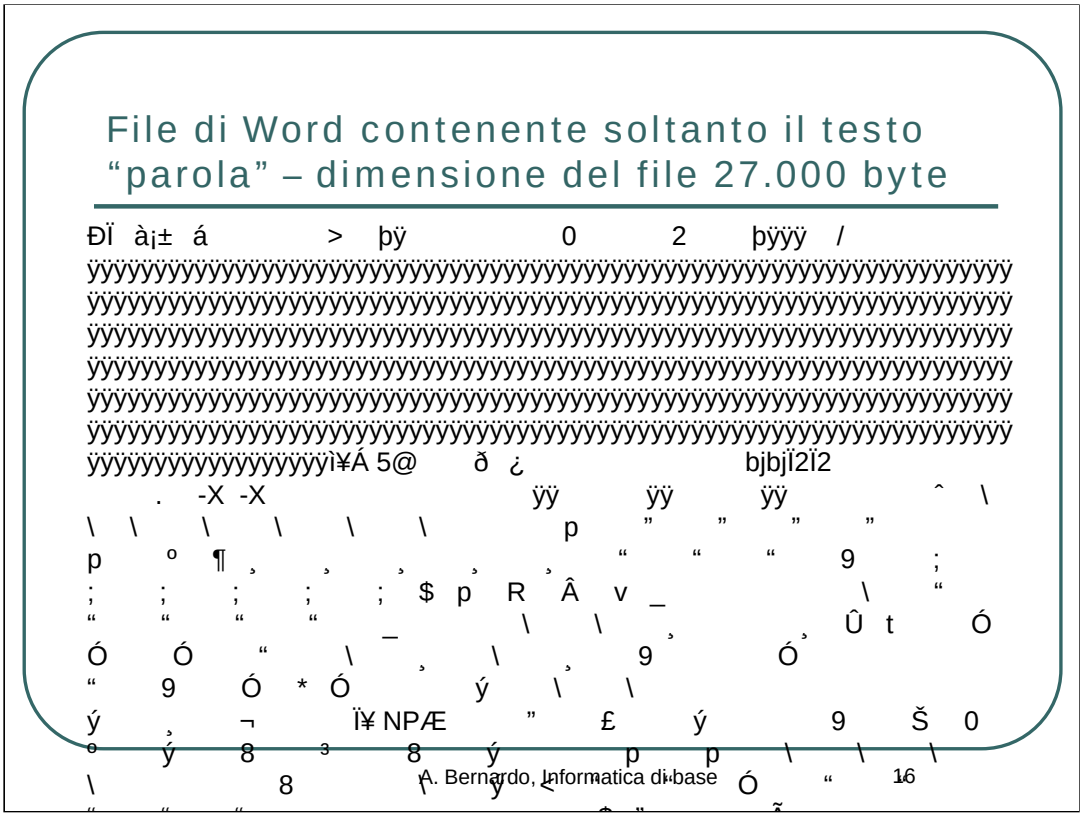
Il formato **.doc** generato da Microsoft Word permette di personalizzare il testo con numero caratteristiche di formattazione. Lo svantaggio è che si tratta di un formato proprietario le cui specifiche non vengono diffuse pubblicamente e la casa produttrice ha il diritto di cambiarle a propria discrezione. Questo formato può essere letto anche con altri programmi, ma senza la garanzia di interpretarlo correttamente. Un file con una sola parola occupa una quantità di memoria superiore a 20kb. Se aprite il file con il blocco note troverete, oltre alla parola che avete scritto, numerosissimi simboli la maggior parte incomprensibili per chi non conosce le specifiche del formato, altri indicano la versione del programma il font utilizzato, il proprietario, ...

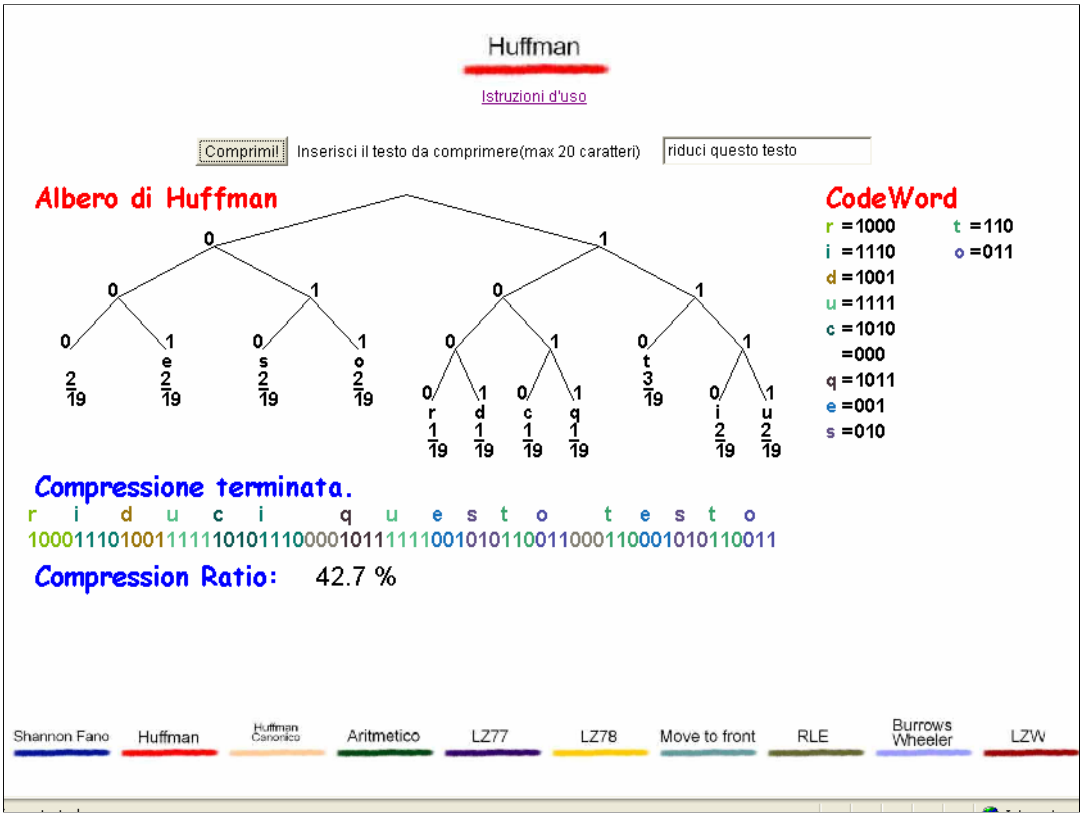
Il formato **RTF** è uno standard formalizzato dalla Microsoft per specificare la formattazione dei documenti con l'intento di scambiarli con altri sistemi di word processing. I file RTF sono file ASCII che oltre al testo contengono comandi speciali che indicano come debba essere formattato il testo (tipo di font, margini delle pagine, …). Il formato RTF rappresenta il modo corretto per distribuire documenti realizzati con Word.

Il formato **.html** (HyperText Markup Language) è quello dei documenti che vengono visualizzati con i browser (Netscape, Internet Explorer, Mozilla, …). E' un linguaggio per la marcatura del testo è aperto e infatti non è associato a nessun programma in particolare, può essere letto e modificato con un qualsiasi editore di testo. Lo standard HTML è stabilito dal Consorzio W3C (The World Wide Web Consortium).

PDF è un linguaggio per la specificazione di documenti da stampare con particolare attenzione a preservare l'aspetto e l'integrità di documenti. **U**tilizzare un unico standard per documenti di diversi tipi (Web, word processor, fogli di lavoro, scanner, fotografie e grafica) mantenendo le caratteristiche originali. **F**irma digitale per la certificazione di autenticità. **G**estione di permessi di accesso e modifica per documenti riservati. **A**ccessibilità per disabili. Il linguaggio PDF è stato realizzato da Adobe nel 1990. La diffusione è dovuta anche alla strategia della casa produttrice che ha realizzato due prodotti: un software gratuito (Acrobat Reader) per la lettura dei documenti e un software a pagamento per la produzione dei documenti.

File di Word contenente soltanto il testo “parola” – dimensione del file 27.000 byte





Questo algoritmo si fonda su alcuni studi effettuati sulla frequenza delle lettere nelle parole.
In un libro scritto in italiano, troveremo sicuramente una frequenza molto alta di lettere A rispetto a lettere Z o U che si presentano meno. Partendo da queste semplici considerazioni, si cerca quindi di codificare le lettere con maggior frequenza come la A con dei codici binari più brevi rispetto a quelli utilizzati per la Z. In questo modo le lettere più frequenti in un testo (che sul pc vengono rappresentate sempre con 8 bits) verranno rimpiazzate da codici di bit più corti.

In un normale file testo, ogni lettera è rappresentata da 8 bit codificati rispettando il codice ASCII. Le parole “riduci questo testo” richiederebbero 19 byte ossia 152 bit:

Adesso applichiamo l'algoritmo di Huffman per calcolare dei nuovi codici da assegnare alle lettere dell'esempio. Il metodo usato è piuttosto semplice. Si deve costruire un albero in cui le lettere più frequenti siano posizionate più vicino alla radice rispetto a quelle con minore frequenza.
Per prima cosa è necessario contare la frequenza di ogni lettera nella nostra stringa:
r(1) i(2) d(1) u(2) …
A questo punto vanno individuate le 2 lettere con minore frequenza che useremo per costruire un nuovo nodo alla base dell'albero. Troviamo nuovamente le due lettere con minore frequenza e facciamo la stessa cosa...
Una volta creato l'albero, bisogna associare ad ogni nodo un bit. E' possibile associare lo 0 a tutti i nodi di sinistra e 1 a tutti quelli di destra. E' possibile anche fare il contrario senza compromettere il risultato finale. Nel nostro caso scegliamo lo 0 a sinistra e l'1 a destra ed ecco il risultato finale.

La codifica delle immagini

- Mappa di pixel – formato bitmap: l'immagine viene memorizzata come insieme di pixel, ognuno con le sue caratteristiche (colore, luminosità, trasparenza) raggruppate in una matrice di pixel
- Descrizione geometrica – formato vettoriale: l'immagine viene archiviata in forma di descrizione geometrica e una serie di caratteristiche di apparenza (dimensione della penna, colore, tipo di linea, …)

Le immagini vengono divise in punti, detti pixel, ciascun punto viene codificato con un numero che corrisponde a un particolare colore.

Le immagini vengono codificate come sequenze molto lunghe di bit; per interpretare queste sequenze occorre conoscere le dimensioni dell'immagine (il numero di pixel della base e il numero di pixel dell'altezza), la **risoluzione** che si misura in **dpi**, dot per inch, cioè punti per pollice quadrato e il numero di colori.

L'immagine vettoriale ha il pregio di poter essere raffigurata a qualsiasi risoluzione. La descrizione vettoriale si adatta però a figure geometriche o immagini ottenute per mezzo di forme geometriche elementari.

Paint produce immagini bitmap – vedi esempio

Word e Power Point producono immagini vettoriali – vedi esempio

Immagine bitmap con PAINT

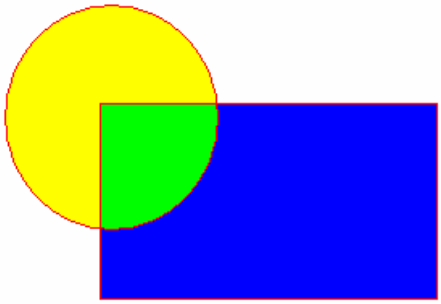
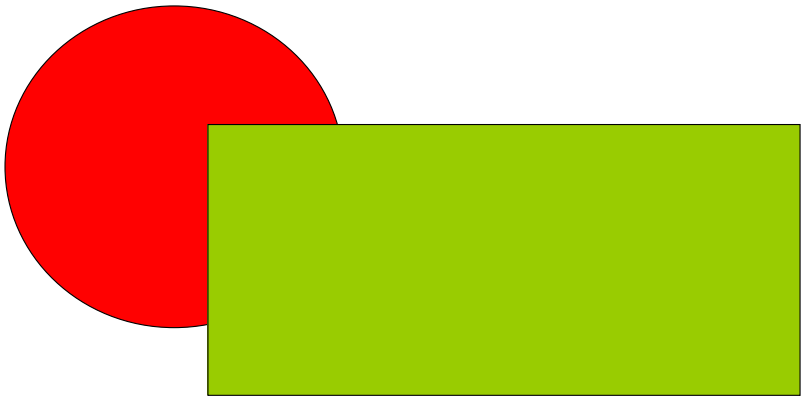


Immagine vettoriale con PowerPoint



Sistema additivo e sottrattivo dei colori

- RGB (Red Green Blu) per visualizzare immagini su schermo
- CMYK (Ciano-azzurro Magenta-rosso Giallo –Black) per stampare le immagini in quadricromia

Pensiamo ora ad un tipico file grafico, ad esempio una fotografia a colori salvata in modalità RGB. Senza addentrarci in discorsi complessi sulla codifica del colore, diremo che questa modalità richiede l'uso di tre **byte** per memorizzare le informazioni di colore e luminosità relative a ciascun punto- o **pixel**-dell'immagine. Per sapere, quindi, quanto spazio occupa su disco questo file grafico, ci basta moltiplicare per tre il numero complessivo di punti da cui è costituita l'immagine. Questo numero si ottiene, a sua volta, moltiplicando il numero delle colonne per il numero delle righe che compongono il rettangolo occupato dall'immagine.

Data allora una foto di 1024 pixel orizzontali per 768 verticali, lo spazio che essa occupa come file su disco è di 1024 x 768 x 3 = 2.359.296 byte, il che equivale a 2.304 Kilobyte o, ancora, a 2,25 Megabyte. E' importante precisare che tale cifra rappresenta lo spazio occupato dal file

grafico in forma **non compressa**.

Per la codifica dei colori vedi esempio in html con Frontpage.

Codice colori in html

```
<html>
<head>
<title>Codifica colori</title>
</head>
<body>
<p><font color="#FF0000">RED</font></p>
<p><font color="#00FF00">GREEN</font></p>
<p><font color="#0000FF">BLU</font></p>
<p><font color="#000000">NERO</font></p>
<p><font color="#FFFFFF">BIANCO</font></p>
<p><font color="#909090">GRIGIO</font></p>
<p><font color="#FFFF00">GIALLO</font></p>
</body>
</html>
```

Formati di immagini

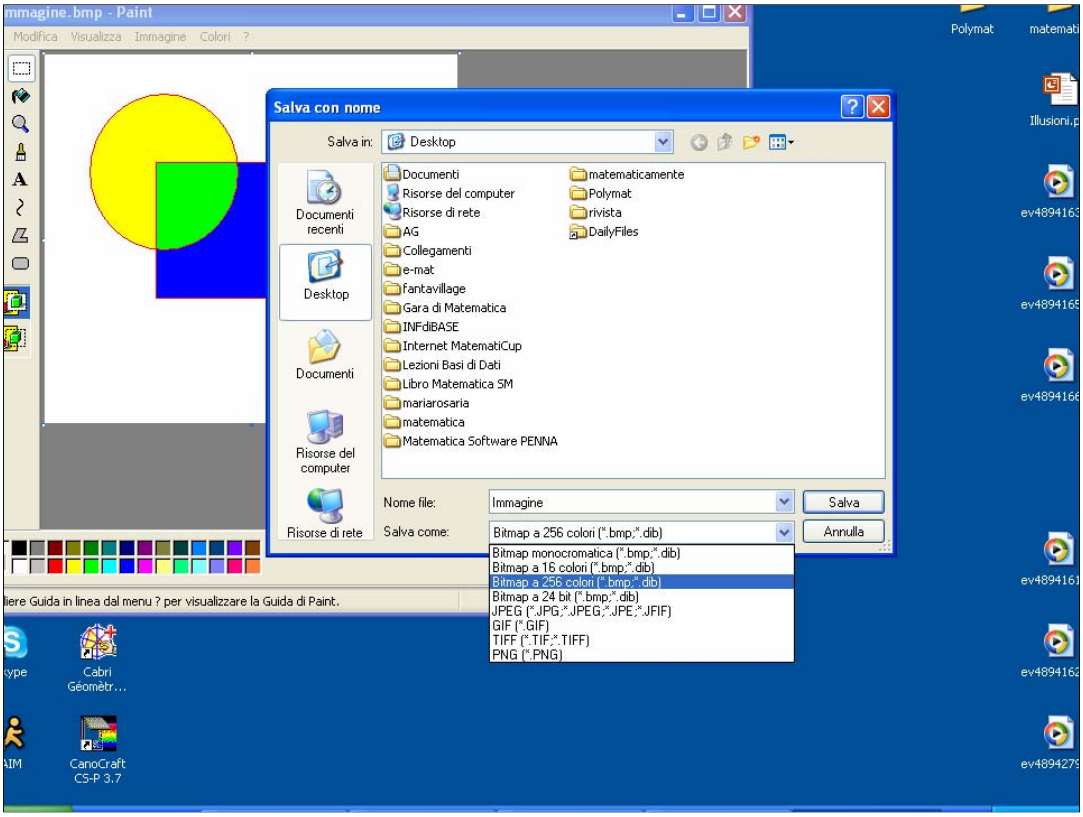
Come per i caratteri si è cercato di stabilire alcuni standard per la codifica delle immagini

- GIF è un formato proprietario permette di archiviare in formato compresso immagini con colori a 8 bit per pixel, ossia 256 colori, permette la codifica di immagini animate, usa un algoritmo di compressione lossless (vedi nota) ossia senza perdita di informazioni
- BMP formato per sistemi operativi Microsoft e OS/2 utilizzato ormai in tutti i sistemi operativi permette di archiviare con diversa profondità di colore: 1, 4, 8, 24 bit
- PNG dovrebbe sostituire il formato GIF ma non ammette l'animazione
- TIFF nato per semplificare lo scambio di immagini tra scanner e programmi applicativi è applicato in diversi ambiti professionali è senza perdita di informazioni (lossy)
- JPG formato di archiviazione con perdita di informazione, può comprimere con un rapporto elevatissimo, viene visualizzata in RGB a 24 bit per pixel

Un formato che è in grado di restituire, al termine della decompressione, un'immagine esattamente uguale – pixel per pixel- all'originale com'era prima che venisse compresso, viene normalmente definito lossless, in italiano potremmo tradurre **senza perdita** oppure **non distruttivo**. Viceversa, un formato di compressione che non può assicurare una reversibilità assoluta, viene definito in inglese lossy, ovvero, in italiano, **con perdita** o anche **distruttivo**. La cosa che si perde o non si perde è la fedeltà all'originale dell'immagine ripristinata.

I grafici e gli impaginatori di professione devono conoscere perfettamente le caratteristiche dei formati di compressione che adoperano, se vogliono ottenere il meglio dalle manipolazioni che effettuano sui file. Sarebbe infatti un grave errore salvare e risalvare un file in un formato lossy

come il JPG, per poi utilizzarlo alla fine in un formato "senza perdita" come il TIF. E' invece corretto fare il contrario, ovvero salvare quante volte si vuole un lavoro in corso d'opera in un formato non distruttivo, per poi salvarlo solo alla fine, se necessario, in un formato distruttivo. La regola (e la logica) vuole, insomma, che l'archiviazione in un formato lossy sia sempre l'anello conclusivo della catena di trasformazioni a cui è sottoposto un file.



Formati in cui si può salvare un’immagine realizzata con Paint

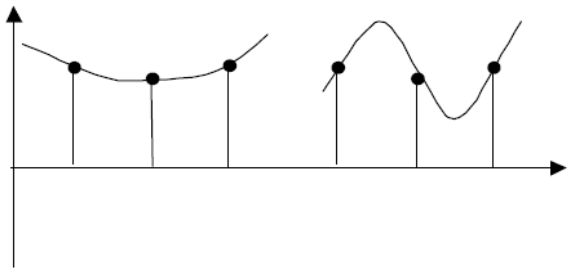
File audio

Il suono è un segnale continuo, per essere memorizzato deve essere campionato ottenendo così un segnale digitale. I parametri che caratterizzano il campionamento sono

- **1. Il numero di canali:** la modalità mono ha uno solo canale mentre quella Stereo ne ha due separati, sinistro e destro. Ovviamente un segnale stereo occupa il doppio di uno mono. Nelle applicazioni più recenti il numero di canali è notevolmente aumentato (surround).
- **2. La risoluzione:** Rappresenta il numero di bit utilizzati per rappresentare i campioni. Solitamente si utilizzano 8 o 16 bit per campione, nel primo caso si hanno 256 valori possibili, relativamente pochi perché offrono una qualità del suono inferiore a quella di un nastro, nel secondo si hanno circa 65.000 valori.
- **3. La frequenza di campionamento:** È il numero di campioni al secondo, può variare da 11 khz adatta alla registrazione della voce, a 22 khz adatta alla registrazione di un nastro fino a 44 khz per una registrazione a qualità cd. Questo parametro deve esser scelto in modo da poter ricostruire il segnale originario da quello campionato.

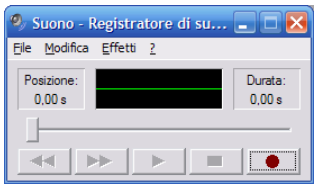
Il campionamento

- Segnali diversi possono dare luogo allo stesso segnale campionato, com'è mostrato in figura.



File audio e compressione

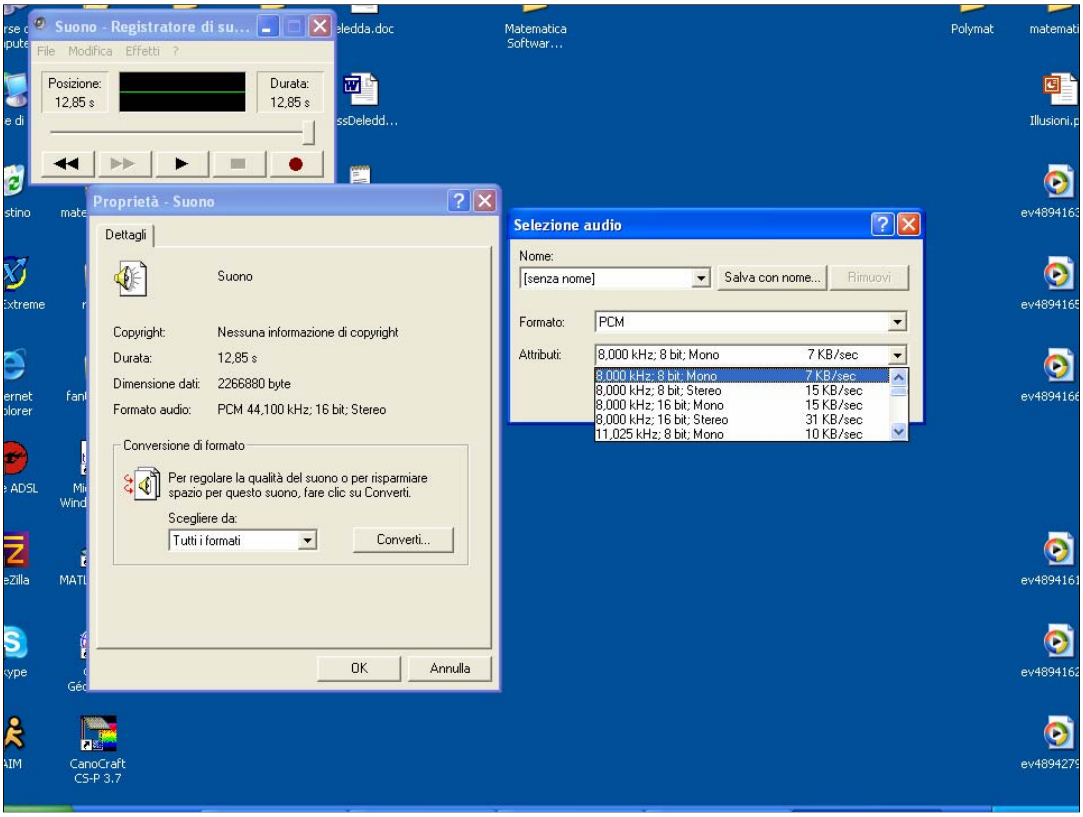
- Analogica (dischi in vinile)
- Digitale (CD anni ’80)
  - WAVE (formato non compresso)
  - MP3 (formato compresso)
  - MIDI (formato vettoriale)
    - Streaming audio



Streaming audio: RAM, RM, ASF, ASX

Lo streaming è il trasferimento in rete dei dati audiovisivi in tempo reale; tutto questo senza tempi d’attesa derivanti dal download completo del file sull’Hard Disk del computer. Con lo streaming, infatti, non viene scaricato l’intero file audio prima di consentirne l’ascolto, ma la riproduzione inizia per ogni blocco di due secondi d’ascolto; nel frattempo viene scaricato il successivo. Si possono verificare momentanee interruzioni nella riproduzione, nel caso in cui il traffico nella rete risulti congestionato. Le due principali tecnologie d’audio streaming utilizzate sono Real

(attraverso il real player), e Windows Media (Windows Media player). La tecnologia dello streaming audio ha permesso alle principali emittenti radiofoniche di presentare sui propri siti web i loro programmi trasmessi via etere, cosicché chiunque potesse ricevere una trasmissione “radio” dal Pc.



Formati dei file audio con il Registratore di suono di Windows