

L'elaborazione di documenti

Il documento e il documento elettronico

Un **documento** è una frase scritta in un qualche linguaggio che abbia un contenuto, una struttura, una semantica e che può essere in relazione con altri documenti.

Un **documento elettronico** è un documento la cui rappresentazione fisica è in forma di bit all'interno di un sistema informatico.

Esempi di documenti:

Testi debolmente strutturati:

romanzi, racconti, poesie, saggi, articoli, eccetera.

Testi fortemente strutturati: elenchi telefonici, schede cliniche, programmi di computer, eccetera.

Ipertesti:

testi contenenti collegamenti interni che ne permettono una lettura non sequenziale (vedi più avanti).

Non-testi:

immagini, schemi, progetti, fotografie, filmati, animazioni, eccetera.

Operazioni tipiche sui documenti

Creazione:

del contenuto, della struttura, dei metadati.

Cancellazione:

del contenuto, della struttura, dei metadati.

Acquisizione:

via tastiera, rete, scanner, penna ottica, eccetera.

Modifica:

aggiornamento, correzione, pulitura, montaggio, strutturazione, eccetera.

Condivisione:

stampa, trasmissione, conversione, eccetera.

Interazione:

esecuzione di istruzioni di I/O contenute all'interno del documento.

Navigazione:

accesso a documenti correlati mediante una delle relazioni che li collegano.

Componenti dei documenti e operazioni

I programmi di elaborazione di documenti e, in particolare, gli editor di testo, operano a vari livelli sui componenti di un documento, ad esempio:

carattere:

unità minima di rappresentazione, corrispondente solitamente ad un tasto della tastiera (ed ad un valore di un opportuno codice di rappresentazione). Operazioni tipiche sui caratteri sono: selezione del tipo di carattere (Arial, Times New Roman, Courier ecc) della dimensione (9pt, 10pt., 12 pt., 14 pt., ecc), dello stile (normale, **grassetto**, *corsivo* ecc) e del colore (**rosso**, nero, **verde** ecc) .

parola:

sequenza di caratteri divisa da separatori di parola dalle altre parole. Come separatori di parole si usano di solito spazi, fine linea, fine pagina, e segni di punteggiatura ma ci sono eccezioni (ad esempio nei linguaggi di programmazione). Conteggio, controllo ortografico e divisione delle parole (hyphenation) sono operazioni tipiche sulle parole.

frase:

sequenza di parole.

paragrafo:

frase delimitata da separatori di paragrafo. Un ritorno a capo, o CR (carriage return) è un separatore di paragrafo. Sistemi operativi diversi utilizzano separatori diversi per i paragrafi.

Ad esempio:

1. DOS e *Windows*: Carriage Return + Line Feed (ASCII 13 e ASCII 10)
2. Mac: Carriage Return (ASCII 13)
3. Unix: Line Feed (ASCII 10)

Queste differenze fanno sì che lo stesso documento, usato su sistemi operativi diversi anche con lo stesso tipo di applicazione, possa dare dei problemi di conversione.

I programmi di elaborazione testi inoltre usano introducono altre strutture quali sezioni, colonne, oggetti grafici, eccetera, che richiedono ulteriori separatori specifici.

Componenti dei documenti e operazioni

Per alcuni sistemi di elaborazione di documenti, orientati alla stampa, è importante il concetto di **pagina**.

Le pagine sono di dimensioni prefissate, ma esistono formati diversi a seconda degli scopi ed anche del paese. Ad esempio, in Europa esistono i formati A, definiti a partire dal formato A0 (84x118.8 cm). Il formato più comune è A4 (21x 29.7 cm). In USA invece sono più comuni i formati lettera e legale. Altri tipi di pagina possono essere usati per scopi particolari (buste, biglietti da visita, eccetera)

Per alcuni sistemi di elaborazione di documenti, orientati alla visualizzazione su schermo, è importante invece il concetto di **pagina visualizzabile** (o **schermata**). Dato che gli schermi sono di capacità diverse per quanto riguarda dimensioni, risoluzione, numero di colori. Una schermata che risulta adatto ad uno schermo può risultare inadeguata per un altro.

Ipertesti

Un testo può essere considerato come una sequenza di caratteri, strutturati opportunamente in parole, frasi, paragrafi ecc., che è inteso per un accesso ed una lettura sequenziale in una qualche direzione (ad esempio da destra a sinistra, dall'alto al basso) che dipende dalla cultura di riferimento.

Un **ipertesto** è un testo che contiene dei collegamenti diretti (detti *link*, o collegamenti ipertestuali) fra alcune sue parti. Tali collegamenti permettono un accesso ed una lettura non sequenziale dell'ipertesto, navigando attraverso di esso. Queste pagine sono un esempio di ipertesto: ad esempio, cliccando su una linea dell'indice ci si porta alla pagina che contiene la trattazione dell'argomento specificato nell'indice.

I documenti ipertestuali, esistenti da molto tempo, si sono diffusi negli ultimi anni grazie all'avvento del *World Wide Web* (WWW), dato che le pagine di un sito *Web* sono tipicamente documenti ipertestuali multimediali. In effetti lo stesso WWW può essere visto come un ipertesto multimediale distribuito.

I documenti ipertestuali vengono creati usando i cosiddetti linguaggi di marcatura (di *markup*). Questi sono dei formalismi che permettono di usare dei marcatori (detti anche tag) per definire la struttura logica del documento, le relazioni che legano documenti diversi e gli aspetti di presentazione del documento (stile, formattazione, eccetera). *Standard Generalized Markup Language* (SGML, ISO 8879) è un linguaggio per descrivere linguaggi che codificano documenti.

Fra i linguaggi di marcatura più diffusi ricordiamo HTML (che è una specifica istanza di SGML).

Uso di un editore di testi: funzioni di base

I programmi di videoscrittura (o editori di testi) sono fra i più diffusi programmi applicativi e permettono di scrivere, modificare, formattare documenti di vario genere,

I programmi per l'elaborazione di testi (o editori di testi), quali ad esempio *Microsoft Word* oppure *Corel WordPerfect*, sono di norma di uso estremamente semplice in quanto tutte le principali funzioni sono disponibili da opportuni menu ed inoltre è disponibile un manuale ed un sistema di aiuto on-line (anch'esso accessibile tramite menu). L'autoapprendimento di questi programmi applicativi è quindi abbastanza facile. Ricordiamo qui alcune caratteristiche comuni ai vari editori di testi.

Tutti gli editori di testi permettono le seguenti funzioni:

a capo automatico:

una volta che sono stati impostati i margini del documento (usando i comandi grafici su un opportuno righello) quando arriviamo alla fine di una riga il cursore, che indica la posizione nella quale verrà inserito il prossimo carattere, va a capo automaticamente, eventualmente dividendo in modo opportuno una parola.

modifiche al testo: per modificare una porzione di testo basta spostare il cursore (usando il mouse) nella zona nella quale si vuole fare la modifica e quindi aggiungere il testo (digitandolo da tastiera) o modificare quello esistente. Per cancellare del testo basta selezionare il testo che si vuole cancellare con il cursore (trascinando il mouse mentre si tiene premuto il pulsante sinistro del medesimo) e quindi usare il tasto *Delete* (Cancella) o *Backspace*. La parte selezionata del testo può essere copiata in un'altra parte del documento o in un altro documento usando le funzionalità di *Cut* (Taglia) e *Paste* (Incolla) disponibili dal menu *Edit* (Modifica).

Uso di un editore di testi: controlli e visualizzazione

controllo ortografico:

se questa funzionalità è attiva, le parole inserite nel testo sono confrontate con quelle presenti in un dizionario memorizzato nel computer. La lingua di riferimento ovviamente deve essere selezionata in precedenza, e questo può essere fatto dal menu *Tools* (Strumenti). Le parole che non sono presenti nel dizionario sono segnalate come possibili errori (di solito mediante una sottolineatura in un colore diverso da quello del testo). È disponibile anche una funzionalità di correzione automatica che permette di ottenere suggerimenti su possibili correzioni della parola errata ed anche di realizzare la correzione in modo automatico. Inoltre è disponibile un thesaurus che permette di ottenere sinonimi e contrari.

ricerca e sostituzione :

tramite gli opportuni comandi *Search* (Trova) e *Replace* (Sostituisci) del menù *Edit* (Modifica) è possibile cercare le occorrenze di una sequenza di caratteri all'interno del testo ed eventualmente sostituirla con un'altra sequenza di caratteri. I comandi Trova e Sostituisci possono essere usati anche per cambiare lo stile.

comandi di visualizzazione:

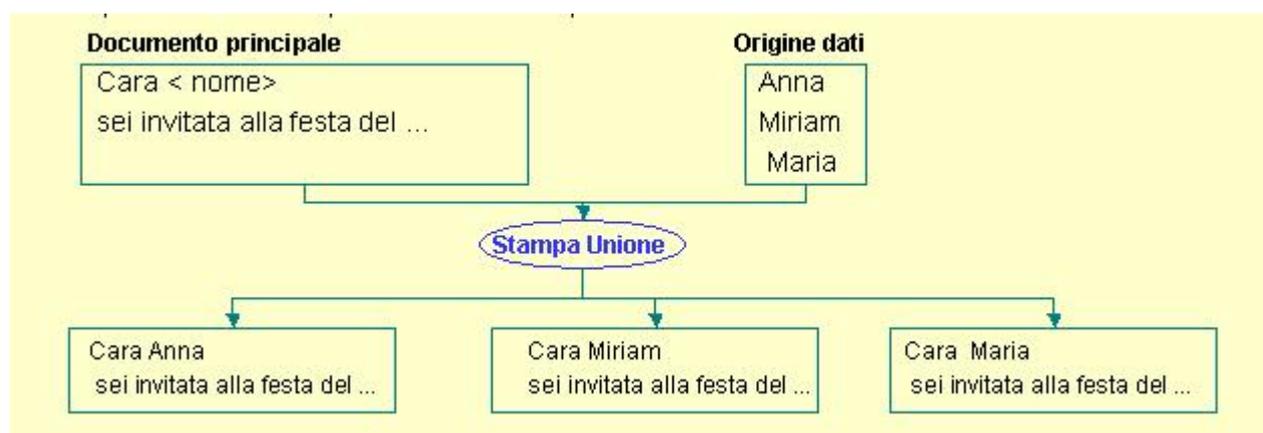
il menù *View* (visualizza) permette di visualizzare il documento secondo varie modalità che permettono di evidenziare la struttura del documento (sostanzialmente mostrando i titoli ai vari livelli) oppure di visualizzare il documento normale secondo vari formati di presentazione.

Uso di un editore di testi: stampa unione

comandi relativi al file:

come tutti i programmi che operano su dei file, anche gli editori di testi permettono (dal menu

file) di aprire, chiudere, salvare e stampare il file contenente il documento. Di particolare interesse è la stampa unione che permette di stampare dati provenienti da due file. Più precisamente, si ha un documento principale che contiene la struttura del testo da stampare (ad esempio un modello di lettera) ed un documento (detto origine dei dati) contenente i dati da includere nel documento principale (ad esempio i dati dei vari destinatari della lettera). Mediante opportuni campi unione, di solito indicati racchiudendo del testo fra <>, si indicano nel documento principale i tipi di dati che devono essere prelevati dall'origine dei dati ed i punti dove tali dati devono essere inseriti. Al momento della stampa unione verranno quindi generate (e stampate) varie copie del documento principale, ottenute istanziandolo con tutti i dati del file origine dei dati nel modo specificato dai campi unione. Ad esempio



Uso di un editore di testi: formattazione

Vi sono vari comandi che permettono di definire il formato di un documento. Oltre ai possibili formati sui caratteri, già visti, ricordiamo qui le seguenti funzioni, accessibili dal menu Formato:

margini:

si possono definire i margini, ovvero gli spazi bianche che circondano il testo usando gli opportuni comandi sul righello;

numeri di pagina:

si possono definire secondo vari formati e sono inseriti automaticamente dal programma;

elenchi:

si possono inserire elenchi numerati, puntati etc. quali quello contenuto in questa pagina;

tabelle:

(vedi più avanti);

stili:

è possibile memorizzare tutte le informazioni relative alla formattazione di un documento e creare così uno stile (al quale sarà dato un nome). Tutte le volte che si vorranno usare tali formattazioni per un nuovo documento sarà sufficiente usare lo stile memorizzato in precedenza (rintracciabile dal menu stile, nel menu formato, usando il nome dello stile scelto).

Uso di un editore di testi: immagini, tabelle

Di solito gli editori di testi permettono anche di inserire immagini, grafici e tabelle. Ad esempio, in *Word* è possibile:

1. inserire un'immagine selezionando Immagine nel menu Inserisci. Si ottiene così un menu a tendina dal quale si possono selezionare i vari tipi di immagine che si possono usare. Ad esempio, l'immagine può essere contenuta in una file e copiata (oppure collegata), oppure l'immagine può essere disegnata usando delle forme predefinite (accessibili selezionando Forme). Notare che quando si usano le forme per realizzare delle figure convesse (cerchi, ellissi, rettangoli ecc) se si vuole inserire del testo entro il perimetro della figura, il colore della

medesima deve essere impostato a nessun colore. Per fare questo basta cliccare due volte sulla figura stessa e quindi usare il menu che appare.

2. Inserire una Tabella come, ad esempio, la seguente

Nome	Cognome	Via	Città
Mario	Rossi	Vivaldi	Roma
Anna	Bianchi	Corelli	Firenze

Per inserire una tabella si usa il comando Inserisci Tabella del menu Tabella. Il formato della tabella può essere scelto usando il comando Formattazione Automatica fra vari formati predefiniti. È possibile cambiare le dimensioni usando il comando Altezza e larghezza Celle, e si possono fare operazioni sugli elementi delle varie tabelle (ad esempio, ordinamenti per i valori di un certo campo).

Oggetti OLE

Con i moderni editor di testo è possibile inserire nel documento non solo testo ma anche grafici, fogli di lavoro, immagini, video e animazioni, creando così documenti composti. Questo è possibile usando la tecnologia **OLE** (*Object Linking and Embedding*) o un altro sistema simile chiamato OpenDoc.

Un oggetto OLE è un generico oggetto (testo, immagine, informazione multimediale) che, tramite OLE viene sempre gestito dall'applicazione che lo ha creato: ovvero, quando viene selezionato un oggetto OLE in un documento viene invocata l'applicazione che ha creato l'oggetto OLE e che, in generale, può essere diversa dall'applicazione in uso per elaborare il documento. Ad esempio, se abbiamo creato un grafico con uno spreadsheet e lo abbiamo inserito come oggetto OLE in un file che stiamo editando con un editore di testi, facendo doppio clic sul grafico attiviamo lo *spreadsheet*. Ci sono due modalità per inserire e gestire un oggetto OLE

Incorporamento (*Embedding*).

I dati dell'oggetto da inserire vengono copiati nel documento composto che stiamo creando. Facendo doppio clic sull'oggetto incorporato si attivano i menu dell'applicazione che lo ha creato.

Collegamento (*Linking*).

L'oggetto da inserire non viene copiato e nel documento composto viene inserito solo un collegamento a tale oggetto. Facendo doppio clic sull'oggetto collegato si attiva l'applicazione che lo ha creato (sull'oggetto collegato).

Rappresentazione interna dei documenti

Ogni documento elettronico ha una o più rappresentazioni interne (in termini di byte, o caratteri) ed una o più rappresentazioni esterne (in termini di pixel). Internamente, i caratteri che costituiscono un documento, una volta codificati opportunamente, sono memorizzate in strutture logiche dette **file** che, come abbiamo visto, sono gestite dal sistema operativo.

Per rappresentare i caratteri si usano i seguenti codici:

Il codice **ASCII**,

utilizza 7 bit per i caratteri, permettendo così di rappresentare 128 valori diversi ($2^7 = 128$). L'ottavo bit viene utilizzato come codice di controllo. Esistono anche codici ASCII a 8 bit e vari produttori hanno usato estensioni diverse (questo spiega i problemi di conversione tra file di sistemi diversi, ad esempio fra *Windows* e *OS-Macintosh*).

Il codice **EBCDIC**

utilizza 8 bit e permette quindi di rappresentare $2^8 = 256$ valori diversi.

Il codice **Unicode**

utilizza 16 bit. I possibili valori non sono ancora stati tutti assegnati (34168 su 655236). Unicode raddoppia lo spazio occupato da testi scritti in caratteri latini, ma permette la rappresentazione dei caratteri di molti alfabeti non latini (arabo, ebraico, cinese, ecc) e di caratteri speciali (matematici, tecnici, elementi di disegno ecc).

Il codice ISO Latin 1

è un'estensione a 8 bit del codice ASCII ed è molto diffuso negli usi di rete. ISO Latin 1 corrisponde ai caratteri Unicode i cui primi 8 bit sono tutti zero.

Formati proprietari e non

Il codice **ASCII** si può usare per rappresentare qualsiasi struttura dati o programma, usando le particolari sequenze di caratteri tipiche di quella struttura. Tuttavia molte applicazioni usando un formato privato e proprietario, di solito binario, dei documenti che creano. Anche se si adducono spesso ragioni di efficienza per tali formati in realtà la vera ragione è commerciale, dato che il formato proprietario impedisce che tale formato sia usabile anche da altre applicazioni.

Esempio:

Il formato .doc, prodotto con il programma di scrittura *Microsoft Word*, è un formato proprietario. Se guardiamo un documento .doc come file di testo (ad esempio usando un editor puramente testuale quale il Notepad disponibile in *Windows* solitamente dal menu accessori nel menu ottenibile cliccando su Start), vediamo una sequenza di caratteri illeggibile. Riusciamo a leggere un file .doc usando il programma *Word*, perchè tale programma interpreta correttamente le varie sequenze di caratteri. Per ragioni di interoperabilità e portabilità dei documenti tuttavia si stanno affermando formati di interscambio.

Esempio:

Word ed altri programmi di scrittura sono capaci di produrre file in un formato chiamato RTF (Rich Text Format) che è leggibile anche come file di testo e che è comprensibile a varie applicazioni.

Il linguaggio HTML, nato nel contesto del *World Wide Web*, si sta affermando come un importante sistema di rappresentazione di documenti. A causa di alcuni suoi limiti è stato introdotto un nuovo linguaggio per documenti chiamato XML. HTML e XML definiscono formati "aperti", non proprietari e leggibili.

Alcuni formati

Esistono molti formati diversi di file di dati, quasi sempre distinguibili dall'estensione .xxx che segue il nome del file

Esempi di estensione:

.txt o .asc
file in formato ASCII

.doc
file in formato MS *Word*

.rtf
file in *Rich Text Format*

.htm o .html
file in formato HTML

.ps
file in formato *PostScript*

.eps
file in formato *Encapsulated PostScript*

.pdf
Portable Document Format (Acrobat)

.zip

	file compresso (usare <i>unzip</i> per decomprimere)
.gz	file compresso (gunzip per decomprimere)
.z o .Z	file compresso (uncompress per decomprimere)
.tar	archivio realizzato col programma TAR
.tar.z	archivio realizzato con TAR e poi compresso
.tgz	archivio realizzato con TAR e poi compresso
.uu	file ASCII codificato con uuencode
.gif	file contenente immagine in formato GIF
.jpg	file contenente immagine in formato JPEG
.pic	file contenente immagine in formato PIC
.wav	file contenente audio
.au	file contenente audio
.mpg	file contenente filmato MPEG
.mov	file contenente filmato <i>QuickTime</i>

Compressione dei file

Spesso prima di spedire o trasferire dei file, oppure prima di memorizzarli su disco, se ne effettua la **compressione**. Tale tecnica permette di mantenere il contenuto informativo dei dati riducendo però lo spazio di memoria occupato. È possibile la compressione perché i file normalmente contengono delle ridondanze che possono essere eliminate usando opportuni algoritmi. Per file prodotti da programmi di videoscrittura si possono ottenere riduzioni dello spazio occupato fino al 50%, mentre per file di grafica bitmap si può arrivare al 90%. Ovviamente esiste un limite alla possibilità di compressione, oltre il quale si perde il contenuto informativo.

Alcuni formati sono già compressi per cui sono inutili ulteriori compressioni. Per esempio è inutile comprimere i file pdf, gif o mp3.

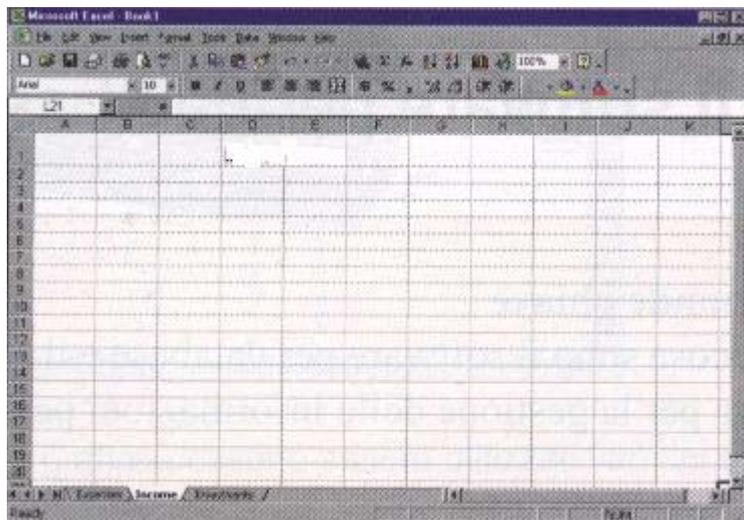
Per *Windows* uno dei programmi di compressione *shareware* più diffusi si chiama Winzip (è possibile reperire tale programma in rete). Per Unix il metodo più comune per produrre archivi compressi è usare i programmi TAR e GZIP.

Ovviamente, prima di usare un file compresso questo dovrà essere decompresso: a tale scopo si usano gli stessi programmi usati per la compressione con modalità diverse.

I fogli elettronici

I **fogli elettronici** (detti anche *spreadsheet*) sono strumenti che permettono il trattamento e l'analisi di dati numerici mediante tabelle (o griglie) nelle quali è possibile operare sulle singole caselle, sulle colonne, sulle righe o su più caselle collegate fra loro. Si possono effettuare varie operazioni numeriche, dalle più semplici (quali la somma dei valori contenuti in una colonna) a quelle più

complesse quali si hanno in ambito statistico e finanziario. I fogli elettronici permettono inoltre di definire grafici (secondo vari formati) in modo abbastanza immediato. I più diffusi fogli elettronici sono Microsoft Excel, Corel Quattro Pro e Lotus 1-2-3. Un foglio elettronico si presenta come nella figura seguente:



Immissione dei dati nei fogli elettronici

I dati presenti in un **foglio elettronico** sono molto più strutturati di quelli presenti in un elaboratore di testi: l'area di lavoro di un foglio elettronico infatti è organizzata in colonne e righe. Ogni colonna è identificata da un nome riportato nella parte superiore (A, B, C ecc) mentre le righe sono identificate da numeri presenti nel lato sinistro (1,2,3 ecc). È possibile inserire opportune etichette all'inizio delle righe e delle colonne, così come in altri punti del foglio, semplicemente digitandole da tastiera.

Il punto di intersezione fra una riga e una colonna è detto **cella**, e la sua posizione è ottenuta dalla colonna e dalla riga che la individuano (ad esempio A4, B7 ecc). Un intervallo è costituito da un gruppo di celle adiacenti, per esempio quelle celle comprese fra A3 e D6

È possibile **immettere dei valori** in una cella semplicemente posizionando il cursore sulla cella e quindi digitando il valore. È possibile anche immettere contemporaneamente valori in gruppi di celle mediante operazioni di taglia e incolla, oppure trasferendoli per mezzo di un programma da un altro file. I valori possono essere molte cose: valori monetari, percentuali, gradi di temperatura, misure di angoli, eccetera.

Cosa si può fare con un foglio elettronico

Il motivo principale dell'utilità di un foglio elettronico consiste nella possibilità di immettere nella caselle formule, funzioni e nella possibilità di ricalcolare con estrema facilità nuovi valori.

Ad esempio, supponendo di avere nella colonna A, nelle righe da 1 a 10, dei valori numerici possiamo inserire nella cella A11 una formula del tipo = SOMMA(A1:A10). Questa formula ci dice che il valore della casella 11 è costituito dalla somma dei valori presenti nella prime 10 righe della colonna A. Se cambiamo uno di questi valori il valore della somma in A11 viene aggiornato automaticamente.

Nelle formule possiamo usare delle funzioni predefinite quali ad esempio la somma, la media di un insieme di valori, l'ammortamento, eccetera. Sono possibili molte funzioni di vario tipo: finanziarie, matematiche e trigonometriche, statistiche, eccetera, tutte selezionabili dal menù Inserisci.

La caratteristica essenziale di un foglio elettronico è la possibilità di usare all'interno di una cella riferimenti ad altre celle (come ad esempio A1 e A10 nella cella A11, nell'esempio precedente). Questo permette di poter ricalcolare valori in modo immediato: basta immettere un nuovo valore in un operando per avere automaticamente aggiornato il risultato in tutte le formule che lo usano. Da questa facilità di ricalcolo deriva la possibilità di effettuare la cosiddetta analisi condizionale, ovvero si possono esaminare più scenari diversi semplicemente immettendo valori diversi e quindi ricalcolando istantaneamente i risultati che ci interessano.

È possibile anche inserire nei fogli elettronici vari tipi di grafici quali istogrammi, grafici a barre, a torta, eccetera.