Da File System a Content System

Nuovi paradigmi per la gestione dei documenti

Emanuele Goldoni
Università degli Studi di Pavia

Un'interfaccia dovrebbe usare
l'allocazione dinamica della memoria,
le liste puntate, lo hashing
o qualsiasi altra tecnica disponibile,
ma non dovrebbe mai e poi mai
scaricare sugli utenti
le limitazioni dell'implementazione software,
dando origine a vincoli arbitrari [...]

Jef Raskin, *Interfacce a misura d'uomo*

Introduzione

Gli ultimi trent’anni di storia dell’informatica, sin dalle prime versioni del Disk Operating System, hanno portato gli utenti a pensare che il sistema di gestione dei documenti informatici basato su albero delle directory e nomi di file sia l’unico possibile. Al contrario questo paradigma, vicino al modo di pensare dei programmatori, è tuttaltro che intuitivo e versatile, poco scalabile al crescere del numero di documenti da gestire e limitante in tutte le operazioni di ricerca e recupero dei dati stessi. In questo documento verranno quindi messi in luce, attraverso esempi, i limiti sopra citati dei file system oggi in uso e verranno anche proposto nuovi paradigmi, realmente *user centred*, in fase di sviluppo o già proposti negli anni passati.

Il file system gerarchico

Il paradigma del file system ad albero, utilizzato da tutti i sistemi operativi più diffusi, deve la sua grandissima diffusione ai vantaggi di natura pratica offerti ai progettisti ed agli sviluppatori. L'implementazione di un semplice file system basato su directory, sottodirectory e file è infatti estremamente facile da implementare: il sistema utilizzato, ad esempio, dai lettori mp3 è implementabile in poche ore e si appoggia su concetti basilari di programmazione quali liste e puntatori. Inoltre una struttura ad albero permette di individuare univocamente un file attraverso una sola stringa, contenente il percorso completo (cioè la sequenza, in ordine gerarchico, delle cartelle e sottocartelle) ed il nome del documento, semplificando quindi ulteriormente il lavoro di chi deve scrivere programmi.

Il dilemma del nome

Quando stiamo cercando di salvare il nostro lavoro, i nomi di file sono una seccatura, perché dobbiamo interrompere quello che stiamo facendo (archiviare il nostro lavoro) e inventarci lì per lì un nome per il file. Come sottolinea lo stesso Jef Raskin, «creare un nome è un compito oneroso: ci viene richiesto di inventare un nome che sia unico, facile da ricordare, che rispetti le convenzioni del sistema che stiamo usando, e di farlo così, sui due piedi, nel giro di qualche secondo. Inoltre, in quel momento, il nostro fuoco dell'attenzione non è il nome del file, ma salvare il nostro lavoro. Quando poi si tratta di ritrovare qualcosa, i nomi dei file sono un supplizio: dopo qualche settimana (o meno) scopriamo che, forse, il nome che avevamo inventato non era poi così significativo, e che l'abbiamo scordato [...]».
Con che nome dovrò salvare questo documento? *Interazione Uomo Macchina - Da File System a Content System* sembra essere una buona scelta e sicuramente domani sarò in grado trovarlo al primo colpo. Tuttavia, tra alcuni mesi, potrei non ricordare più il titolo esatto ma solo il fatto di aver scritto una relazione per il progetto di Interazione Uomo Macchina: meglio quindi *Relazione corso Interazione Uomo Macchina - Da File System a Content System*? Ma tra due anni potrei solo ricordare di aver fatto una relazione per un corso di HCI: cercando tuttavia i file che contengono *relazione* all'interno del nome, potrei ottenere come risultato decine di file e, utilizzando invece HCI, nessun risultato.

Il (dis)ordine delle directory

Gli ultimi tempi sono stati caratterizzati da un aumento esponenziale del numero medio di file archiviati all'interno dei computer. Tra le cause di questo fenomeno vi è sicuramente l'impiego dell'informatica in un grandissimo numero di attività quotidiane ma, soprattutto, l'introduzione di strumenti che consentono, con estrema facilità, di produrre grandi quantità di documenti multimediali (macchine digitali, videocamere, cellulari) anche senza disporre di particolari competenze.
Questo fenomeno ha ulteriormente evidenziato tutti i limiti dell'attuale metodo di archiviazione dei documenti informatici, qualunque sia la loro natura. Infatti, se solo inventare un nome significativo per un documento di testo di qualche pagina è impegnativo, diviene invece improponibile il pensare di rinominare le centinaia di fotografie digitali che, ogni anno, ciascuno di noi è in grado di realizzare. Da qui la pratica diffusa di suddividere i documenti in cartelle, ciascuna con un nome significativo (o presunto tale) e in grado di semplificare le operazioni di ricerca dei documenti.
Questo sistema di archiviazione dei documenti in directory e sottodirectory, spesso utilizzato per "mettere ordine" al crescere del numero di documenti, è però fortemente limitante. Prendiamo, ad esempio, il seguente albero delle directory:

Se ho festeggiato l'ultimo compleanno di mio fratello in vacanza, dove troverò le foto? Saranno in *Compleanni* → *Parenti* o, piuttosto, in *Vacanze* → *2007*? E in che anno sono stato in vacanza a Madonna di Campiglio? Era il 2005 o il 2006?
Lo stesso dicasi, ad esempio, per un ufficio: meglio raggruppare le fatture per anni o per clienti/fornitori? E se invece volessi dividerle per importi in base a determinati scaglioni fiscali? Inoltre, una volta adottata una determinata soluzione, è molto difficile cambiare! Se, anziché divise per anno, volessi le foto delle vacanze catalogate per località, dovrei scorrere tutti i file singolarmente.
Un documento infatti può essere associato a più concetti ma la necessità di inserimento in un'unica directory, l'univocità della posizione dell'albero comporta la rottura di alcuni di questi collegamenti. Da questo discendono i problemi sopra citati nel recupero dei documenti archiviati se non si adotta, in fase di ricerca, lo stesso "punto di vista" utilizzato per la memorizzazione.

L’estensione senza ragione

Uno dei problemi meno compresi dagli utenti è la dissociazione tra contenuto e *filetype*: poco importa infatti all'utente se la richiesta di preventivo è stata scritta con un word processor e stampata per essere spedita via fax o posta ordinaria piuttosto che inviata direttamente per email.
A chi non è capitato almeno una volta di cercare per ore un documento scaricato da internet tra i file pdf, salvo poi scoprire che si trattava di un file di Word? Questo è perfettamente comprensibile: durante la lettura, l’attenzione è focalizzata sul testo e non sul programma che si sta utilizzando per leggere. Dal punto di vista dell’utente, le estensioni non hanno quindi senso: esistono i documenti, le immagini, i filmati e non i doc, i txt, gli rtf, i pdf, le png, le bmp, le jpg etc.. Per il computer la differenza è invece sostanziale.

Gli approcci alternativi

Abbiamo quindi visto come l’attuale sistema di gestione dei documenti informatici, basato sull’utilizzo di una gerarchia di directory e un nome univoco dotato di estensione, sia fortemente limitante e non esente da problemi. Di seguito verranno illustrati nuovi paradigmi,

Il Canon Cat

Il primo approccio, proposto dall’esperto di interazione Jef Raskin, è stato implementato nel Canon Cat, un personal computer per la videoscrittura progettato dallo stesso Raskin 20 anni fa. Il Cat era basato su un’interfaccia interamente testuale e non faceva assolutamente uso di mouse, icone o elementi grafici e, invece di utilizzare la tradizionale interfaccia a linea di comando o menu di sistema, si appoggiava ad una apposita tastiera con comandi attivabili tenendo premuto il tasto "Use Front". Alcuni concetti introdotti da questo sistema nel 1989 solo in tempi recenti hanno iniziato ad essere presenti nei sistemi operativi Mac e Windows e altri ancora oggi non hanno corrispondenze nel panorama tecnologico. Tra questi vi è il tasto DISK, che permetteva di gestire il salvataggio su disco, il caricamento da disco, la copia dei dati su un supporto diverso con un unico tasto ed in maniera del tutto trasparente per l’utente, e l’eliminazione del concetto di file. Con il Cat infatti, tutti i dati erano visti come un unico, lungo *stream* di testo suddiviso in pagine.
L’idea di fondo del Canon Cat è quella di utilizzare un unico, grande documento di testo, strutturabile a piacere grazie ad appositi delimitatori, in grado di consentire una ricerca *full text* e mostrare direttamente i termini cercati all’interno del contesto. La descrizione completa di questo sistema è allegata in appendice.
Sebbene lo stesso Raskin non consideri questo meccanismo utilizzabile tout court per la catalogazione delle immagini e, allo stesso modo, dovrebbe essere modificato per gestire la grande quantità di documenti che gli utenti non producono ma si limitano a leggere e memorizzare (pagine web, pdf etc), l’idea di fondo rimane comunque valida e rimane la dimostrazione di come innovativi paradigmi di interazione possano esistere e funzionare (e, sperabilmente, meglio di quelli esistenti!).
A questo proposito, è opportuno notare come i motori di ricerca più diffusi si siano evoluti, mostrando per ogni pagina il frammento di testo all’interno del quale compare il termine cercato anziché, come in un primo tempo, la descrizione della pagina indicata nei metatag.

I motori di ricerca

E’ proprio il più grande archivio di file del mondo (leggasi Internet) a fornire un secondo, lampante, esempio di come il tradizionale meccanismo dei file system sia inadeguato e poco versatile. Chi naviga in Internet digitando ogni volta l’indirizzo completo? In fondo, a chi interessa sapere quale è il nome file o la posizione all’interno del sito? E se, a maggior ragione, non sappiamo se l’informazione da noi cercata esiste, come possiamo sapere dove si trova, in che file e che formato è stato utilizzato per memorizzarla? Come ha sottolineato lo stesso Raskin, il miglior nome di un file è il suo stesso contenuto. In questo i motori di ricerca, Google in primis, sono veramente all’avanguardia: consentono di cercare, attraverso un’interfaccia sostanzialmente unica, tra testi, immagini e video una particolare informazione; in altre parole, il concetto di file o estensione scompare! Personalmente ho potuto constatare come molti utenti neofiti, avendo il browser impostato per aprire direttamente il motore di ricerca Google, siano in breve tempo giunti a ritenere che Google stesso sia Internet. In quest’ottica, il concetto di URL stesso è superfluo: gli utenti sopra citati utilizzano infatti il box di ricerca di Google anche in questo caso, digitando per intero l’indirizzo. Non è un caso che il browser incluso nel progetto OLPC (di cui parleremo in seguito), destinato a persone senza alcuna cultura informatica, utilizzi Google come pagina iniziale e abbia invece del tutto eliminato la barra degli indirizzi!

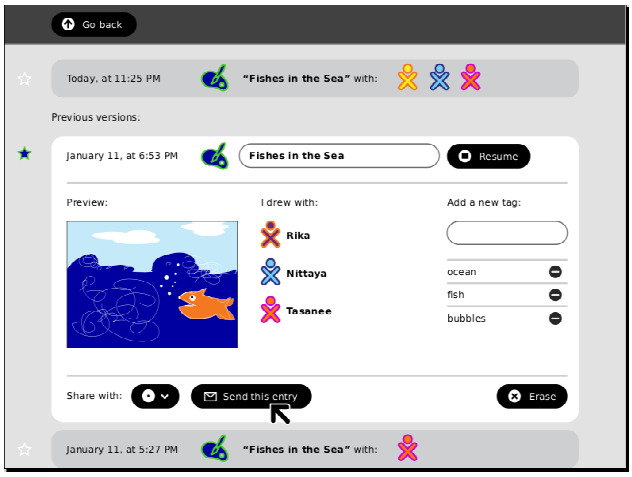
Il progetto OLPC

Nel Gennaio 2005 lo MIT Media Lab ha lanciato un nuovo progetto di ricerca per lo sviluppo di un computer portatile da 100 $. Una associazione non-profit, One Laptop Per Child, ha quindi preso in carico lo sviluppo del progetto, vedendo in esso una tecnologia ed un prodotto in grado di rivoluzionare il modo di educare i ragazzi, specialmente coloro che vivono in zone rurali o paesi meno sviluppati. Il software a bordo di questi computer utilizza un’interfaccia grafica sviluppata da zero e completamente innovativa, così come una tastiera leggermente modificata, un trackpad ed un monitor in grado di operare in più modalità. In particolare, per quel che riguarda l’interfaccia utente Sugar, i progettisti hanno voluto abbandonare la tradizionale metafora del desktop, preferendo concetti che meglio si adattano all’idea di ambiente collaborativo per l’apprendimento. Esempi di questo *paradigm shift* sono il passaggio da Desktop a Neighborhood, da Barra dei Menu a Frame, da Applicativo software a Attività, da File a Oggetto, da File system gerarchico a Diario (Journal).
Nel capitolo introduttivo delle specifiche del OLPC troviamo il seguente paragrafo, che merita di essere riportato.

> *The concept of the Journal, a written documentation of everyday events, is generally understood, albeit in various forms across cultures. A journal typically chronicles the activities one has done throughout the day. We have chosen to adopt a journal metaphor for the file system as our basic approach to file organization. While the underlying implementation of such a file system does not differ significantly from some of those in contemporary operating systems, it also holds less importance than the journal abstraction itself.*

In poche righe sono riassunti concetti fondamentali e che, raramente, assumono un ruolo così centrale nella progettazione: l'indipendenza dell'interfaccia utente dall'implementazione sottostante e l'attenzione agli aspetti cognitivi e l'influenza della cultura sul modo di leggere la realtà.

Figura 1.1 Proposta di interfaccia per il Journal dell'OLPC



Tra i vantaggi di questo innovativo approccio vi è, ad esempio, l'abbandono del concetto di apertura/chiusura di un file: come in un diario, si può in ogni momento interrompere il lavoro su una pagina, tornare ad una precedente e riprendere quanto di stava facendo tempo prima o visualizzare l'evoluzione nel tempo di una particolare attività. La possibilità di avere le attività organizzate temporalmente, di sfruttare i metadati per la catalogazione, la ricerca e l'ordinamento, nonché un sistema di backup più intuitivo, sono alcuni tra i vantaggi di questo nuova interfaccia.

I tag ed il Web 2.0

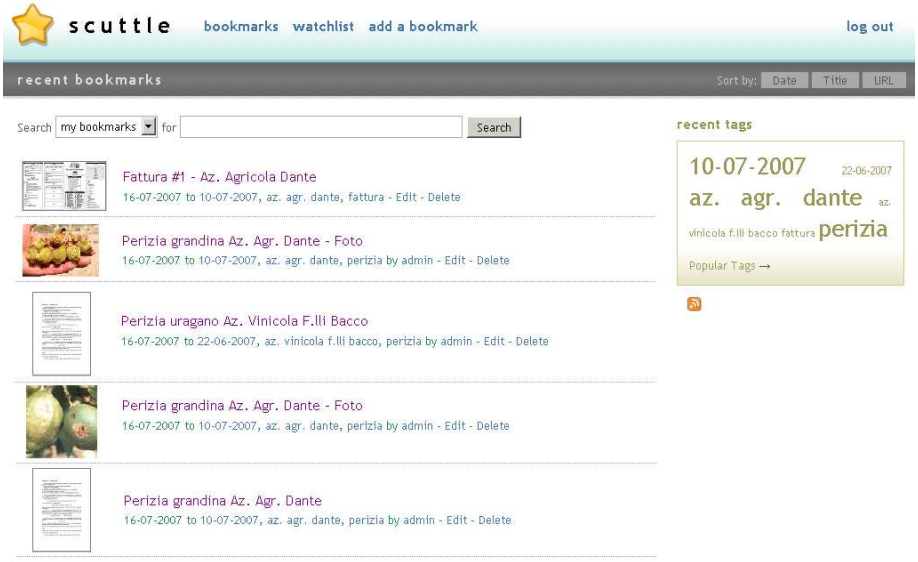
Siti come Flickr, YouTube o del.icio.us, comunemente ascritti al fenomeno Web2.0, hanno spopolato e basato la loro fortuna su un concetto tuttaltro che innovativo: i *tag*. Un *tag* altro non è infatti che una keyword, o parole chiave. Come detto prima, il miglior nome per un file è il suo stesso contenuto o, in alternativa, un insieme di parole in grado di sintetizzarlo.
Quando un ricercatore apre il sito della IEEE o ACM ed effettua un'interrogazione, altro non fa che elencare alcune parole chiave e il sistema effettuerà una comparazione con tutti i termini presenti nel titolo degli articoli e nell'insieme di keyword indicate dagli autori stessi. I database scientifici, i sistemi di *social bookmarking* e di *video o foto sharing* hanno quindi permesso all'utente di superare il concetto di file sostituendolo con il contenuto stesso. A chi interessa con quale nome YouTube archivia i video sui propri server? A nulla, se il video è identificabile con il contenuto stesso!

Il tagged file system

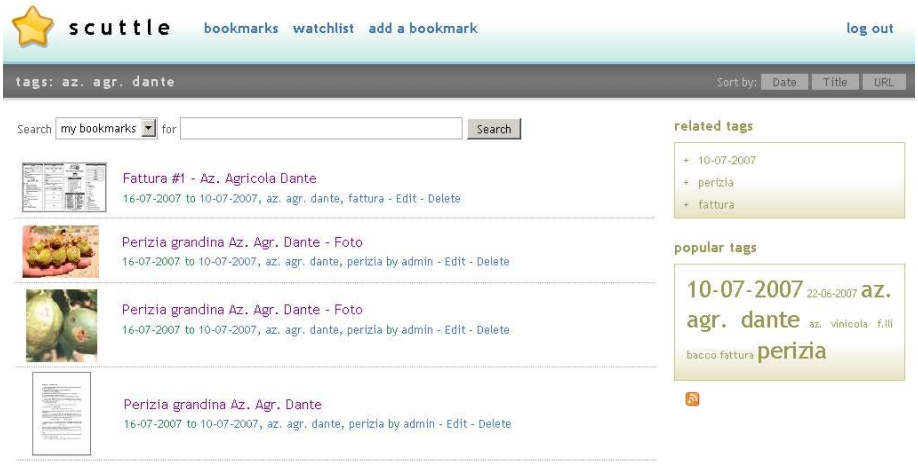
L’idea di associare ad ogni documento uno o più *tag* (o keyword o label) significativi, in grado di descriverne il contenuto, si è rapidamente estesa sul web e, dai contenuti multimediali (Flickr.com è stato il pioniere) e bookmark, è oggi passata anche alla posta elettronica. Ora manca solamente il “salto” più importante: il passaggio al desktop. Quella di seguito mostrata è solo una simulazione grafica ma, in futuro, diventerà probabilmente realtà. Con un sistema di questo tipo, non esisteranno più directory e nomi di file e l’utente potrà navigare tra i documenti in modo estremamente versatile e rapido.

La ricerca tramite tag: caso d’uso

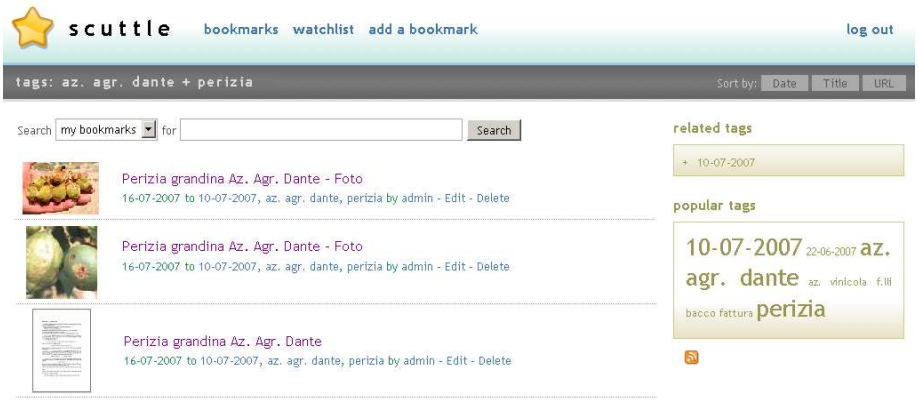
Supponiamo ad esempio di sbirciare, solo per un istante, all’interno del personal computer di Giorgio, un noto perito agrario. Questi viene spesso chiamato dalle aziende agricole per valutare i danni causati da grandinate, temporali o periodi di siccità; ogni volta deve scattare foto, osservare, stendere una relazione da inviare all’assicurazione e preparare quindi la fattura per il cliente.



Ecco l’elenco di tutti i file più recenti. Sulla destra una *tag-cloud* mostra i contenuti più frequenti. Poiché stiamo cercando le informazioni inerenti la perizia effettuata per conto dell’Az. Agr. Dante, clicchiamo sul *tag* corrispondente.



Ottimo, manca poco... Sulla destra è possibile vedere tutti i *tag* correlati: poiché a noi interessa la perizia e non la fattura, optiamo per il primo termine



Ed ecco finalmente tutti i documenti relativi a questo lavoro! Ora però meglio andarsene, prima che Giorgio se ne accorga...

I vantaggi del tagged file system

Un *tagged file system* permette quindi di superare il concetto di nome del file e quello di albero delle directory. Il nome del file è infatti sostituito da un titolo, un elenco di parole chiave e un'eventuale "istantanea" del contenuto, così da permetterne una più rapida individuazione. Inoltre, da una struttura gerarchica passiamo quindi ad una multi-dimensionale. Se quindi nel file system tradizionale è utilizzato solo un percorso univoco per raggiungere il file, con questo nuovo sistema il documento può essere ritrovato partendo da qualsiasi punto della rete semantica. Se, nell'esempio precedentemente illustrato, non mi fossi ricordato dell'azienda agricola in questione ma solo della data di intervento, avrei potuto partire da quest'ultima (10-07-2007) e quindi passare a perizia, riuscendo comunque ad rintracciare i file cercati.
Con questo nuovo approccio proposto, il computer passa quindi da *machine centric* a *user centric*: in altre parole, l'utente non è più costretto a pensare come nel modo in cui ragiona il PC, ma viceversa!

I rapporti con il Semantic Desktop

Certo il sistema non è esente da errori: gli errori di battitura nella digitazione dei *tag* (perizzia), l'utilizzo alternativo di plurali e singolari (alcune volte potrei utilizzare vacanza e, altre, vacanze), la scelta di sinonimi (vacanza, ferie o villeggiatura) o situazioni analoghe rendono più complicata la ricerca. Ed è anche in questa direzione che alcuni progetti di *semantic desktop* si stanno muovendo: mostrando infatti, in risposta ad una ricerca, tag semanticamente correlati a quello cercato (plurali, termini vicini per analogia, sinonimi, termini più generali etc.), i tempi e i "vicoli ciechi" incontrati verrebbero ridotti notevolmente, con conseguente maggior soddisfazione dell'utente.

Bibliografia

Bloehdorn S., Görlitz, Simon Schenk, Völkel M., **TagFS - Tag Semantics for Hierarchical File Systems,** Proceedings of the 6th International Conference on Knowledge Management, 2006 (I-KNOW 06)

Jef Raskin, **Interfacce a misura d'uomo**, Apogeo, 2003

Kase O., Lemire D., **Tag-Cloud Drawing: Algorithms for Cloud Visualization**, to appear in proceedings of Tagging and Metadata for Social Information Organization, 2007 (WWW 2007)

*OLPC project***, OLPC Human Interface Guidelines,** http://wiki.laptop.org/go/OLPC_Human_Interface_Guidelines, 2006