DATI STRUTTURATI, NON STRUTTURATI e
SEMI-STRUTTURATI

David Loshin

Dati semplicemente semi-strutturati

Pubblicato il 17 ottobre 2005

Abbiamo per lo più familiarità con i **dati strutturati**—dati che sono stati accuratamente modellati, organizzati, composti e formattati in modo da essere facilmente elaborati e amministrati. Gli esempi più frequenti comprendono i *database* e configurazioni più comuni come i fogli di calcolo, i file a formato fisso, i file di log, ecc. Fortunatamente è molto facile lavorare sui dati strutturati. In questo caso è possibile scrivere programmi che elaborano con facilità i dati, organizzandoli, analizzandoli e visualizzandoli.

Probabilmente abbiamo familiarità anche con i **dati non strutturati**. Infatti, ne state leggendo proprio in questo momento. I dati non stutturati comprendono tutta quella massa di informazione che non si può sistemare facilmente nelle tavole di un *database*. La forma più riconoscibile di dati non strutturati è costituita dal testo in documenti quali gli articoli, le presentazioni di diapositive o i *message components* nelle e-mail.

Ci sono poi contenuti di tipo intermedio che sono chiamati **dati semi-strutturati**. Questo termine si riferisce a insiemi di dati in cui c'è una certa struttura implicita che viene sempre mantenuta, ma non è di natura abbastanza regolare per avere i requisiti necessari al tipo di gestione e di automazione che si applica di solito ai dati strutturati. Siamo quotidianamente bombardati da dati semi-strutturati, in ambienti sia tecnici che non tecnici. Per esempio, le pagine web si conformano a certe disposizioni tipiche e i contenuti inseriti in file HTML spesso hanno tra i marcatori un certo numero di metadati. Questo comporta automaticamente per i dati presentati determinate particolarità. Un esempio di natura non tecnica può essere costitituito dai cartelli stradali collocati lungo un'autostrada. Benché ci siano sistemi locali diversi usati in paesi diversi, dopo avere osservato alcuni cartelli, siamo normalmente in grado di capire qual è la nostra uscita.

Questo è ciò che rende interessanti i dati semi-strutturati—benché non ci sia nessuna regola rigida di formattazione, la presenza di una certa regolarità basta per poter estrarre *qualche* informazione interessante.

---

http://www.b-eye-network.com/view/1761

David Loshin

Simple Semi-structured Data

Published: October 17, 2005

We are mostly familiar with structured data—the data that has been neatly modeled, organized, formed, and formatted into ways that are easy for us to manipulate and manage. The most frequent examples include databases, as well as more mundane frameworks such as spreadsheets, fixed-format files, log

files, etc. Fortunately, structured data is relatively easy to work with. Here, we can write programs that easily work with the data by organizing, analyzing and displaying it.

You are probably familiar with unstructured data as well. In fact, you are reading it right now. Unstructured data incorporates the mass of information that does not fit easily into a set of database tables. The most recognizable form of unstructured data is text in documents, such as articles, slide presentations or the message components of emails.

There is an intermediate classification of content called "semi-structured data." This refers to sets of data in which there is some implicit structure that is generally followed, but not enough of a regular structure to "qualify" for the kinds of management and automation usually applied to structured data. We are bombarded by semi-structured data on a daily basis, both in technical and non-technical environments. For example, web pages follow certain typical forms, and content embedded within HTML often have some degree of metadata within the tags. This automatically implies certain details about the data being presented. A non-technical example would be traffic signs posted along highways. While different areas use their own local protocols, you will probably figure out which exit is yours after reviewing a few highway signs.
This is what makes semi-structured data interesting—while there is no strict formatting rule, there is enough regularity that *some* interesting information can be extracted.