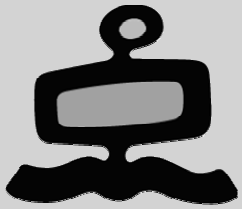


InformaticaUmanistica

Lezione 2
Gestione del testo

Pasquale Savino

ISTI - CNR

UNIVERSITÀ DI PISA

Sommario

- **La gestione dei dati in una Biblioteca Digitale**
  - Acquisizione
  - Rappresentazione
  - Indicizzazione
  - Ricerca
  - Conservazione
- **Gestione del testo**
- **Gestione delle immagini**

Gestione dei dati in una Biblioteca Digitale

Acquisizione

- Per essere gestiti all'interno della Biblioteca Digitale i documenti devono essere degli oggetti digitali
- Se abbiamo un testo su carta o una foto o un filmato su nastro, ecc. dobbiamo effettuare una operazione di acquisizione e trasformazione del documento in un formato digitale
- Tale operazione richiede l'uso di apparecchiature hardware e software specifiche per ogni tipo di dato
- È possibile creare il documento direttamente in formato digitale. Anche in questo caso abbiamo bisogno di strumenti hardware e software specifici.

Rappresentazione

- Il processo di acquisizione permette di ottenere documenti in formato digitale.
- I documenti digitali vengono rappresentati tramite opportuni formati, diversi per ogni tipo di dato.
- È possibile che il formato prodotto dal sistema di acquisizione non sia il più adatto, per cui può essere necessario trasformare il documento in una rappresentazione diversa.
- Importanza degli standard.

Indicizzazione

- Il processo di indicizzazione permette di descrivere il contenuto del documento e ne permette la ricerca.
- L'indicizzazione è un processo che ci è familiare: si pensi agli indici dei libri, ai cataloghi di opere (quadri, musica, film), ecc.
- Si possono distinguere due tipi di indicizzazione
  - Manuale: richiede molto tempo ed il risultato può dipendere da chi effettua l'indicizzazione. Permette di descrivere in modo dettagliato la semantica dell'oggetto digitale. Molto spesso viene facilitata tramite l'uso di thesauri ed ontologie.
  - Automatica: veloce e poco costosa. Può risultare imprecisa ed in alcuni casi è impossibile.

Ricerca

- La fase di ricerca consiste in tre parti principali:
  - Formulazione delle interrogazioni
  - Elaborazione delle interrogazioni in modo da recuperare quelle che meglio soddisfano i bisogni informativi dell'utente
  - Visualizzazione dei risultati
- La interrogazioni possono essere di tipo Booleano o a ricerca libera
- L'esecuzione delle interrogazioni può permettere di recuperare gli oggetti che corrispondono esattamente alla richiesta o che hanno un certo grado di similitudine con la richiesta.
- I risultati possono essere visualizzati in ordine di similitudine decrescente con l'interrogazione (se questo è disponibile)

Conservazione

- La conservazione consiste nell'archiviare su supporti di memorizzazione a lungo termine gli oggetti digitali.
- Al fine di permettere una effettiva conservazione dei dati è necessario predisporre un'organizzazione per la salvaguardia dei dati.
- Inoltre, è necessario fornire una descrizione completa delle caratteristiche dell'oggetto digitale, quali ad es. il formato, la data di digitalizzazione, il software utilizzato per la digitalizzazione e quello necessario per la visualizzazione.

Gestione del testo

Gestione di libri

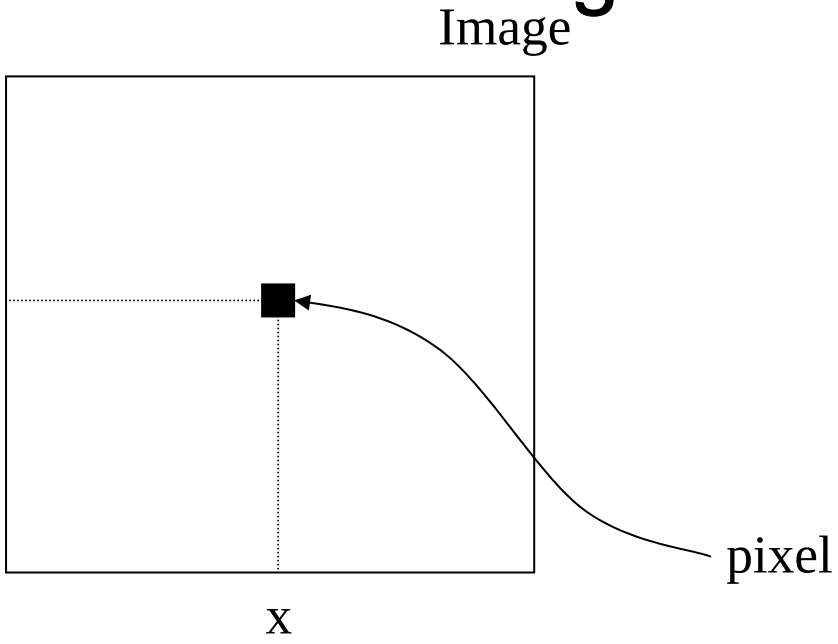
- **Automazione accesso ai cataloghi delle Biblioteche tradizionali**
- **Utilizzo di Cataloghi elettronici**
- **I servizi della Biblioteca rimangono gli stessi delle Biblioteche tradizionali, la ricerca dei libri risulta più veloce, ed è possibile effettuare ricerche complesse (ad es. libri scritti congiuntamente da due autori, oppure i libri su un certo argomento scritti in un dato periodo, ecc.)**
- **Si utilizzano tecnologie tradizionali per la gestione dei cataloghi**
- **È importante uniformare i cataloghi di varie biblioteche per permettere ricerche di libri su più cataloghi**

Gestione di documenti testuali

- **Il primo passo per una transizione dalle Biblioteche tradizionali alle Biblioteche Digitali prevede che la biblioteca abbia i documenti in forma elettronica, non solo i cataloghi**
- **La forma più semplice di contenuto (ma anche quella di più facile utilizzo) è il testo**
- **Documenti testuali ottenuti in modi diversi**
  - Creati direttamente per accesso on-line
  - Convertiti da stampe
  - Digitalizzati dalle tracce audio di film o programmi televisivi

Acquisizione di documenti come immagine

- **Le singole pagine sono acquisite come immagini tramite uno scanner**
- **Ogni singola pagina viene rappresentata come una sequenza di punti (pixels)**
- **Ad ogni pixel viene assegnato un valore (nero, bianco, grigio, colore), rappresentato con un codice binario**
- **Si possono applicare tecniche di compressione della codifica per ridurre l'occupazione dell'immagine**



Un'immagine è una funzione di due variabili spaziali f(x,y)

Per un'immagine a colori f(x,y) è un vettore con tre valori, uno per ognuno dei tre colori principali (rosso, blu, verde) corrispondenti all'intensità del pixel (x,y)

$f(x,y)=(f_{rosso}(x,y), f_{blu}(x,y), f_{verde}(x,y))$

Metodi di scanning


18 ASTRONOMY FOR ALL

of ancient times, Hipparchus, compiled the first really important star catalogue about this period, which is still referred to when questions arise concerning alterations in

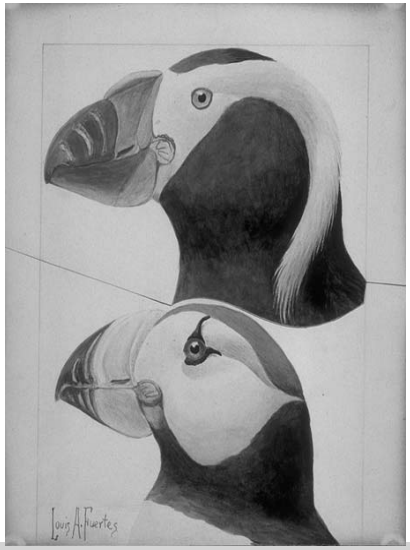
Fig. 13.—An Ancient Astronomer engaged in Observation.

the brightness of the stars (Fig. 17). One hundred and fifty years after the beginning of the Christian era Ptolemy wrote his comprehensive work on star-lore as then known, including very accurate calculations of solar and lunar eclipses, and set up his definition of the universal system. (*See* Chap. X., p. 111.)

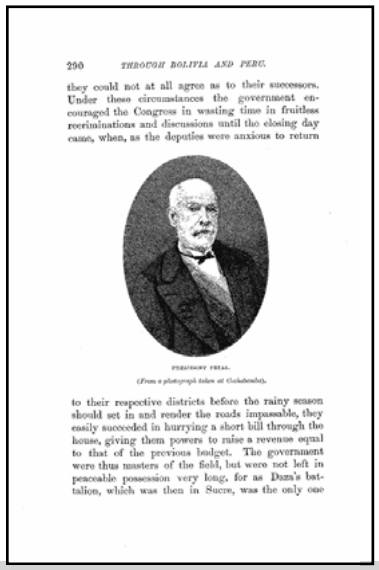
During the decline of arts and letters in Greece the grand old schools of savants were blotted out in the political and

Bitonal

Grayscale



Color


290 THROUGH BOLIVIA AND PERU.

they could not at all agree as to their successors. Under these circumstances the government encouraged the Congress in wasting time in fruitless recriminations and discussions until the closing day came, when, as the deputies were anxious to return

to their respective districts before the rainy season should set in and render the roads impassable, they easily succeeded in hurrying a short bill through the house, giving them powers to raise a revenue equal to that of the previous budget. The government were thus masters of the field, but were not left in peaceable possession very long, for as Daza's battalion, which was then in Sucre, was the only one

Special Treatment

Qualità dell'immagine acquisita

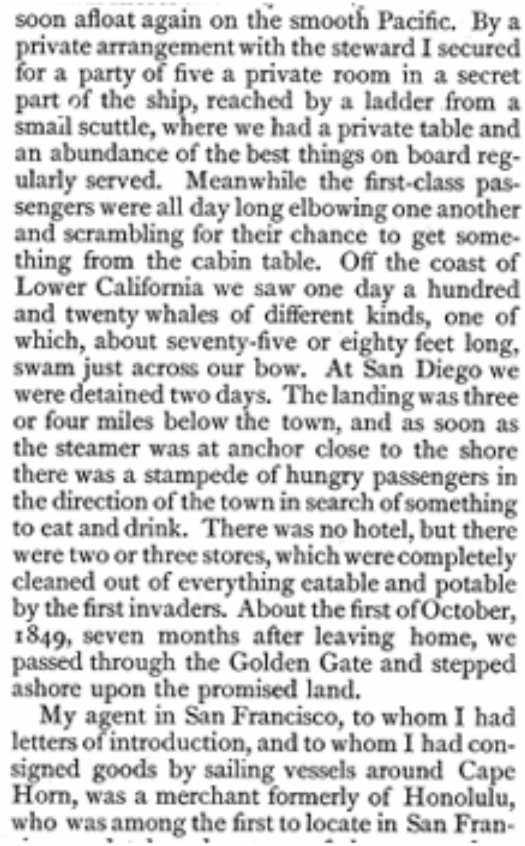
- **Dipende da**
  - Risoluzione dello scanner (numero di pixel per cm o numero di pixel per pollice – ppi)
  - bit depth (numero di bit per pixel). Maggiore è il numero di bit per pixel, maggiore sarà il numero di colori rappresentabili.
  - Miglioramento dell'immagine. Ad es. con eliminazione dei difetti dell'immagine, miglioramento dei colori, ecc.
  - Gestione del colore.
  - Tecnica di compressione dell'immagine.
  - Capacità di calcolo del sistema.
  - Abilità e perizia dell'operatore.

Riconoscimento caratteri

- **Processo di trasformazione di una immagine contenente testo in caratteri**
- **La qualità del riconoscimento dipende dalla qualità dell'immagine**
- **Utilizza tecniche di "image processing" combinate con tecniche linguistiche (ad es. utilizzo di dizionari)**
- **Il risultato è affetto da errori**
- **In una Biblioteca Digitale è opportuno, in generale, conservare sia l'immagine originale che il testo riconosciuto.**
- **Il testo può essere utilizzato per la ricerca per contenuto del documento**

Acquisizione

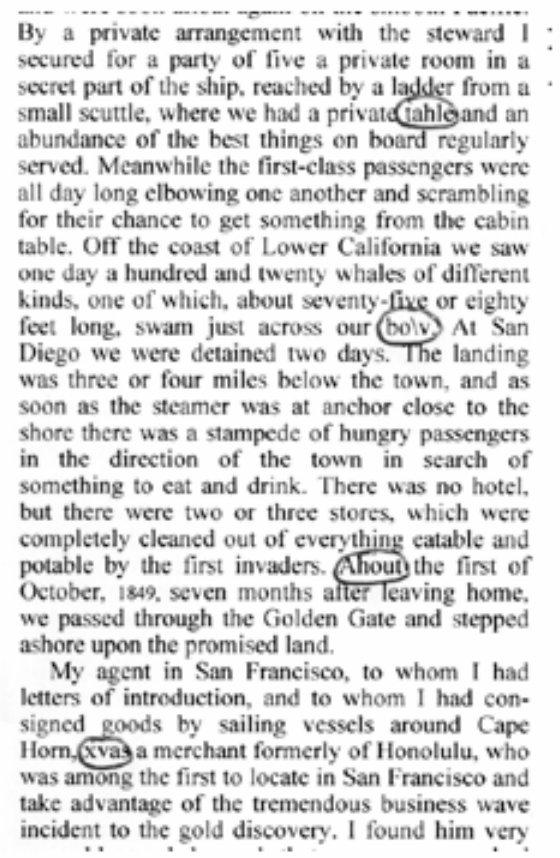
OCR Comparison



soon afloat again on the smooth Pacific. By a private arrangement with the steward I secured for a party of five a private room in a secret part of the ship, reached by a ladder from a small scuttle, where we had a private table and an abundance of the best things on board regularly served. Meanwhile the first-class passengers were all day long elbowing one another and scrambling for their chance to get something from the cabin table. Off the coast of Lower California we saw one day a hundred and twenty whales of different kinds, one of which, about seventy-five or eighty feet long, swam just across our bow. At San Diego we were detained two days. The landing was three or four miles below the town, and as soon as the steamer was at anchor close to the shore there was a stampede of hungry passengers in the direction of the town in search of something to eat and drink. There was no hotel, but there were two or three stores, which were completely cleaned out of everything eatable and potable by the first invaders. About the first of October, 1849, seven months after leaving home, we passed through the Golden Gate and stepped ashore upon the promised land.

My agent in San Francisco, to whom I had letters of introduction, and to whom I had consigned goods by sailing vessels around Cape Horn, was a merchant formerly of Honolulu, who was among the first to locate in San Fran-

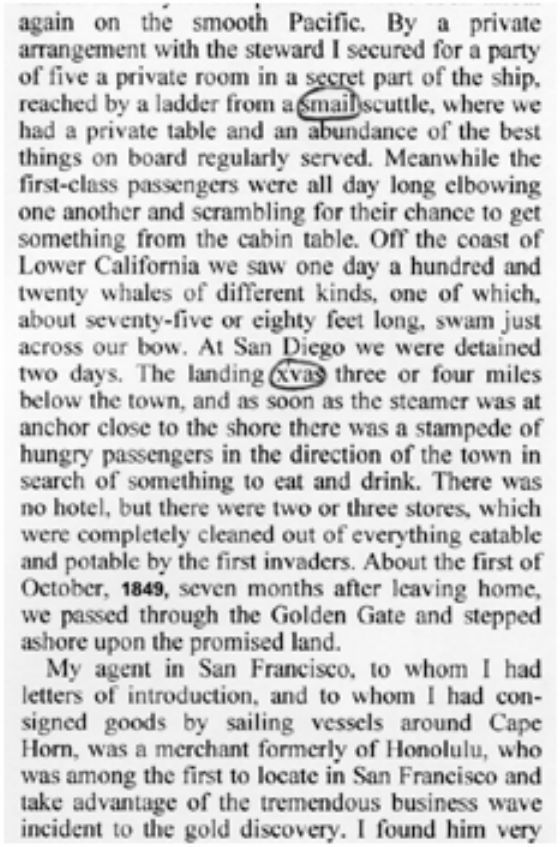
Image of Original Text



By a private arrangement with the steward I secured for a party of five a private room in a secret part of the ship, reached by a ladder from a small scuttle, where we had a private tahle and an abundance of the best things on board regularly served. Meanwhile the first-class passengers were all day long elbowing one another and scrambling for their chance to get something from the cabin table. Off the coast of Lower California we saw one day a hundred and twenty whales of different kinds, one of which, about seventy-five or eighty feet long, swam just across our bo\v. At San Diego we were detained two days. The landing was three or four miles below the town, and as soon as the steamer was at anchor close to the shore there was a stampede of hungry passengers in the direction of the town in search of something to eat and drink. There was no hotel, but there were two or three stores, which were completely cleaned out of everything eatable and potable by the first invaders. Ahout the first of October, 1849, seven months after leaving home, we passed through the Golden Gate and stepped ashore upon the promised land.

My agent in San Francisco, to whom I had letters of introduction, and to whom I had consigned goods by sailing vessels around Cape Horn, xvas a merchant formerly of Honolulu, who was among the first to locate in San Francisco and take advantage of the tremendous business wave incident to the gold discovery. I found him very

OCR Results from 300 dpi Image File



again on the smooth Pacific. By a private arrangement with the steward I secured for a party of five a private room in a secret part of the ship, reached by a ladder from a smail scuttle, where we had a private table and an abundance of the best things on board regularly served. Meanwhile the first-class passengers were all day long elbowing one another and scrambling for their chance to get something from the cabin table. Off the coast of Lower California we saw one day a hundred and twenty whales of different kinds, one of which, about seventy-five or eighty feet long, swam just across our bow. At San Diego we were detained two days. The landing xvas three or four miles below the town, and as soon as the steamer was at anchor close to the shore there was a stampede of hungry passengers in the direction of the town in search of something to eat and drink. There was no hotel, but there were two or three stores, which were completely cleaned out of everything eatable and potable by the first invaders. About the first of October, 1849, seven months after leaving home, we passed through the Golden Gate and stepped ashore upon the promised land.

My agent in San Francisco, to whom I had letters of introduction, and to whom I had consigned goods by sailing vessels around Cape Horn, was a merchant formerly of Honolulu, who was among the first to locate in San Francisco and take advantage of the tremendous business wave incident to the gold discovery. I found him very

OCR Results from 600 dpi Image File

OCR Comparison of 6 PT Type

Acquisizione

Original Image File:

The above reproduction represents the first number of a periodical published by American emigrants belated in Panama in 1849. The original consisted of four pages, about six by ten inches in size, and was printed on light blue paper. So far as we are informed, but four numbers appeared, the others being published on the 3d, 10th, and 17th of March. This publication throws interesting side lights on the Panama trip, of which there is an account. Lists of arrivals are printed in each number. Washington's Birth-

Results of OCR from 600 DPI File:

The above reproduction represents the first number of a periodical published by American emigrants belated in Panama in I34~. The original cousisted of four pages, about six by ten incbes in size, andwas printed 3(8 ligbtblue paper. So far as we are informed, b,tfournumbers appeared, tim others being published onthe 3d, ioth, and ~ytb of March. This publication throws interesting side lights on the P3033333 trip, of which there is an account. Lists of arrivals are printed in each nu3ober. Washington's Birthday, it appears, was duly celebrated by American citizens with a procession and a banquet, and

Results of OCR from 300 DPI File:

Tias above osposslsalisaa osyoehs'slb lbs llsat 'sasatabse of A porisaliassi psbliblatal by Aassoita's sasigosaslla bslAlssl i's Pa'sassd isa 0849. 'lts soigisill hsahiblsal sf fsso psgsa, alassl ais by lsis inasbea in sias, and a'sas psisassal so lighl blass papso. las faa ad ate ate latfoossasal, basl fs'so ossasabass a1spsAssal. alas ohlastb issiog psblislssal s's lts vl~, aslt, asasi sylla of Massla. Tlsib p'sblisalista llaada'sh i'slssshlioag hide higlalb sta llts PsCalsA 10115, of

Rappresentazione di documenti testuali

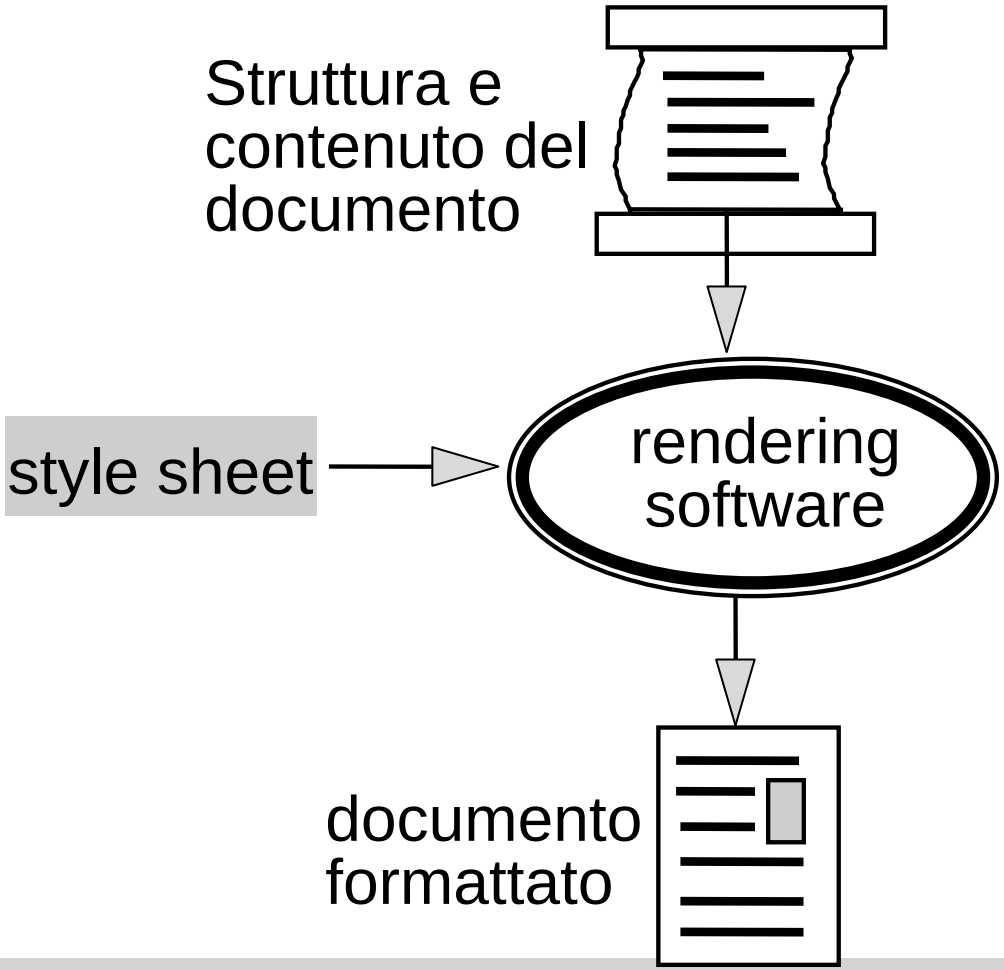
- **Struttura**
  - Descrive la divisione del testo in vari elementi sia fisici (caratteri, parole) che logici (titolo, autori, capitoli, ecc.)
  - La struttura viene spesso rappresentata da linguaggi di markup
  - Linguaggi di Markup
    - ➔ **SGML (Standard Generalized Markup Language)**
    - ➔ **HTML/XML**
- **Visualizzazione**
  - Descrive il modo in cui il documento viene visualizzato sullo schermo
  - Linguaggi di visualizzazione
    - ➔ **TeX, PostScript, PDF**

Testo

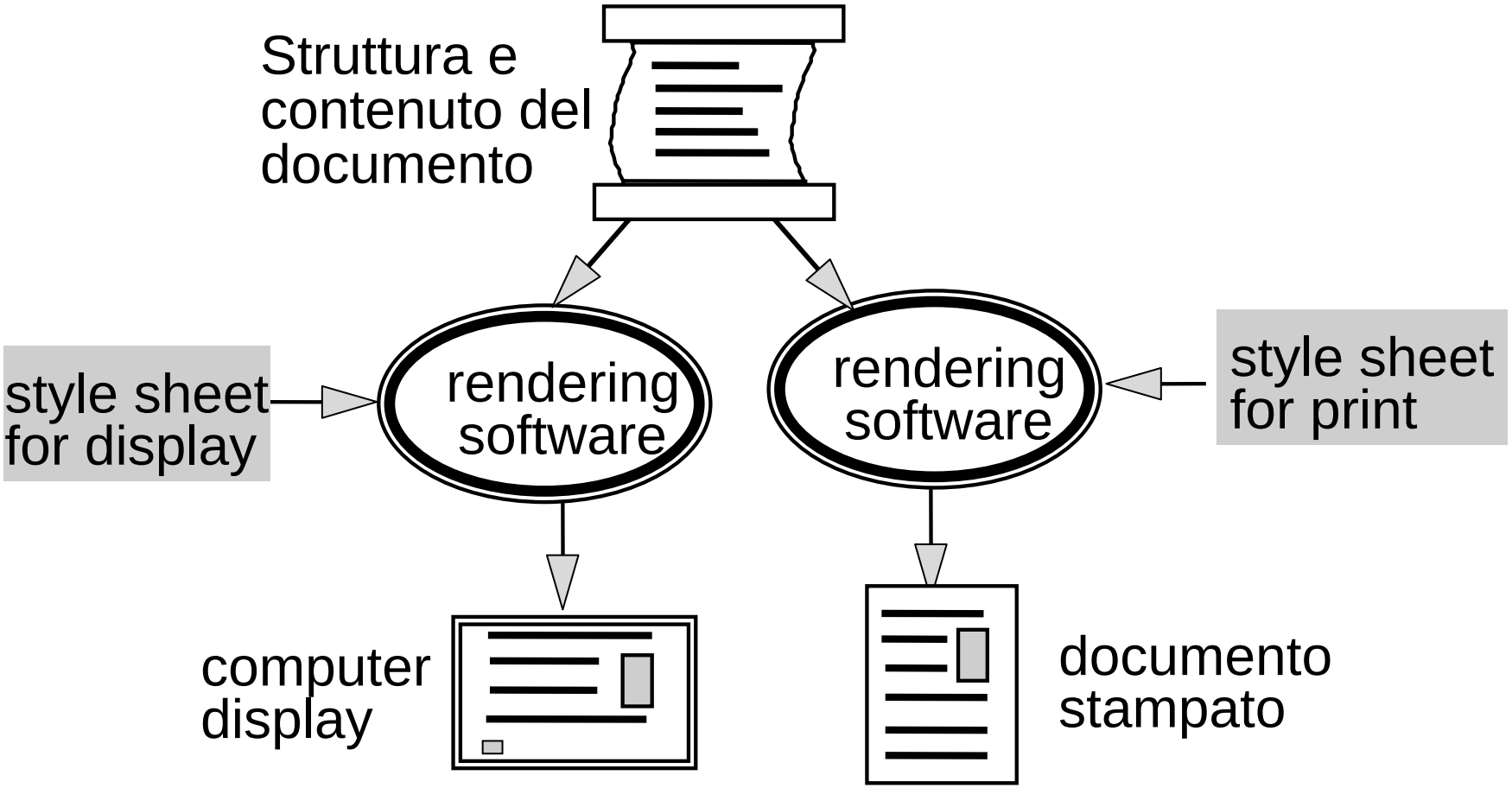
- **La ricchezza del testo**
  - **Elementi: lettere, scripts, simboli**
  - **Struttura: parole, frasi, paragrafi, titoli, tabelle**
  - **Presentazione: fonts, layout, disegni**
  - **Casi particolari: simboli matematici, musica**

- ***Le Biblioteche Digitali devono rappresentare tutte queste varianti***

Markup e Style Sheets



Alternative Renderings



Esempio di markup

```
<body><table border="0" cellpadding="0" cellspacing="0" width="100%">

  <tr>

    <td width="150" colspan="2" align="center"> </td>

    <td align="center">

    <!--webbot bot="Navigation" s-type="banner" s-rendering="graphics"
    S-Orientation B-Include-Home B-Include-Up U-Page S-Target --></td>

  </tr>

  <tr>

    <td width="100%" colspan="3" style="border-bottom-style: none;
    border-bottom-width: medium"></td>

  </tr>

</table><table border="0" cellpadding="0" cellspacing="0" width="100%">

  <tr>

    <td style="border-bottom-style: solid; border-bottom-width: 1"
    align="right">

  </tr>
```

Linguaggi di markup

SGML (Standard Generalized Markup Language)

Un linguaggio che permette la definizione di nuovi linguaggi di markup. Permette di rappresentare la struttura logica e di layout dei documenti

XML (eXtensible Markup Language)

Versione semplificata di SGML, utilizzata principalmente per la rappresentazione di informazione online

DTD (Data Type Definition)

Una specifica definizione di linguaggio di markup per una particolare classe di documenti rappresentati in SGML.

HTML (Hypertext Markup Language)

Un particolare linguaggio di markup con link ad altri oggetti

XML Example (Metadata)

```
<?xml version="1.0"?>
<!DOCTYPE dlib-meta0.1 SYSTEM "http://www.dlib.org/dlib/dlib-meta01.dtd">
<dlib-meta0.1>
    <title>Digital Libraries and the Problem of Purpose</title>
    <creator>David M. Levy</creator>
   <publisher>Corporation for National Research Initiatives</publisher>
    <date date-type = "publication">January 2000</date>
    <type resource-type = "work">article</type>
```

continued on next slide

XML Example (Metadata)

continued from previous slide

```
<identifier uri-type = "DOI">10.1045/january2000-levy</identifier>
   <identifier uri-type =
"URL">http://www.dlib.org/dlib/january00/01levy.html</identifier>
   <language>English</language>
   <relation rel-type = "InSerial">
      <serial-name>D-Lib Magazine</serial-name>
      <issn>1082-9873</issn>
      <volume>6</volume>
      <issue>1</issue>
   </relation>
   <rights>Copyright (c) David M. Levy</rights>
</dlib-meta0.1>
```

Page-Description Languages

- **Lo scopo è quello di presentare i documenti elettronici con una qualità simile a quella dei documenti a stampa**
- **I primi metodi di formattazione del testo erano specifici per la stampa**
- **Attualmente sono altrettanto importanti le problematiche relative alla visualizzazione su schermo**
- **Vedremo brevemente tre diversi strumenti**
  - TeX – Produzione e formattazione di document
  - PostScript – Stampa di alta qualità
  - Portable Document Format (PDF)

TeX

- **Linguaggio sviluppato agli inizi degli anni ’80 da Donald Knuth**
- **Al contenuto del documento vengono aggiunti una serie di comandi che danno le direttive di formattazione e visualizzazione.**
- **Contiene istruzioni specializzate per la notazione matematica**
- **Include un sistema specifico (Metafont) per il disegno di font**

PostScript

- **Linguaggio grafico sviluppato dalla Adobe Systems, utilizzato principalmente per la creazione di rappresentazioni grafiche di document da stampare**
- **Molti programmi di gestione documenti possono produrre una rappresentazione PostScript del documento da inviare a device di stampa**
- **Vi possono essere piccole variazioni dovute ai vari interpreti PostScript**
- **Utilizzato anche per la memorizzazione e lo scambio di documenti**

Portable Document Format (PDF)

- **Sviluppato dalla Adobe come linguaggio di memorizzazione di pagine di documenti in un formato portabile su diversi sistemi**
- **Utilizzato principalmente per documenti creati in forma elettronica**
- **Documenti acquisiti da scanner (bit-map) possono essere estremamente grandi in PDF**
- **Questo implica che in alcune situazioni il PDF può essere poco adatto ad essere usato nelle Biblioteche Digitali**
- **I lettori di file PDF sono gratuiti, mentre i programmi di generazione di PDF sono a pagamento**

Indicizzazione del testo

- **Il processo di indicizzazione associa (weighted) index terms ai documenti**

- **Gli Index terms possono essere**
  - Parole scelte all'interno di un vocabolario controllato
  - Parole estratte automaticamente
  - Frasi estratte automaticamente
  - Altri metadati

Indicizzazione del testo

- **Tratteremo il caso dell'indicizzazione automatica. In questo caso, si è verificato sperimentalmente che la migliore indicizzazione si ottiene usando termini singoli con un peso legato all'importanza del termine.**
  - È importante scegliere opportunamente il metodo di calcolo del peso di ogni termine.
- **L'indice viene utilizzato durante la fase di ricerca dei documenti.**
- **L'indice può essere utilizzato per determinare quali sono i documenti che contengono un termine o una lista di termini**
- **L'uso dei pesi dei termini può essere utilizzato per misurare il grado di similitudine tra i documenti trovati e l'interrogazione.**

Indicizzazione del testo

- Il processo di indicizzazione prevede i seguenti passi:
  - **Eliminazione delle parole più comuni (ad es. articoli, congiunzioni, ecc.) contenute in una stop-word list.**
  - **Riduzione delle parole rimanenti alla radice (ad es. i verbi sono trasformati nella forma all'infinito, i plurali in singolari, ecc.)**
  - **Le parole rimanenti vengono utilizzate come vocabolario**
- Supponiamo che i termini del vocabolario utilizzato siano $(t_1, \ldots, t_n)$
- Indichiamo con w_{ij} il peso (o rilevanza) del termine *i* nel documento *j*
- Il documento d_j sarà quindi rappresentato con il vettore

  $((t_1,w_{11}), \ldots , (t_i,w_{ij}), \ldots , (t_n,w_{nj}))$

Calcolo della rilevanza dei termini

- **Si possono utilizzare diversi metodi per il calcolo della rilevanza. Un metodo molto utilizzato è la cosidetta funzione *tfidf* (term frequency inverse document frequency) che si basa sull'assunzione che:**
  - Quante più volte un termine è presente in un documento, tanto più è significativo per quel documento. Quindi con "term frequency (*tf*)" indichiamo il numero di volte che un termine compare nel documento.
  - Se un termine è ripetuto in molti documenti risulta poco utile per discriminare un documento dall'altro. Si utilizza quindi una misura del numero di occorrenze del termine nell'intera collezione: la "document frequency (*df*)" misura il numero di documenti che contengono il termine.
- **Il peso del termine è dato da w = tf * log (N/df), dove N è il numero di documenti archiviati.**

Indicizzazione del testo

- Quindi l'indicizzazione dei documenti produce la seguente matrice.

	d_1	…	d_i	…	d_m
t_1	w_{11}	…	w_{1i}	…	w_{1m}
…	…	…	…	…	…
t_k	…	…	w_{ki}	…	…
…	…	…	…	…	…
t_n	w_{n1}	…	w_{ni}	…	w_{nm}

- Normalmente questa matrice viene acceduta creando un indice sui termini.

Ricerca di documenti testuali

- **Ricerca Esatta**
  - Vengono cercati tutti i documenti che contengono **esattamente** le parole specificate nell'interrogazione
  - Molto spesso i sistemi che permettono una ricerca esatta, consentono di formulare interrogazioni Booleane, nelle quali le parole che compongono l'interrogazione sono connesse con operatori AND, OR, NOT
  - Ad es: (Biblioteche OR Archivi) AND Digitali
  - In alcuni casi è possibile specificare che le parole devono essere vicine, oppure che devono far parte della stessa frase, ecc.

Ricerca di documenti testuali

- **Ricerca Approssimata**
  - Si eliminano le parole troppo comuni.
  - Le parole dell'interrogazione vengono ridotte alla radice
  - Vengono recuperati anche i documenti che non contengono tutte le parole presenti nell'interrogazione
  - Viene calcolata la rilevanza del documento per l'interrogazione. Spesso si usa una misura basata su tf*idf
  - I documenti trovati vengono ordinati in base alla rilevanza (viene fatto un ranking dei risultati)
  - Esempi di interrogazioni
    - ➔ **http://www.google.it/**

Ricerca dei documenti

- **Dopo aver formulato l'interrogazione si possono utilizzare diversi modelli per determinare quali sono i documenti rilevanti e quale è il grado di rilevanza.**
  - Modello Booleano
  - Modello basato sull'uso di una logica Fuzzy
  - Modello Vettoriale
  - …

Modello Booleano

- **Una interrogazione può contenere gli operatori logici and/or/not**
  - L'interrogazione "digital and library" recupera tutti i documenti associati con entrambi i termini presenti nell'interrogazione.
  - L'interrogazione "digital or library" recupera tutti i documenti associati con almeno uno dei termini presenti nell'interrogazione.
  - L'interrogazione "library and not digital" recupera tutti i documenti che contengono la parola "library" ma non contengono la parola "digital".
- **Si possono costruire interrogazioni molto complesse tramite la combinazione degli operatori And, Or, Not.**

Fuzzy logic model

- **Estende il modello Booleano in modo da assegnare un peso ai documenti ritrovati.**
- **Supponiamo che un documento D contenga i termini t_1 and t_2, rispettivamente con i pesi w_1 and w_2**
- **Consideriamo l'interrogazione t_1 and t_2**
  - La rilevanza di D è data da $\min\{w_1,w_2\}$
  - Se la query contiene molti termini, tutti rilevanti tranne uno, la rilevanza di D risulta determinata dall'unico termine poco significativo
- **Consideriamo l'interrogazione t_1 or t_2**
  - *La rilevanza di D è data da* $\max\{w_1,w_2\}$
  - Se la query contiene molti termini, tutti poco rilevanti tranne uno, la rilevanza di D risulta determinata dall'unico termini molto rilevante.

Modello vettoriale

- **I documenti e le interrogazioni sono rappresentati con vettori dei termini, per ognuno dei quali viene specificato il peso.**

  $d_j = ((t_1,w_{11}), \ldots , (t_i,w_{ij}), \ldots , (t_n,w_{nj}))$

  $Q = ((tq,wq_1), \ldots , (tq_i,wq_i), \ldots , (tq_n,wq_n))$

  dove i w_{ij} ed i wq_i sono i pesi dei termini nel documento e nell'interrogazione

- **La similarità tra il documento d_j e la query Q, è data dal prodotto scalare dei due vettori**

$$sim(Q,d_j) = \frac{\sum_{i=1}^{n} w_{ij} \times wq_i}{\sqrt{\sum_{i=1}^{n} (w_{ij})^2} \times \sqrt{\sum_{i=1}^{n} (wq_i)^2}}$$