InformaticaUmanistica

Biblioteche Digitali

Pasquale Savino

ISTI - CNR

UNIVERSITÀ DI PISA

Obiettivi del corso

Il corso ha lo scopo di fornire le **basi teoriche e sperimentali** relative alle tecniche ed alle metodologie per la **organizzazione, creazione e gestione** di una Biblioteca Digitale.

- Panoramica generale del settore delle Biblioteche Digitali, con una visione approfondita di alcuni degli aspetti più rilevanti
- Analisi delle aree di ricerca più promettenti nel settore
- Analisi dei diversi utilizzi applicativi delle Biblioteche Digitali, evidenziando in particolare le applicazioni al settore dei Beni Culturali
- Utilizzo sperimentale delle Biblioteche Digitali, evidenziando le problematiche relative alla loro creazione e gestione.

Programma del corso

- **Introduzione alle Biblioteche Digitali**
- **Metadati**
- **Esempi di Biblioteche Digitali (con esercitazioni)**
- **Architettura e tecnologie di base delle Biblioteche Digitali**
- **Progettazione di una Biblioteca Digitale (con esercitazioni)**
- **Nuove tendenze**

Materiale didattico

- **Lucidi, disponibili sul sito Web del corso**
- **Ian Witten, David Bainbridge, “How to Build a Digital Library”, Morgan Kaufmann Publishers**
- **Michael Lesk, “Practical Digital Libraries”, Morgan Kaufmann Publishers**
- **William Y. Arms, “Digital Libraries”, The MIT Press**

Lezioni e ricevimento

- **Lezioni**
  - Giovedì – 10:15 – 12:00
  - Venerdì – 12:15 – 13:00
- **Ricevimento**
  - Venerdì – 15:30 – 17:00 previo appuntamento, presso ISTI-CNR, Via Moruzzi, 1 (Area della Ricerca – San Cataldo)

InformaticaUmanistica

Introduzione alle Biblioteche Digitali

UNIVERSITÀ DI PISA

Sommario [1/2]

- **Cenni storici**
  - Vannevar Bush
  - Dalle Biblioteche ai Cataloghi Automatizzati
  - Gli OPAC accessibili via Web
  - Le Biblioteche Digitali
- **Cos'è una Biblioteca Digitale**
  - Definizione
  - Confronto tra BD e database, sistemi IR, WWW, biblioteca tradizionale
  - Vantaggi delle BD
  - Alcuni esempi di Biblioteche Digitali

Sommario [2/2]

- **Cosa ha permesso la nascita delle Biblioteche Digitali**
  - Evoluzioni tecnologiche
- **Tipologie di Biblioteche Digitali**
  - Biblioteche Pubbliche e Biblioteche Specializzate
  - Tipi di documenti trattati
    - ➔ **Libri**
    - ➔ **Documenti testuali**
    - ➔ **Immagini**
    - ➔ **Audio/video**
    - ➔ **……..**

Cenni storici

Vannevar Bush (1890-1974)

- **Direttore dell'US Office of Scientific Research and Development**
- **Ha predetto diverse evoluzioni tecnologiche**
  - L'idea del "MEMEX" (1945) conteneva molte delle idee sulle quali si basano il Web e le Biblioteche Digitali
  - Il MEMEX aveva lo scopo di fornire ai ricercatori la possibilità di scambiarsi facilmente informazioni e di avere accesso alla totalità delle conoscenze disponibili

Memex

- **Integrazione di computer, tastiera e scrivania**
- **"archivio meccanizzato privato e biblioteca"**
  - Eliminazione degli aspetti che rendevano ripetitivo e inefficace il retrieval di informazione
  - Implementazione basata sull'uso di microfilm
- **Indicizzazione associativa**
  - "il processo di legare due elementi insieme è l'aspetto più importante"
  - Preludio agli ipertesti...

Memex

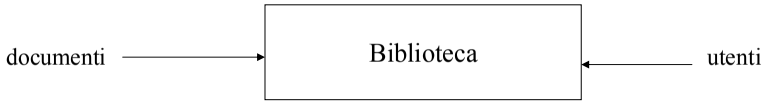
- **L'informazione poteva essere indicizzata utilizzando associazioni tra I vari elementi, considerando anche l'adattamento alle esigenze degli utenti**
  - WWW non lo fornisce ancora oggi
- **Bush osservava che I nuovi strumenti modificano il modo in cui facciamo le cose ed espandono le nostre esigenze**
  - L'impatto del WWW e delle Biblioteche Digitali non è ancora completamente noto
- **L'articolo di Bush non prevedeva l'uso di sistemi di ricerca free-text**
  - knowledge trails only; Yahoo minus keyword searching

Dalle Biblioteche tradizionali alle Biblioteche Digitali

- **Biblioteche tradizionali**
  - Inizialmente erano luoghi dedicati alla conservazione dei documenti
  - Con l'aumento della quantità di documenti immagazzinati, si è reso necessario creare degli strumenti di ricerca efficaci
  - L'avvento dei calcolatori ha permesso di automatizzare e rendere più efficienti ed efficaci gli strumenti di ricerca
- **Biblioteche digitali**
  - I documenti stessi sono in forma digitale
    - ➔ **Ricerca basata sul contenuto**
    - ➔ **Conservazione degli oggetti digitali**
    - ➔ **Protezione da accessi indesiderati, ecc.**

Le Biblioteche tradizionali

- **La biblioteca svolge il ruolo di "mediatore" tra gli oggetti "portatori di informazione" (documenti) e gli "utilizzatori dell'informazione"**



- **DOCUMENTO: Qualsiasi oggetto utilizzabile a fini di consultazione, ricerca, informazione**

Funzioni delle biblioteche

- **SELEZIONE**
- **ACQUISIZIONE**
- **DESCRIZIONE**
- **ACCESSO ====>**
  Strumenti per la ricerca dei documenti o delle informazioni sui documenti
- **CONSERVAZIONE**

Strumenti per la ricerca

- ANTICAMENTE: Ordinamento fisico dei documenti
- NELL'EPOCA MODERNA: Ordinamento delle descrizioni dei documenti

=====> I CATALOGHI

Strumenti per la ricerca

- **Ordinamento fisico**
- **I documenti sono collocati secondo un certo criterio:**
  - ➔ **Data di "arrivo" (liste inventariali)**
  - ➔ **Soggetto (classificazione)**
  - ➔ **Tipo di documento**

  ...*e possono essere ricercati soltanto in base a quel criterio*

Strumenti per la ricerca

- **Il catalogo:**
  - Ciascun documento è descritto con un insieme di elementi significativi, scelti secondo le regole della descrizione bibliografica (Titolo, autori, … soggetto, … collocazione); la descrizione è riportata su una o più schede
  - Ciascuna scheda è intestata con uno degli elementi bibliografici ritenuti utili a ricercare il documento ====> Punti di accesso
  - Le schede intestate sono ordinate nel catalogo secondo i punti di accesso
  -
- ***Il catalogo permette la ricerca di un documento conoscendo uno qualsiasi dei suoi punti di accesso***

La varieta' delle -teche

[depositi==>collezioni, raccolte]

- **Biblioteche**
- **Emeroteche**
- **Cineteche**
- **Discoteche**
- **…….===> Mediateche**
- **Archivi**
- **Musei**

La varieta' delle -teche

- **Ciascuna "-teca" raccoglie e organizza una speciale tipologia di documenti**
- **Ciascuna tipologia di documenti ha proprie regole di descrizione**

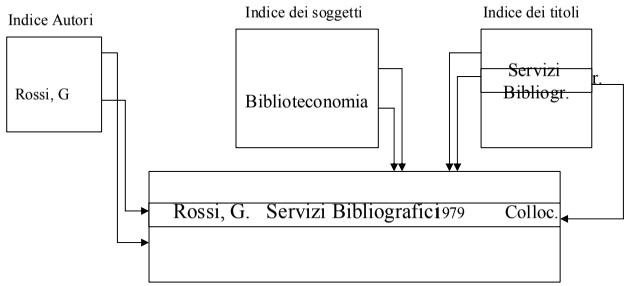
…..ma rimangono valide le regole di organizzazione dei cataloghi

I cataloghi “automatizzati”

- **L’uso dei calcolatori ha reso le prestazioni dei cataloghi più potenti e più flessibili**

I cataloghi “automatizzati”

- **Indici :** liste ordinate **di elementi particolari della descrizione bibliografica,**
- **utili a identificare il documento di interesse : Autori - titoli - soggetti - collane -**



Archivio di “record” bibliografici

I cataloghi “automatizzati”

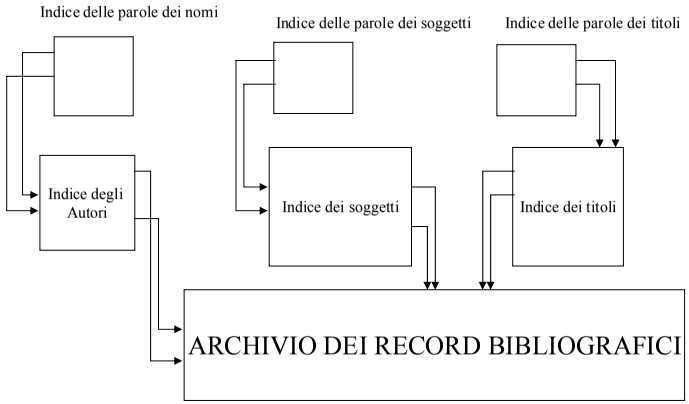
- Nei cataloghi a schede si può ricercare un documento solo attraverso uno dei suoi punti di accesso

- Nei cataloghi automatizzati, invece, si possono fare interrogazioni definendo più punti di accesso:

AUT = **Rossi, G.**
AND
TIT = **Servizi bibliografici**

I cataloghi “automatizzati”

I cataloghi di ultima generazione costruiscono “indici degli indici”



I cataloghi "automatizzati"

Ricerche con dati non certi:

AUT = **Rossi** AND TIT = **Servizi**

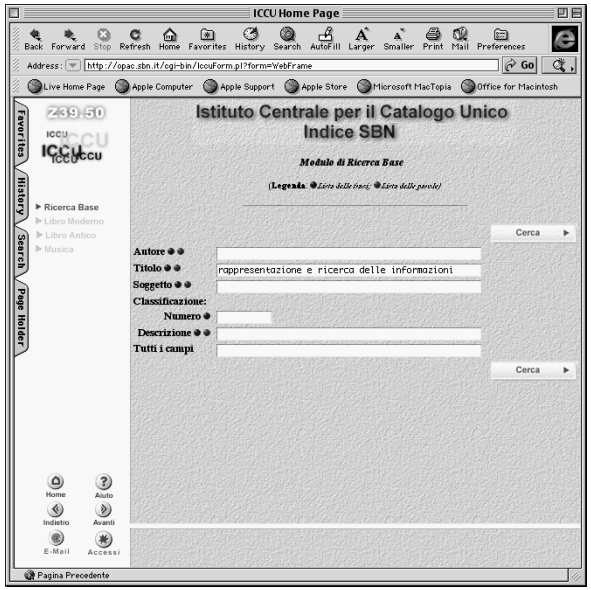
TIT = **bibliogr#**

PAROLA = **biblio#** AND DATA > **1999**

Il catalogo delle biblioteche automatizzate

- **INTERFACCIA UTENTE-SISTEMA:**
- Il modo in cui il sistema informativo si presenta all'utente per istruirlo e guidarlo a svolgere le operazioni

Il catalogo del Servizio Bibliografico Nazionale: WWW.sbn.it



Cos'è una Biblioteca Digitale

Definizione informale

Una Biblioteca Digitale è una collezione organizzata di oggetti digitali accessibili in rete, ed un insieme di servizi che ne permettono la conservazione, l'accesso e la ricerca, oltre che l'organizzazione e la manutenzione della collezione. La collezione può contenere dati di tipo diverso, quali ad esempio testi, immagini, audio, video, ecc.

Cos'è una BIBLIOTECA DIGITALE ? OGGETTI DIGITALI

Dalle descrizioni dei documenti

===> ai DOCUMENTI

Dai cataloghi ====> ai **"DEPOSITI" distribuiti sulla rete contenenti:**

- **Oggetti digitali (testi, suoni, immagini tridimensionali, video, fotografie, film, …)**
- **Metadati (descrizioni degli oggetti)**

Gli oggetti digitali

- **Una Biblioteca Digitale permette di archiviare "documenti" contenenti vari tipi di dati ……**
  - Testo
  - Immagini
  - Video
  - Audio
  - 3D objects
  - Virtual-reality worlds
  - ……..
  - Composizione dei tipi di dati precedenti

I principali servizi di una Biblioteca Digitale

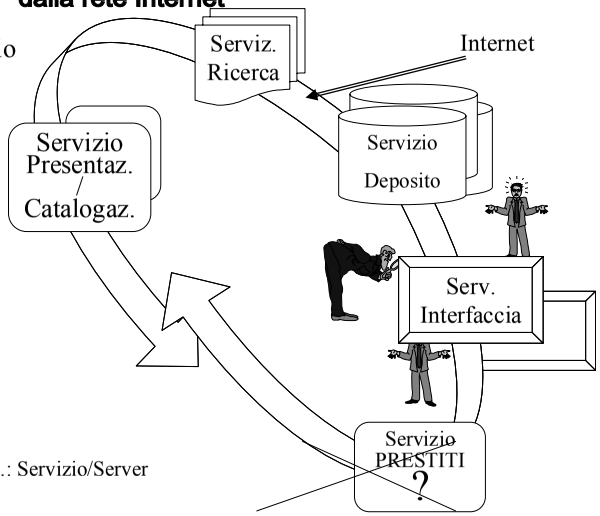
- **Accesso e recupero**
  - Cataloghi
  - Riferimenti
  - Indici
- **Conservazione**
- **Gestione**
  - Controllo dell'accesso
  - Condivisione dei dati
  - Gestione della collaborazione tra gli utenti
    - ➔ **E.g. collaborative filtering, catalogazione, …..**
  - ……
- **Una Biblioteca Digitale dovrà fornire almeno tutti I servizi offerti da una Biblioteca tradizionale**

Che cos'è una Biblioteca Digitale?

Un sistema distribuito per la gestione di risorse digitali accessibili dalla rete Internet

Gli utenti accedono al servizio Interfaccia WEB attraverso il Protocollo HTTP

La comunicazione fra i servizi del sistema avviene attraverso la rete,con l'uso di un protocollo specifico del sistema di DL



N.B.: Servizio/Server

In cosa differisce una Biblioteca Digitale da un DataBase

- **Un database tradizionale SQL ha la relazione come contenitore base dell'informazione:**

```
select name
from employee, project
where employee.deptnumber = "25" AND
        project.number = "100"
```

- **I database sfruttano strutture e relazioni note**
- **Il retrieval dei DBMS non è probabilistico**

In cosa differisce una BD da un sistema di IR tradizionale?

- **La differenza è meno netta che per I DBMS**
- **I sistemi IR systems possono essere considerati precursori delle BD**
- **I sistemi di IR si sono occupati tradizionalmente di documenti testuali ma loro evoluzioni trattano anche documenti multimediali**
  - Match esatto - Boolean, text pattern searching
  - Match non esatto - probabilistic, vector space, clustering
- **Le BD possono essere considerate un superset dei sistemi IR**

In cosa differisce una BD dal WWW?

- **La differenza principale è legata all'organizzazione dei documenti**
  - Il WWW non impone alcuna organizzazione dell'informazione, mentre nelle BD tutte le operazioni sono soggette a procedure ben definite.
  - In particolare, nelle BD sono definiti precisi schemi di metadati che permettono l'accesso ai documenti
- **Alcuni motori di ricerca del web (Yahoo, Lycos) tentano di aggiungere una qualche organizzazione ai documenti trattati**
  - Comunque, non tutti i documenti del web sono gestiti
  - La maggior parte dei motori di ricerca si basa su text search (Altavista, Google)

In cosa differisce una BD dal WWW?

- **Un'altra differenza sostanziale è legata al controllo degli inserimenti**
  - I documenti nel web possono essere inseriti da chiunque, mentre in una DB l'inserimento è permesso solo a particolari utenti
  - I motori di ricerca selezionano i documenti da indicizzare tra quelli presenti nel web, mentre nelle BD tutti i documenti vengono indicizzati sulla base dei criteri definiti
- **Le DB sono soggette ad un maggior controllo (per gli inserimenti, gli accessi e le ricerche) del WWW, ed hanno un insieme di utenti ben preciso**

Quali sono le differenze tra una DB ed una Biblioteca Tradizionale (BT)?

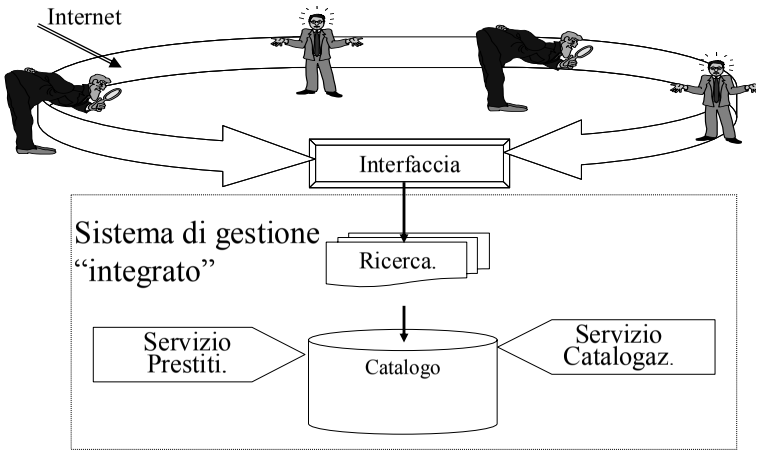
- **Le BT gestiscono oggetti (documenti) fisici**
  - Anche se le BT utilizzano delle schede elettroniche per individuare i documenti, questi si trovano in una ben precisa posizione fisica
  - Questo porta a delle ovvie implicazioni
    - ➔ **Gli oggetti possono esistere solo in un luogo**
    - ➔ **Un solo utente per volta può accedere all'oggetto**
    - ➔ **L'oggetto può essere acceduto solo recandosi fisicamente presso la Biblioteca o attraverso meccanismi di distribuzione postale**

Quali sono le differenze tra una DB ed una Biblioteca Tradizionale (BT)?

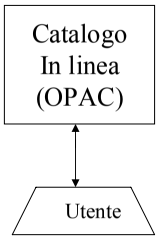
- **Biblioteca Digitale**
  - Rimuovendo la restrizione della esistenza fisica degli oggetti si ottengono notevoli vantaggi
    - ➔ **Accessi multipli, trasmissione elettronica**
  - ma anche alcune complicazioni ……
    - ➔ **Proprietà intellettuale, diritti di accesso, etc.**
- **Una BT offre anche vantaggi dal punto di vista sociale ed educativo**
  - Molte BT offrono servizi aggiuntivi che non possono essere offerti da una BD (ambiente di conversazione, comunicazione tra i lettori, …) …. almeno per ora

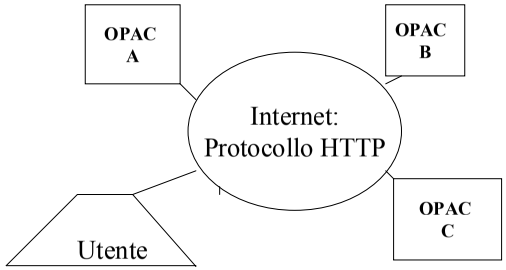
Cos'è una BIBLIOTECA DIGITALE ?

Analogie con un sistema di gestione per biblioteche "tradizionali"



Dalla comunicazione "diretta"……



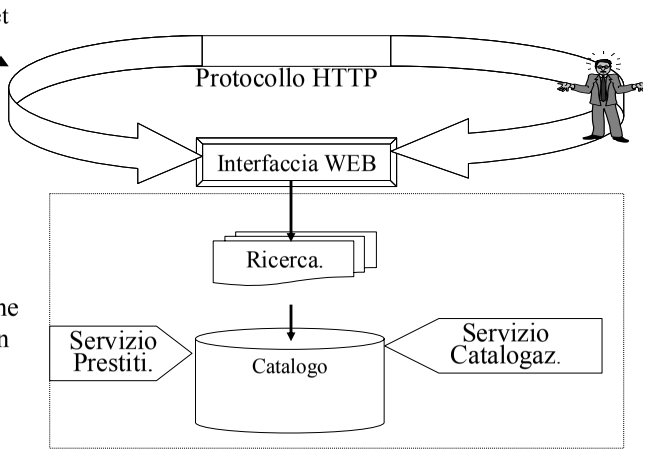


….alla comunicazione in rete via WEB: protocollo HTTP

.

Gli utenti generici accedono via WEB soltanto ai servizi di Informazione, Ricerca/Recupero

I vari moduli software che realizzano i "servizi" non hanno esigenze di comunicazione in rete



Cos'è una BIBLIOTECA DIGITALE ?

Analogie con un sistema di gestione per biblioteche "tradizionali"

I servizi "di base" di una biblioteca digitale:

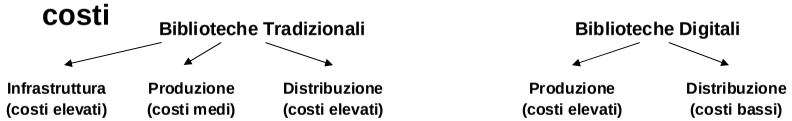
- Interfaccia
- Presentazione degli oggetti digitali ("Acquisto" e catalogazione)
- Deposito
- Ricerca/Browsing/Recupero degli oggetti digitali

Vantaggi delle Biblioteche Digitali

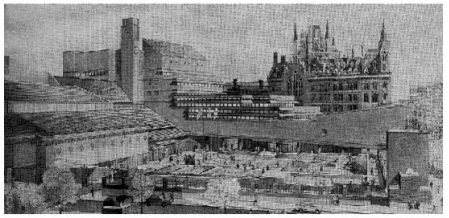
- **Vantaggi per gli utenti**
  - Si costruisce un BD nella convinzione di poter fornire una migliore distribuzione dell'informazione
    - ➔ **Informazione disponibile dove è necessario**
    - ➔ **Maggiori quantità di informazione disponibile**
    - ➔ **Possibilità di selezionare facilmente quello che interessa**
    - ➔ **Possibilità di utilizzare media diversi (testo, immagini, audio, video, ecc.)**
    - ➔ **L'informazione può essere condivisa**
    - ➔ **L'informazione è sempre aggiornata**
    - ➔ **Accesso 24/24**

Vantaggi delle BD

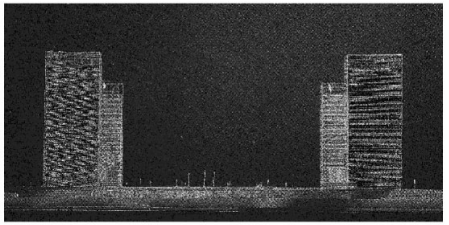
- **Vantaggi economici**
  - Le biblioteche convenzionali risultano sempre più costose
    - ➔ **Infrastrutture**
    - ➔ **Personale**
    - ➔ **Pubblicazione**
  - Attualmente anche le BD hanno costi elevati ma …..
    - ➔ **I costi sono destinati a scendere, in particolare i costi di archiviazione e distribuzione**
- **Le BD richiedono la definizione di nuovi modelli dei costi**



Each of these libraries cost more to build than the cost of scanning its books



The British Library, London £450M



The Bibliothèque de France, Paris FFr 5B

from Lesk, http://community.bellcore.com/lesk/columbia/session1/

Alcuni esempi di Biblioteche Digitali

- **Esempi di cataloghi di Biblioteche Digitali (Chemical Abstracts, Inspec, Biblioteca del CNR di Pisa)**
- **Biblioteche Pubbliche**
  - Library of Congress
  - New York Public Library
  - Boston Public Library
- **Biblioteche Specializzate**
  - NTRSDL
  - Documentari storici (DB ECHO)

Cosa ha permesso la nascita delle Biblioteche Digitali

Un po' di storia

- **I primi tentativi di archiviare informazione bibliotecaria con i computer datano dalla fine degli anni '60. I problemi riguardavano**
  - Alto costo dei computer
  - Interfaccia Uomo/Macchina scadente
  - Mancanza di reti di comunicazione
- **Primi risultati**
  - Library of Congress ha memorizzato le schede MARC (formato per machine readeable cataloguing)
- **Architettura dei sistemi**
  - Piccole quantità di informazione memorizzata su un computer centrale
  - Gli utenti erano connessi a terminali remoti di bassa qualità e con bassa velocità di connessione al computer centrale

Evoluzioni tecnologiche [1/3]

- **Archiviazione elettronica sta diventando sempre più economica rispetto alla carta**
  - Gli edifici delle Biblioteche tradizionali impegnano circa ¼ dei costi
  - Le Biblioteche tradizionali hanno problemi di espanzione (alti costi, tempi lunghi, ecc.)
  - I costi di archiviazione elettronica tendono a diminuire (circa 30% per anno)

Evoluzioni tecnologiche [2/3]

- **Miglioramento della qualità dei display**
  - Miglioramento della risoluzione
  - Disponibilità di software di visualizzazione
  - Disponibilità di standard di visualizzazione
- **Reti ad alta velocità**
  - Connessione diffusa
  - Aumento della velocità dei link della rete
  - Aumento della velocità delle connessioni locali
  - In alcuni paesi è più facile (e veloce) ricevere informazione tramite la rete internet che a stampa

Evoluzioni tecnologiche [3/3]

- **Accesso alla Biblioteca**
  - Le Biblioteche tradizionali sono accessibili solo agli utenti dell'organizzazione
    - ➔ **Esistono organizzazioni con biblioteche molto fornite (per es. centri medici specializzati) ma molti utenti non possono accedervi**
  - L'accesso ad una Biblioteca Digitale richiede costi sempre più bassi
    - ➔ **Meno di 1000 Euro per il computer**
    - ➔ **Meno di 10 Euro/mese per la rete**

Tipologie di Biblioteche Digitali

Tipologie

- **Biblioteche Pubbliche e Biblioteche Specializzate**
- **Gestione letteratura "white" e "gray"**
- **Gestione di vari tipi di dati**

Biblioteche Pubbliche e Specializzate

- **Una Biblioteca Pubblica prevede che utenti con interessi diversi possano accedere all'informazione**
  - L'accesso è comunque controllato
  - Gli utenti sono costituiti dal grande pubblico
  - I documenti trattano di argomenti diversi
- **Le Biblioteche Specializzate hanno le seguenti caratteristiche**
  - L'insieme degli utenti è piccolo e con interessi molto focalizzati
  - Analogamente i documenti riguardano argomenti molto focalizzati
  - È importante controllare efficacemente l'accesso (utenti non autorizzati, possibilità di visionare gli oggetti ma non di copiarli, ecc.)

Biblioteche Pubbliche e Specializzate

- **Esempi di Biblioteche Specializzate**
  - tradizionali - NASA LaRC Technical Library
  - digitali - NASA Technical Report Server, ACM Digital Library, ETRDL
- **Biblioteche Pubbliche**
  - tradizionali – Biblioteca comunale di Pisa
  - digitali – Yahoo, Boston public library

White and Grey Literature

- **La distinzione tra le due non è sempre molto chiara**
- **La definizione fornita da Grey Net:**
  - "that type of publication unavailable through normal book-selling channels, often produced in small quantities with limited distribution, promotion, and exploitation"
  - http://www.greynet.org/pages/1/index.htm

White and Grey Literature

- **Grey Net ammette comunque che la pubblicazione elettronica ha cambiato questa definizione, che andrà quindi sostituita**
- **Intuitivamente la letteratura**
  - White: autore e publisher sono di solito diversi, il lavoro è stato revisionato in modo indipendente, l'opera può essere ottenuta facilmente
  - Grey: è possibile che non sia stato revisionato; spesso viene pubblicato direttamente dall'autore o dalla sua organizzazione; può essere difficilmente reperibile

Esempi

- **White**
  - Riviste, libri, proceedings di conferenze, etc.
- **Grey**
  - Rapporti tecnici, rapporti governativi, etc.

Gestione di vari tipi di dati nelle BD

- **Libri**
- **Documenti testuali**
- **Immagini**
- **Audio/video**

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 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

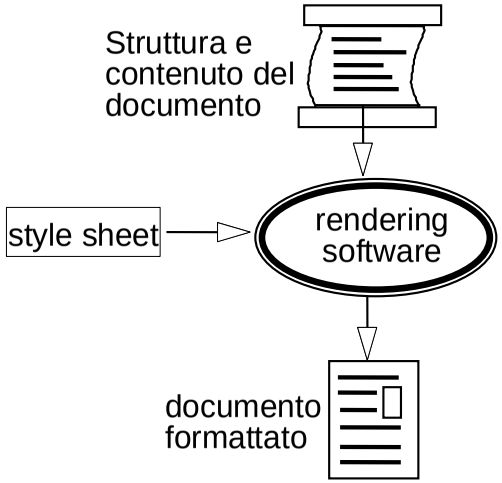
Rappresentazione di documenti testuali

- **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

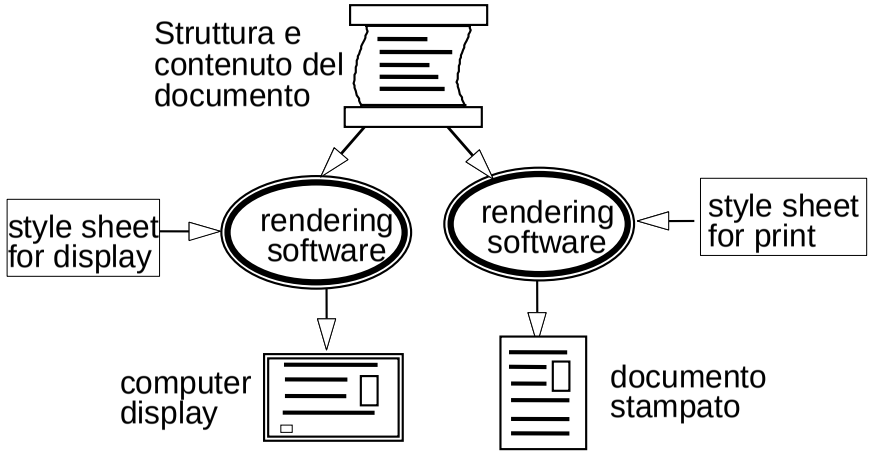
Testo

- **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



Markup Languages

SGML (Standard Generalized Markup Language)

A system for creating markup languages that represent the structure of a document

XML (eXtensible Markup Language)

A simplified version of SGML intended for use with online information

DTD (Data Type Definition)

A markup specification for a class of documents, defined within the SGML framework

HTML (Hypertext Markup Language)

A markup and formatting language with links to other objects

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**

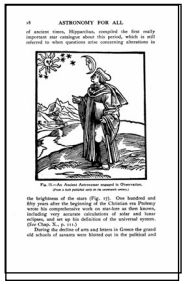
Portable Document Format (PDF)

- 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

Acquisizione di documenti come immagine

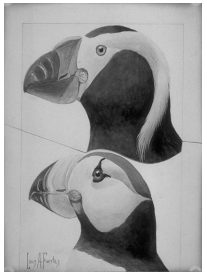
- Le singole pagine sono acquisite come immagini tramite uno scanner
- Ogni singola pagina viene rappresentata come una sequenza di punti (pixels)
- Ad ogni pixel viene assegnato un valore (nero, bianco, grigio, colore), rappresentato con un codice binario
- Si possono applicare tecniche di compressione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

Metodi di scanning

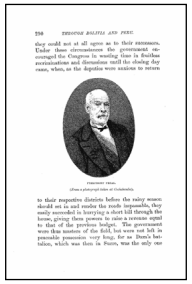


Bitonal

Grayscale



Color



Special Treatment

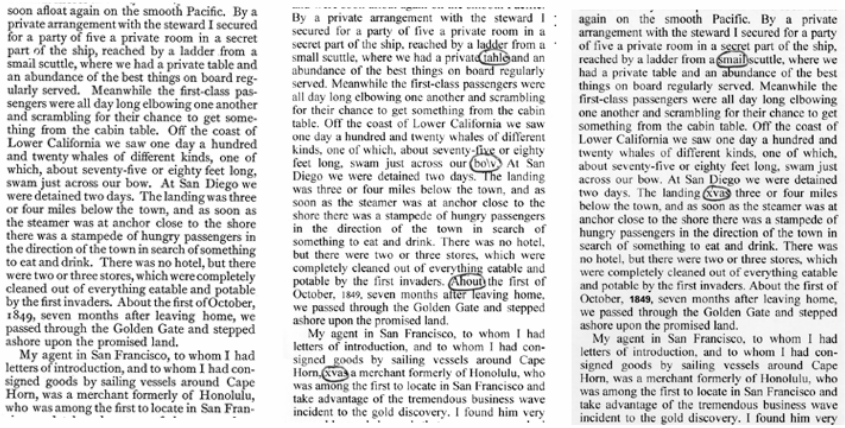
Qualità dell'immagine acquisita

- **Dipende da**
  - Risoluzione dello scanner (numero di pixel per cm^2)
  - bit depth (numero di bit per per pixel)
  - image enhancement
  - color management
  - compression
  - system performance
  - operator judgment and care

Riconoscimento caratteri

- Processo di trasformazione di una immagine in testo
- 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

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 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 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 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 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

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

Gestione del video

- Perchè è importante poter gestire biblioteche digitali di audiovisivi
- Caratteristiche specifiche dell'audio/video
- Applicazioni delle biblioteche digitali audio/video
- Alcuni esempi di biblioteche digitali audio/video

The importance of video

- **Video can be considered today the primary information and communication channel, due to**
  - Richness in information contained
  - Appeal
- **Video libraries will become essential in many application fields**
  - Personal information
  - Distance learning
  - Telemedicine
  - ……

Video characteristics

- **High video production vs print production**
  - TV stations produce 50 Million hours of video per year (25,000 TB)
  - Newspapers and periodicals produce less than 200 TB of data per year
- **Storage and transmission problems**
  - Video is usually compressed
- **Richness in content**
  - Difficulties in automatic extraction of content description

Services of A/V Digital Libraries

- **Digital Video Libraries are more complex than traditional DLs; they require the integration of several specialized technologies**
- **They offer the same services of text digital libraries**
- **Specific characteristics of Indexing and retrieval services**
  - Indexing based on the integration of different technologies for the automatic feature extraction
  - Integration of manual and automatic indexing
  - Retrieval based on different video features

Characteristics of an Audio/Video DL

The need of A/V DLs

- **Nowadays, video is present in many situations**
  - TV broadcasting
  - Professional applications, such as medicine, journalism, advertising, education, training, surveillance, etc.
  - Movies
  - Historical videos
  - Personal videos
- **The combination of audio and video is a very powerful communication channel**
  - approximately 50% of what is seen and heard simultaneously is retained

Advantages of A/V DLs

- **Most of the video material produced is used only once, due to the difficulty to archive it, to protect and to retrieve.**
- **A large video library of distributed and network searchable videos would enable**
  - Preservation of precious and expensive material
  - Reduction of production costs for new videos, through the reuse of existing material
  - Diffusion of knowledge

In general, it will enable the access to information that could have been lost.

A/V vs traditional DLs [1/2]

- **Library creation**
  - Traditional DLs, contain text documents
    - ➔ **Library creation requires automatic acquisition of text, extraction of document content, and indexing**
    - ➔ **This process is well known and many different techniques have been developed**
  - Video is extremely rich in "content" but ….
    - ➔ **the indexing of video content is difficult, expensive, and extremely dependent from the user and the application**
    - ➔ **A possible approach consists in an appropriate integration of automatic content extraction (e.g. speech recognition, image analysis, etc.) and manual indexing**

A/V vs traditional DLs [2/2]

- **Library exploration**
  - Traditional DLs, contain text documents
    - ➔ **Library exploration requires simple interfaces to formulate queries on free text and document metadata.**
  - Video libraries should permit
    - ➔ **To formulate queries on many different "dimensions"**
      - **Text, as extracted from speech and captions**
      - **Images extracted as key frames**
      - **Motion information**
      - **Other features automatically extracted**
      - **Metadata provided manually**

Applications of Audio/Video DL

Who may use A/V DLs?

- **We consider four main categories**
  - Large companies
    - → **Large corporations that may use Digital Video for their internal business, for advertising, promotion, etc.**
  - Media and entertainment
    - → **The most traditional area. Video is one of the key assets.**
  - Education
    - → **Video recording of courses**
    - → **Video used as course material**
  - Others
    - → **Health and medicine**
    - → **Government**
    - → **Surveillance**
    - → **Etc.**

Large companies

- **Audio/video digital libraries are used for**
  - Sales
  - Product launches
  - Marketing
  - Relation with investors
  - Product design (acquisition and analysis of customer's needs)
  - Support for online sales
  - Video archives for internal use
  - Special services for customers, such as web access to specialized video archives, e.g.
    - → **News**
    - → **Economic information**
    - → **Products**
    - → **Materials**
    - → **Etc.**

Media & Entertainment [1/3]

- **Broadcasting companies**
  - Many broadcasters are creating and distributing video programs on the web. A video archive is very helpful to them to add a new service for accessing old video material.
  - Examples:
    - → **ABC News**
    - → **Mediaset**
    - → **RAI**

- Archive of old programs
- Archive of daily programs
- Additional material w.r.t. tv programs

Media & Entertainment [2/3]

- **Video archives**
  - Many national and private organizations own old video material. The digitalization and archiving of this material is beneficial for content owners (for example, they can promote the use of their material) and for users belonging to different categories: e.g. professional users (that need the material to produce their video programs) or researchers or general public.
  - Examples:
    - ➔ **Istituto Luce**

Media & Entertainment [3/3]

- **Movie production companies**
  - Many large movie production companies own a large amount of video material, composed of the films and of related material, such as cuts not used in the final film version, interview, video trials, etc. This material is very helpful for many purposes, from the production of DVD version of the film up to the critical study of the video. Providing access to the general public of this material is also a powerful promotion and advertising channel.
  - Examples:
    - ➔ **MGM**
    - ➔ **20th Century Fox**

Education

- **Digital video used for different purposes**
  - Promotion and advertising
    - ➔ **Online preview of training content**
  - Store and distribute the video courses
    - ➔ **Remote access of the courses**
    - ➔ **Keep track of classroom discussion**
  - Used as course material
    - ➔ **Delivery of video clips to students, either online or in the classroom**
    - ➔ **From remote sites provide students and teachers with on-demand, searchable access to whole programs and video clips**
    - ➔ **Free search and access to the video library con be used by students to find answers to specific questions, to study in depth some topics, etc.**
  - Production of new courses
    - ➔ **Improve the course production procedures, allowing teachers and producers to remotely access the video library**
  - Examples:
    - ➔ **Princeton University**
    - ➔ **Harward Business School**
    - ➔ **University of Arizona**

Other Applications [1/2]

- **Health and medicine**
  - Health and social care info to the general public
  - Information to physicians for special purpose medical procedures
  - Training

Other Applications [2/2]

- **Government**
  - Enhancement of the governmental decision making process, by recording and archiving of public meetings and discussion.
- **Surveillance**
  - A large amount of video is produced for surveillance purposes.
    - ➔ **Required automatic video analysis**
    - ➔ **Archiving for successive search**

The characteristics of Digital video

Types of data managed

- **A digital video is composed of a sequence of frames plus possibly an audio track.**
- **In general, it is possible to view an audio/video document from different perspectives**
  - The audio part can be separated into
    - ➔ **Speech**
    - ➔ **Sound**
  - Sequence of frames (video shot and sequence)
  - Single frames as images
- **From all of them is possible to extract information that can be used for indexing and retrieval purposes**

Digital video characteristics

- **Sequence of frames with a certain frame rate**
  - NTSC 30 frames/sec, PAL 25 f/s, HDTV 60 f/s
  - Minimal change between frames
- **Single frames resolution**
  - 768 x 576 PAL, 720 x 480 NTSC
- **Uncompressed video requires high storage space and bandwidth**
  - For example, one second of uncompressed PAL video requires
    768 x 576 x 16 x25 ~ 172 MByte

Digital video storage and transmission [1/3]

- **The high storage requirements of video imposes the adoption of compression techniques.**
- **High compression rates are possible with video signals, due to the following reasons:**
  - Spatial correlation: correlation among neighboring pixels
  - Temporal correlation: correlation among pixels in different frames
  - A significant part of video data is not perceived

Digital video storage and transmission [2/3]

- **Compression can be divided in two broad categories**
  - **Lossless compression**, that allows one to compress decompress video without any degradation
    - ➔ **Lossless compression provides low compression factors**
    - ➔ **An example of lossless compression is MJEPG, where each frame is compressed using the JPEG format**
    - ➔ **Examples of lossless coding techniques are run-length coding, Huffman coding**

Digital video storage and transmission [3/3]

- **Lossy compression**, where the complete cycle of compression and decompression introduces some degradation of the original video
  - ➔ **Lossy compression allows to obtain high compression factors**
  - ➔ **Examples are the MPEG compression family (MPEG1, MPEG2)**
  - ➔ **Example of lossy coding is DPCM**
    - **DPCM compares adjacent pixels and stores only their difference**

MPEG

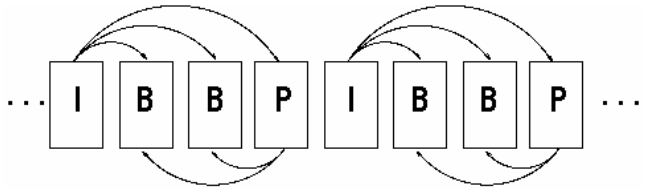
- **MPEG (Moving Pictures Experts Groups)**
  - MPEG1 has a bit-rate up to 1.5Mb/sec
    - ➔ **Designed for storage and retrieval of VHS quality video on CD-ROM**
  - MPEG2 Designed for broadcast video quality
    - ➔ **Bit rate: 2Mbps or higher**
    - ➔ **Used for DVD, cable TV, etc.**
  - MPEG4 is object-based, multi stream
    - ➔ **Variable bit-rates, from <64 kbps, up to 4Mbps and more (in the future)**

MPEG-1 [1/2]

- **Compression based on intra-frame and inter-frame encoding**
- **Intra-frame coding**
  - Each frame is subject to compression
  - Uses DCT compression schema
- **Inter-frame coding**
  - Exploits temporal redundancy
  - Predictive coding
    - **➔current picture is modeled as a transformation of picture at some previous time**
  - Interpolative coding
    - **➔Uses past and future pictures for reference**

MPEG-1 [2/2]

- **MPEG uses three types of frame coding**
  - I frames: intra-frame coding
    - **➔Moderate compression**
    - **➔Access points for random access**
  - P frames: predictive-coded frames
    - **➔Coded with reference to I or P frames**
  - B frames: bi-directionally predictive coded
    - **➔Coded using previous/next I and P frames**
    - **➔High compression**



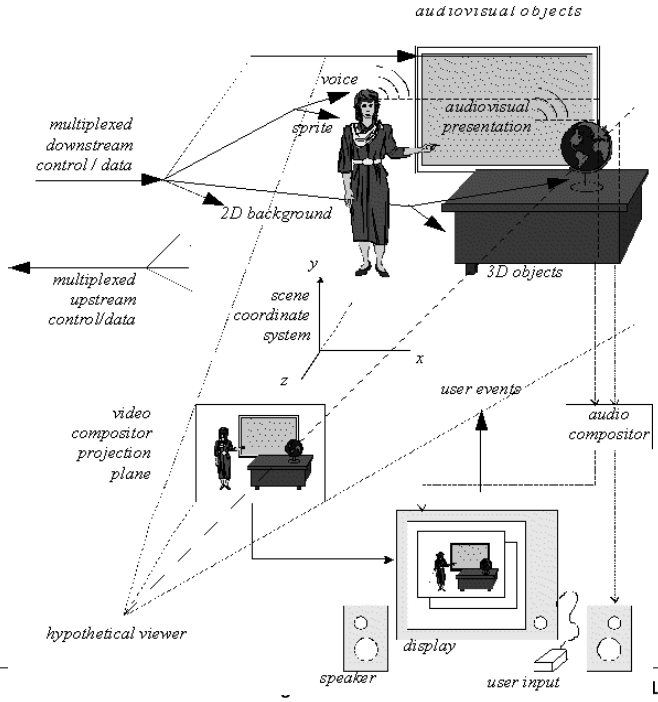
MPEG-4 [1/4]

- **Scalability of bit rate vs quality**
- **Better Audio/Video compression than MPEG-1**
- **Content based coding**
- **Support for efficient streaming**

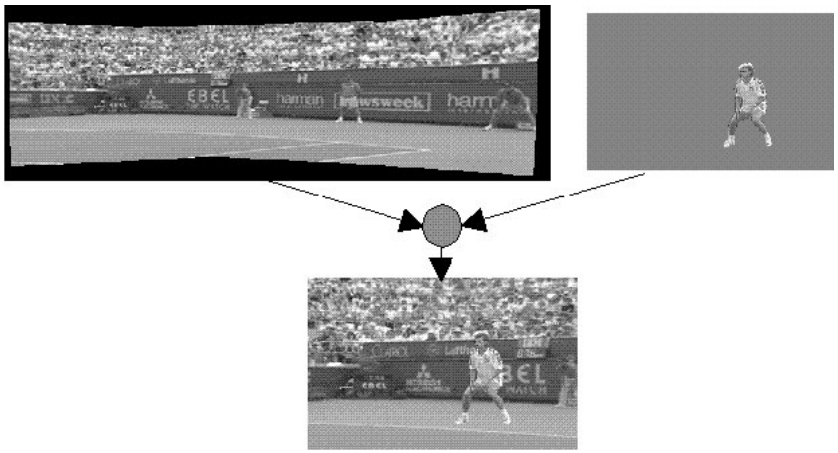
MPEG-4 [2/4]

- **Content based coding**
  - Reusability of object coding
  - Adaptation (different coding for different objects)
  - High quality for interesting parts
  - Possibility of scene composition
    - **➔Integration of natural and synthetic content**
    - **➔Tele-presence**

MPEG-4 [3/4]



MPEG-4 [4/4]



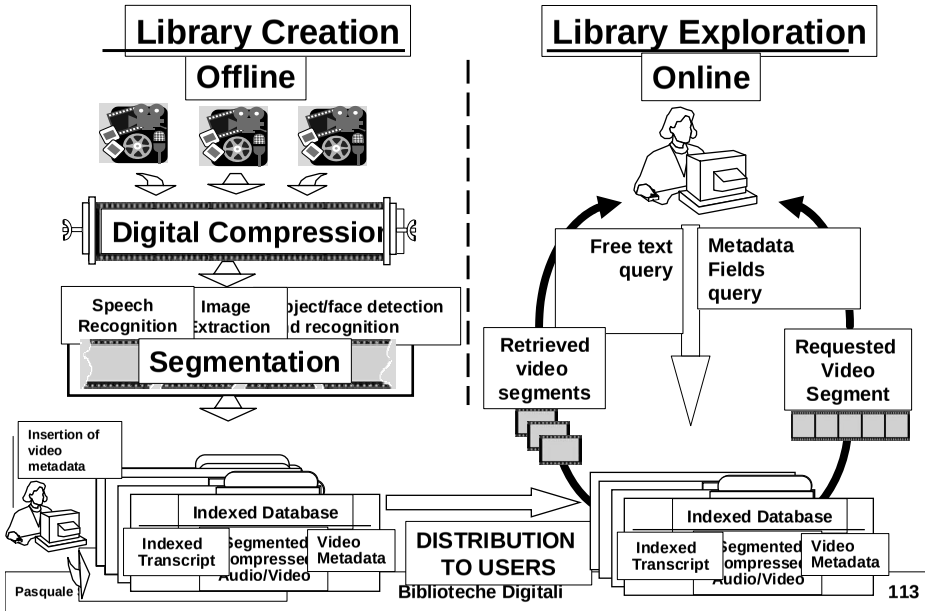
Digital Video representation

- **Video is composed of a sequence of** frames
- **Video is separated into** shots
  - A shot is a sequence of frames separated by a transition
  - Transitions between shots are given by
    - ➔ **Camera break**
    - ➔ **Dissolve**
    - ➔ **Wipe**
    - ➔ **Fade-in, fade-out**
- **A video can be separated into** sequences**, that are semantically meaningful groups of shots, possibly non consecutive**

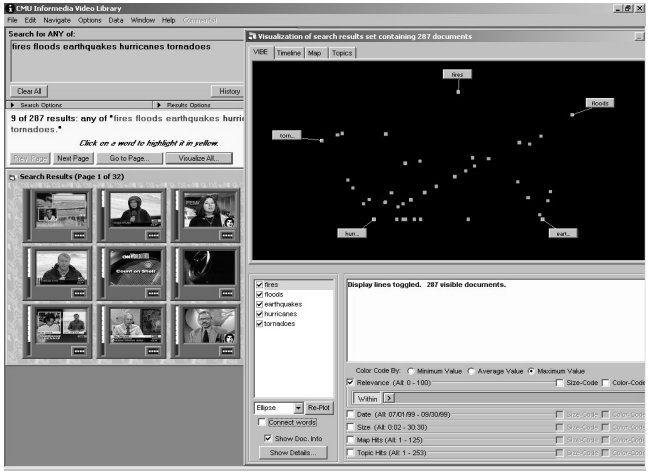
Operations of an A/V Digital Library

- **Video archiving and indexing**
- **Video storage**
- **Content-based search**
- **Video access (visualization and copy)**

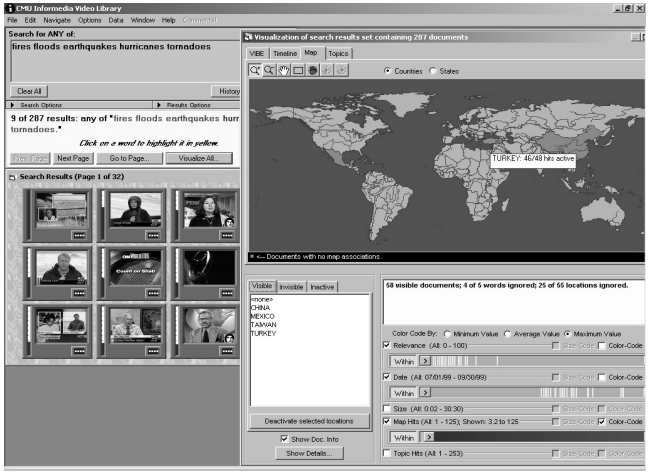
Summary of all phases & operations



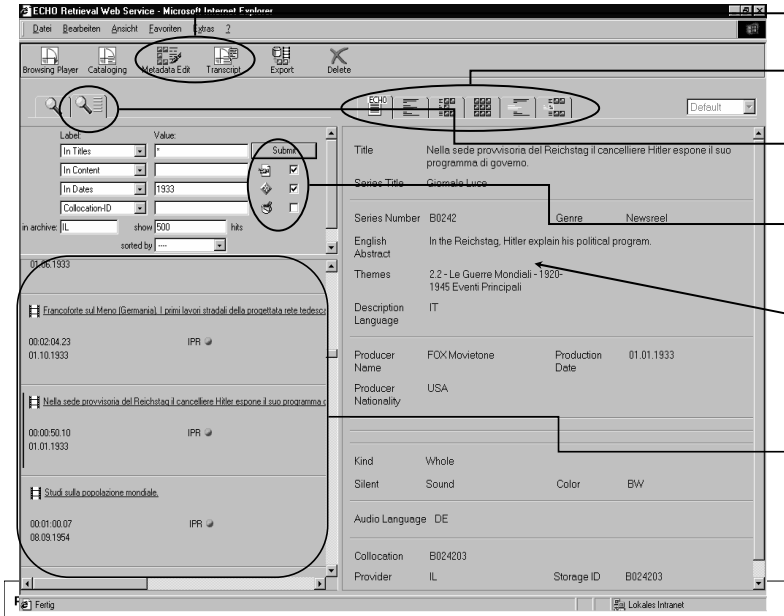
Informedia – an example

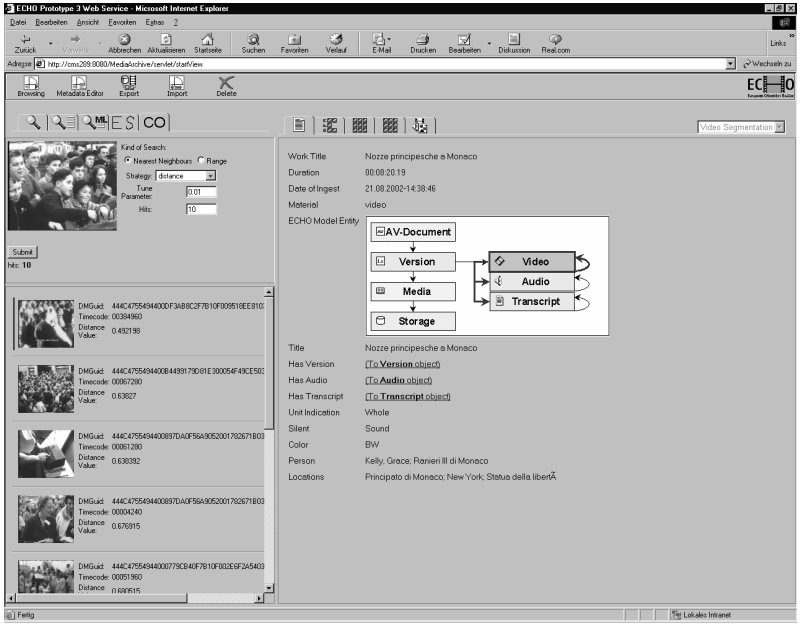


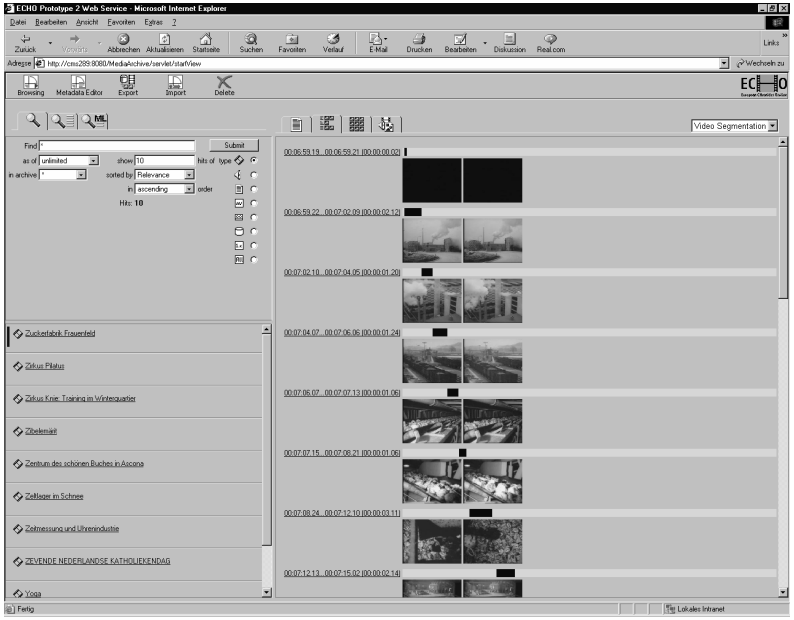
Informedia – an example

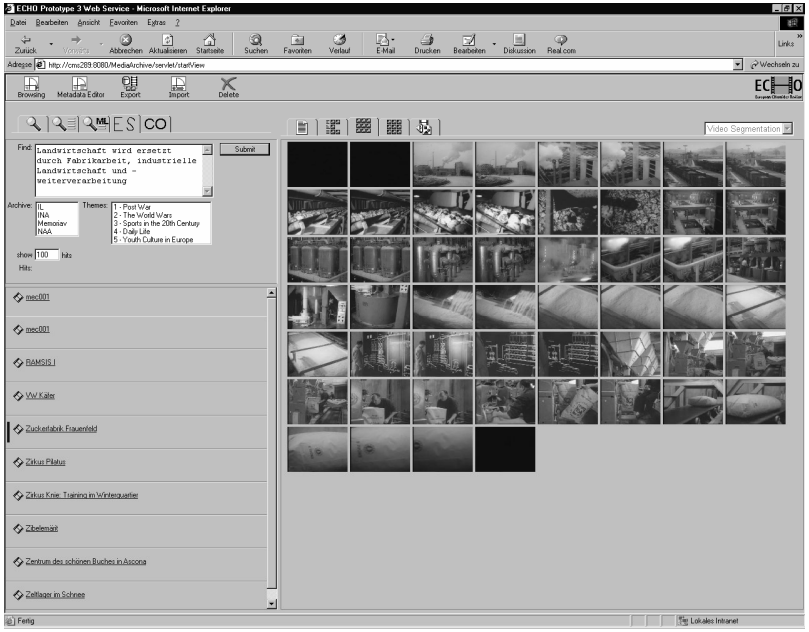


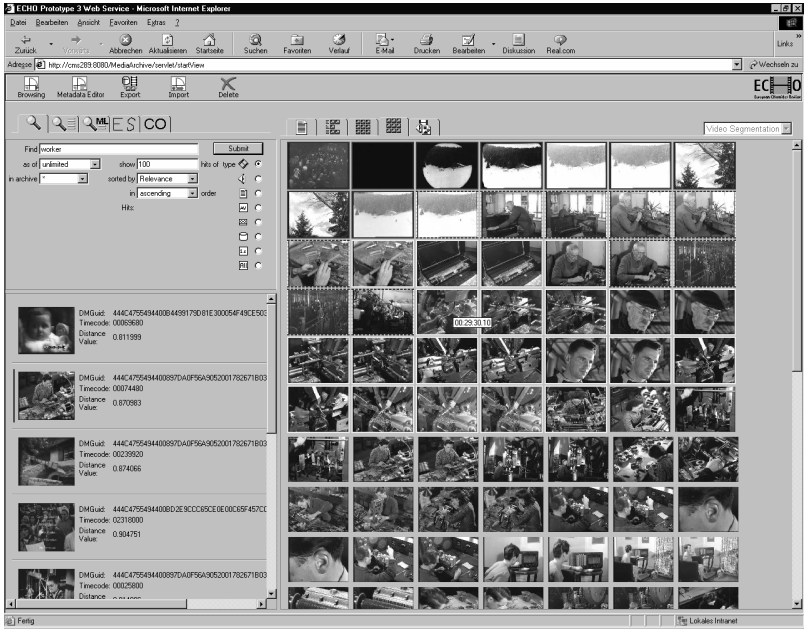
ECHO Retrieval Interface




ECHO Prototype 3 Web Service - Microsoft Internet Explorer
Work Title Nozze principesche a Monaco
Duration 00:08:20:19
Date of Ingest 21.08.2002-14:38:46
Material video
ECHO Model Entity
AV-Document
Version
Media
Storage
Video
Audio
Transcript
Title Nozze principesche a Monaco
Has Version (To Version object)
Has Audio (To Audio object)
Has Transcript (To Transcript object)
Unit Indication Whole
Silent Sound
Color BW
Person Kelly, Grace; Ranieri III di Monaco
Locations Principato di Monaco; New York; Statua della libertÃ


ECHO Prototype 2 Web Service - Microsoft Internet Explorer
Video Segmentation
Zuckerfabrik Frauenfeld
Zirkus Pilatus
Zirkus Knie: Training im Winterquartier
Zibelemärit
Zentrum des schönen Buches in Ascona
Zeltlager im Schnee
Zeitmessung und Uhrenindustrie
ZEVENDE NEDERLANDSE KATHOLIEKENDAG
Yoga


ECHO Prototype 3 Web Service - Microsoft Internet Explorer
Find: Landwirtschaft wird ersetzt durch Fabrikarbeit, industrielle Landwirtschaft und weiterverarbeitung
Video Segmentation
mec001
mec001
RAMSIS I
VW Käfer
Zuckerfabrik Frauenfeld
Zirkus Pilatus
Zirkus Knie: Training im Winterquartier
Zibelemärit
Zentrum des schönen Buches in Ascona
Zeltlager im Schnee


ECHO Prototype 3 Web Service - Microsoft Internet Explorer
Find: worker
Video Segmentation

InformaticaUmanistica

Sommario della prima parte

UNIVERSITÀ DI PISA

Sommario

- **Cenni storici**
  - Vannevar Bush
  - Dalle Biblioteche ai Cataloghi Automatizzati
  - Gli OPAC accessibili via Web
  - Le Biblioteche Digitali
- **Cos'è una Biblioteca Digitale**
  - Definizione
  - Confronto tra BD e database, sistemi IR, WWW, biblioteca tradizionale
  - Vantaggi delle BD
  - Alcuni esempi di Biblioteche Digitali

Sommario (cont.)

- **Cosa ha permesso la nascita delle Biblioteche Digitali**
  - Evoluzioni tecnologiche
- **Tipologie di Biblioteche Digitali**
  - Biblioteche Pubbliche e Biblioteche Specializzate
  - Tipi di documenti trattati
    - → **Libri**
    - → **Documenti testuali**
    - → **Immagini**
    - → **Audio/video**
    - → **……..**