

1

Informazione, informatica e risorse culturali digitali sul web: guida all'uso dei linguaggi ipertestuali

UniMc - BBCC 10/11 – I semestre

Laboratorio di informatica
avanzata per I beni culturali

Pierluigi Feliciati
pierluigi.feliciati@unimc.it

cos'è l'informazione?

L'informazione è oggi oggetto di grande interesse: si parla comunemente, per esempio, di **società dell'informazione** ...

... ma quello di informazione è un **concetto diversificato e complesso**: qual'è l'unità minima di informazione? Nel campo dei BBCC, cosa possiamo definire informazione?

proviamo a capirlo meglio attraverso un percorso storico

cos'è l'informazione?

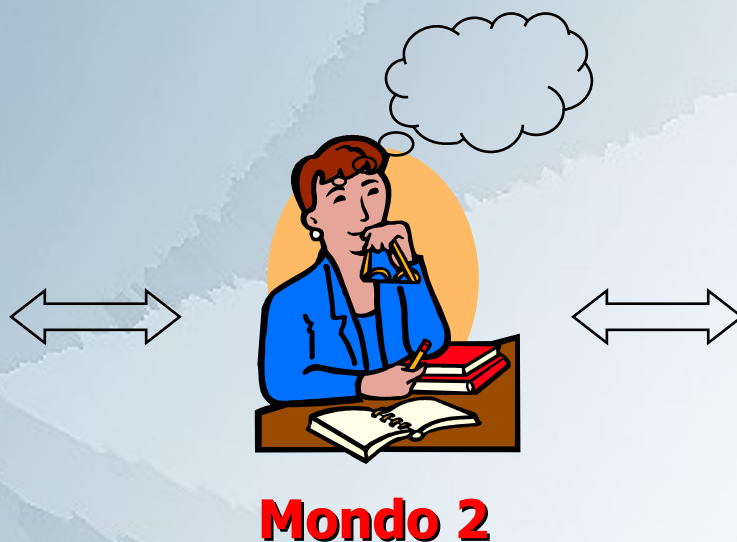
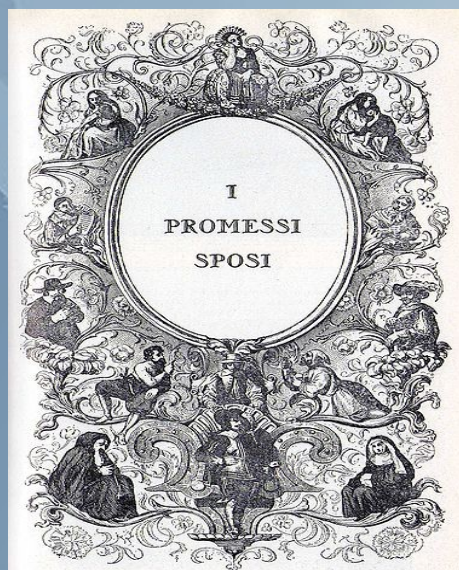
la transizione del business

L'evoluzione economico-industriale della società occidentale può essere interpretata secondo tre fasi:

1. Dapprima l'attenzione è stata rivolta alla produzione di beni materiali, nella società pre-industriale: **gestione della materia**
2. Producendo beni materiali, ci si è anche resi conto dell'importanza di produrre in modo efficiente, nella società industriale: **gestione dell'energia**
3. Impiegando energia, ci si è resi conto dell'importanza di controllare adeguatamente la produzione, nella società post-industriale: **gestione dell'informazione**

informazione e mondi

informazione fisica, conoscenza e opera



*A. Manzoni,
I Promessi Sposi, 1827*

Mondo 3

con l'introduzione dei **computer**, si è arrivati a progettare un **dispositivo programmabile** per la **gestione dell'informazione**

informazione e informatica

l'informatica è

**lo studio sistematico degli algoritmi che
descrivono e trasformano l'informazione
o più semplicemente**

**la scienza della rappresentazione
e dell'elaborazione dell'informazione**

il concetto di informazione

l'informazione è sempre portata da o
trasmessa su o memorizzata in o contenuta in
qualcosa che

NON E' INFORMAZIONE,

quindi:

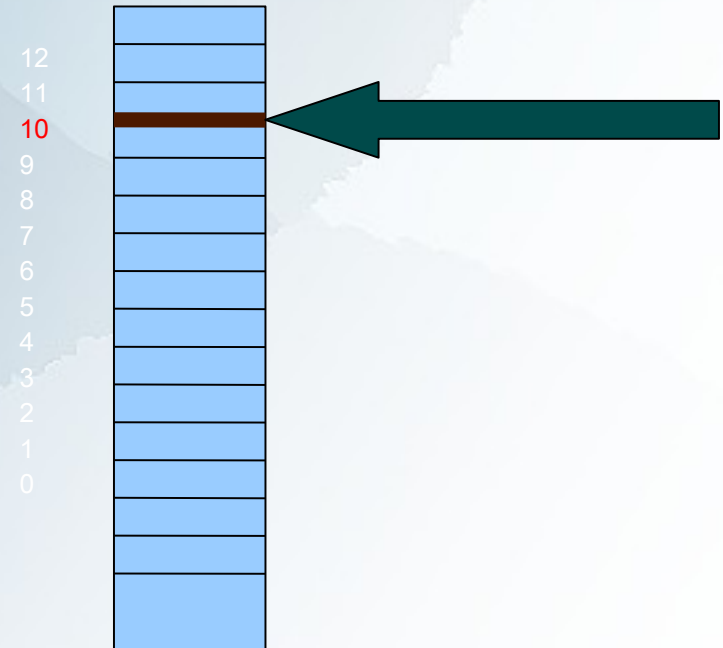
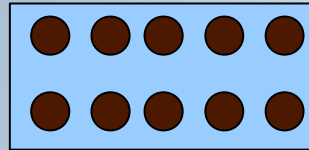
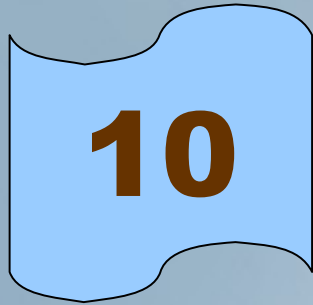
*l'informazione è disponibile solo se è
accessibile il supporto (quel qualcosa)
che la mantiene*

il concetto di informazione

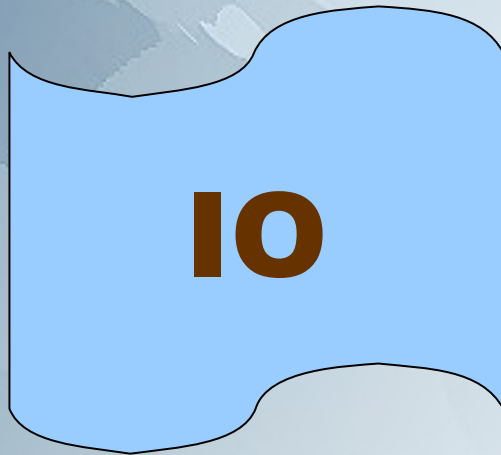
**ogni supporto ha le proprie caratteristiche:
alcuni sono migliori per la trasmissione di
informazione (es. l'aria o un cavo in fibra
ottica), altri per la sua memorizzazione (es.
un CD o una usb key)**

la stessa informazione può essere scritta su
supporti differenti e uno stesso supporto può
portare informazioni differenti

l'informazione e i suoi supporti



**stessa informazione su
diversi supporti....**



**e stesso supporto con
informazioni differenti:
io o dieci?**

informazione e linguaggio: codifica e decodifica

la relazione tra informazione e suoi supporti è complessa

l'insieme delle regole di tale relazione è una delle basi di un **linguaggio**

- ♦ **la codifica** è l'operazione di fissazione dell'informazione su un supporto
- ♦ **la decodifica** è la lettura di informazione da un supporto

informazione, messaggio e risorse digitali

i singoli atomi di informazione (**simboli**), combinati secondo una configurazione ordinata basata su una codifica/decodifica comprensibile, costituiscono **un messaggio**.

Ovviamente, più è ampio l'insieme di simboli usato, maggiore l'informazione di quel simbolo:
se si prova a **ridurre l'unità elementare a due stati**
(sì o no, 1 o 0, acceso o spento...)
si ottiene il **massimo di economia di simboli e di configurazioni possibili,**
semplificando al massimo la comprensibilità

informazione, messaggio e risorse digitali

la **codifica binaria** rappresenta allora l'alfabeto più semplice possibile per la codifica di dati e istruzioni. Per questo i calcolatori attuali si basano su questa codifica, **basata su bit (da binary digit)** combinati in successioni ben formate di **2^n possibilità** dove **n** rappresenta

il **numero di bit che compone la successione.**

Esempio: **$2^3 = 8$ successioni con 3 bit**

000, 001, 010, 100, 100, 101, 110, 111

A una **successione di $2^8 (=256)$** si dà il nome di **byte** da cui i multipli **kilobyte (1024 byte)**, **megabyte (1048576 bytes)** etc

dai bit alle risorse digitali

attraverso codifiche standardizzate, cioè **tabelle di corrispondenza tra successioni e valori**, si **associano alle successioni di bit** **valori** numerici, caratteri alfanumerici, stati del sistema, valori fisici (suoni, immagini..) etc. Ecco perché si parla di **informazione digitale** e di **convergenza al digitale**.

La tendenza attuale è cioè a campionare in digitale tutti i messaggi (**testi, numeri, suono, video, azioni meccaniche, onde radio, etc.**) nello stesso codice digitale.

dai bit alle risorse digitali

La convergenza al digitale implica la necessità di dispositivi **di decodifica** (computer).

Le successioni informative binarie, sia che siano disponibili localmente (cioè nella nostra macchina) oppure in rete, sono dette **risorse digitali**:

risorse per la funzione informativa, di fonte, di contenuto culturale che rivestono;

digitali per il loro formato di codifica (e dunque per la conseguente necessità di disporre di dispositivi di decodifica per accedervi)

informazione e informatica

sempre più l'informatica viene allargata
alle telecomunicazioni

ovvero alla

**trasmissione rapida a distanza
dell'informazione**

IT Information Technology

ICT Information & Communication
Technology

Identificazione e marcatura delle risorse digitali

una volta che prendiamo come oggetto del nostro lavoro e della nostra ricerca **le risorse digitali**, una delle questioni principali è l'identificazione dell'**unità minima**:

in che modo distinguere la **singola** risorsa digitale dagli **insiemi** di risorse digitali?

di certo, è raro che si possano trovare *risorse isolate*, con qualche eccezione per le risorse ad accesso locale (ma anche qui forse potremmo discutere l'unità di misura) e per le *brutte* risorse web.

Identificazione e marcatura delle risorse digitali

le risorse culturali digitali, specie quelle remote, **acquistano gran parte del loro valore informativo e della loro efficacia comunicativa, vale a dire una buona qualità**, solo se sono:

- **univocamente identificabili**
- **correttamente contestualizzate**

Identificazione e marcatura delle risorse digitali

per **identificare** le risorse digitali è necessario **pre-describerle**, quindi **inserire o associare informazioni di riferimento**, corrispondenti per capirci al frontespizio, al titolo, all'autore, alle note sull'edizione, alla registrazione di protocollo...

per **descrivere** (“catalogare” o “inventariare”) le risorse digitali ci si dovrà basare prima di tutto su queste **informazioni/fonte**, poi sull'analisi dei contenuti della risorsa.

Identificazione e marcatura delle risorse digitali

queste informazioni/fonte sono definite appunto

metadati = dati sui dati

qui usiamo questo termine con una sfumatura diversa rispetto a quella che si usa per i *database*, per cui si definiscono **metadati** le **strutture e le regole per i dati**

Dati e metadati

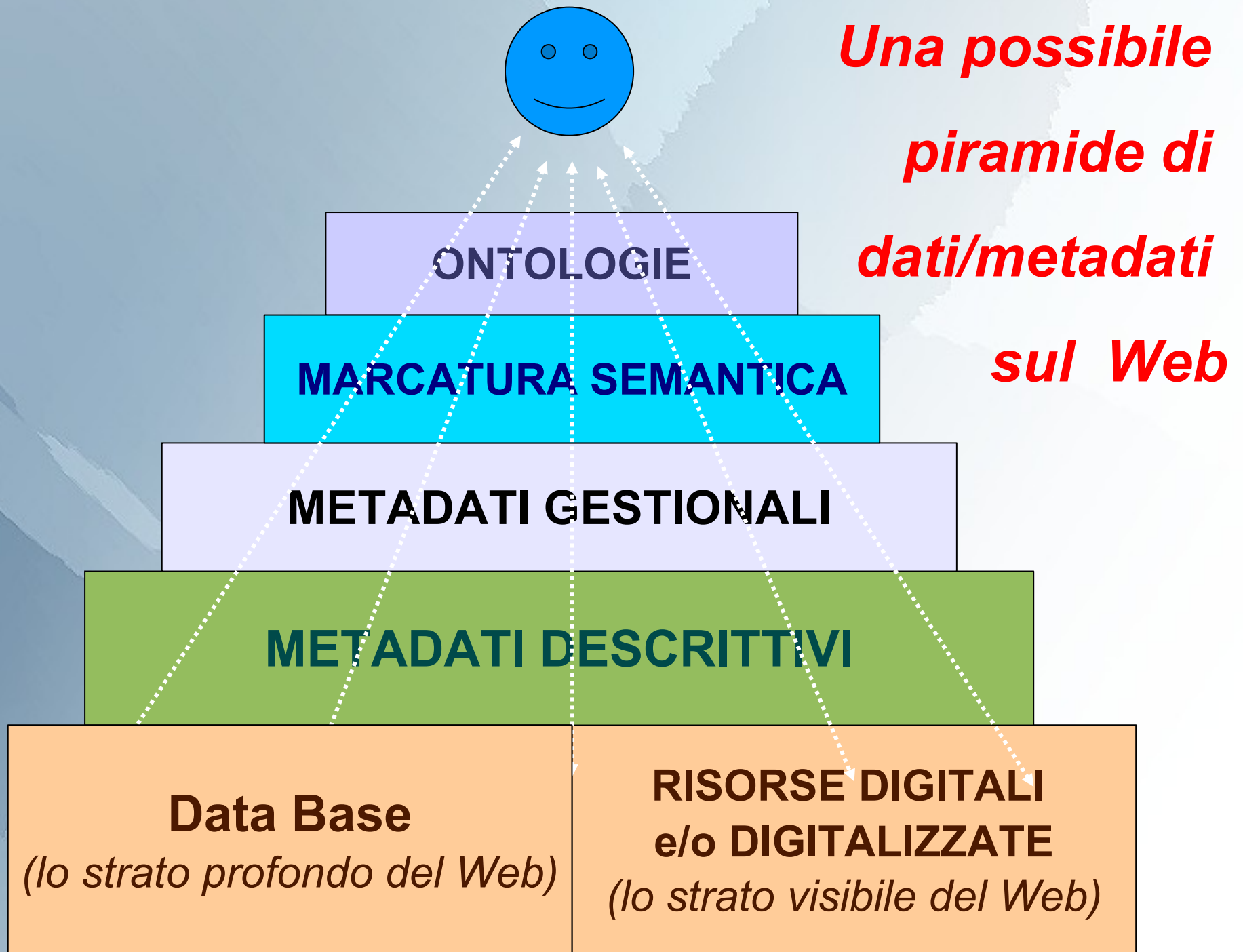
Da quello che si è accennato fin qui, risulta evidente che ogni oggetto/risorsa digitale deve “portarsi dietro” alcuni dati che **non fanno parte del suo contenuto informativo**, ma che sono indispensabili a renderlo rappresentabile tramite specifiche procedure di decodifica.

Le informazioni di base (*nome, formato/estensione di codifica*) **non sono sufficienti** a GESTIRE questi oggetti con efficienza, a organizzarli, a ritrovarli in rete, a conservarli a lungo termine.

Dati e metadati

I libri hanno un frontespizio e un codice ISBN, le opere d'arte in un museo hanno etichette e pannelli esplicativi, oppure schede dettagliate nel catalogo, le medicine in una farmacia hanno informazioni sul principio attivo e codici che distinguono quelle per cui necessita una ricetta da quelle da banco, i documenti archivistici addirittura nascono con associate le informazioni sui vincoli di contesto.

Allo stesso modo gli **oggetti digitali** necessitano di informazioni associate. Queste informazioni (**dati sui dati**) si chiamano appunto **metadati**.



Identificazione e marcatura delle risorse digitali

per garantire le funzioni basilari (**trovare, identificare, selezionare, ottenere**) i *metadati* devono rispondere a **requisiti strutturali e formali condivisi**. L'**adozione di standard** in questo senso diventa infatti particolarmente importante:

- per facilitare lo scambio tra soggetti diversi
- perché in ambiente digitale parte dell'elaborazione delle informazioni è svolta da *software*, che non sono sempre abbastanza furbi da sciogliere le differenze sintattiche e logiche...

Identificazione e marcatura delle risorse digitali

Lo standard più usato per esprimere informazioni (metadati descrittivi) per le risorse digitali è il:

Dublin Core Metadata Element Set

sviluppato in ambito USA - OCLC (Online Computer Library Center) che ha stabilito un vocabolario semantico per descrivere informazioni sulle caratteristiche *core* (**principali**) di un oggetto digitale e categorizzarlo ai fini di una ricerca semplificata da parte dell'utente (*umano o automatico* che sia).

Identificazione e marcatura delle risorse digitali

Il set minimo proposto nel dicembre 1996 è costituito da **15 elementi di base** e si è esteso anche a sottoelementi o qualificatori

Lo standard è in via di sviluppo ma il cosiddetto *core* dei 15 elementi della descrizione è rimasto stabile.

Ciascun elemento è definito usando un set di 10 attributi ricavati dalla norma ISO 11179 (*Specification and standardization of data elements*).

Identificazione e marcatura delle risorse digitali

Dei **dieci attributi**, **sei sono comuni** a tutti gli elementi DC:

- **Versione** 1.1
- **Registrazione di autorità** Dublin Core Metadata Initiative
- **Lingua** en
- **Obbligatorietà** opzionale
- **Tipo di dato** stringa di caratteri
- **Occorrenza massima** illimitata

gli **altri quattro** – essenziali per la coerenza semantico-sintattica sono:

- **Nome**
- **Identificatore**
- **Definizione**
- **Commento**

Identificazione e marcatura delle risorse digitali

gli elementi di Dublin Core (1-8)

<u>nome</u>	<u>identificatore</u>	<u>definizione</u>
Titolo	DC.Title	Un nome dato alla risorsa
Creatore	DC.Creator	Un'entità che ha la responsabilità principale della produzione del contenuto della risorsa
Soggetto	DC.Subject	L'argomento della risorsa.
Descrizione	DC.Description	Una spiegazione del contenuto della risorsa
Editore	DC.Publisher	Un'entità responsabile della produzione della risorsa, disponibile nella sua forma presente
Autore di contr. subordinato	DC.Contributor	Un'entità responsabile della produzione di un contributo al contenuto della risorsa
Data	DC.Date	Una data associata a un evento del ciclo di vita della risorsa, ad es. nella forma YYYY-MM-DD
Tipo	DC.Type	La natura o il genere del contenuto della risorsa, ad es. la lista dei Dublin Core Types

Identificazione e marcatura delle risorse digitali

gli elementi di Dublin Core (9-15)

<u>nome</u>	<u>identificatore</u>	<u>definizione</u>
Formato	DC.Format	La manifestazione fisica o digitale della risorsa (cfr. Internet Media Types [MIME])
Identificatore	DC.Identifier	Un riferimento univoco alla risorsa nell'ambito di un dato contesto (URL, DOI, ISBN...)
Fonte	DC.Source	Un riferimento a una risorsa dalla quale è derivata la risorsa in oggetto
Lingua	DC.Language	La lingua del contenuto intellettuale della risorsa (codice di 2 chr ISO 639 o 2+2 ISO 3166 x paese)
Relazione	DC.Relation	Un riferimento alla risorsa correlata
Copertura	DC.Coverage	L'estensione o scopo del contenuto della risorsa (sua estensione spaziale, temporale o giurisdiz.)
Gestione dei diritti	DC.Rights	Informazione sui diritti esercitati sulla risorsa

I Metadati per strutturare, gestire e conservare

Metadati strutturali

collegano le varie parti di una stessa risorsa o le componenti di più risorse per un'adeguata e completa fruizione, che spesso avviene attraverso la mappatura di schemi di metadati diversi. Questi metadati inoltre forniscono dati di identificazione e localizzazione del documento, come il codice identificativo, l'indirizzo del file sul server, l'archivio digitale di appartenenza e il suo indirizzo Internet.

Ne sono esempio:

1. **SGML** e soprattutto **XML**
2. **MAG** e **METS** (che svolgono anche queste funzioni)

I Metadati per strutturare, gestire e conservare

Metadati amministrativi, gestionali e di conservazione

memorizzano le informazioni sulla creazione e la manutenzione degli oggetti nel sistema di gestione dell'archivio digitale e sono necessari per una corretta esecuzione delle relative attività.

- Data la labilità dell'informazione elettronica, questi tipi di metadati assumono un'importanza preponderante ai fini della conservazione permanente degli oggetti digitali: essi possono documentare i processi tecnici associati alla conservazione permanente, fornire informazioni sulle condizioni e i diritti di accesso agli oggetti digitali, certificare l'autenticità e l'integrità del contenuto, documentare la catena di custodia degli oggetti, identificarli in maniera univoca. Gli standard **METS**, **MPEG21**, **MAG** e **PREMIS** sono i più usati.

Contestualizzazione delle risorse digitali

le risorse sono infatti quasi sempre **associate in insiemi coerenti** per avere la garanzia di essere correttamente **contestualizzate** ed essere inserite e recuperate in una “rete di significati”

I sistemi di risorse possono, soprattutto:

- rispondere a una **logica strutturale definita** (database, gestiti da DBMS)
- **basarsi su collegamenti ipertestuali** (siti e applicazioni web).

Identificazione e marcatura delle risorse digitali

un **database** è una raccolta di dati permanenti, gestita da un elaboratore e suddivisa in

- **Metadati o schema:** definizioni che ci dicono quali regole valgono sui dati, quali valori possono essere validi (vincoli di integrità), come i dati sono strutturati e collegati tra loro. Lo schema può cambiare nel tempo, non ha alcun vincolo con i programmi che accedono al database e deve essere definito prima dei dati.
- **Dati:** rappresentazione (astrazione) dei fatti, ma può anche includere previsioni, supposizioni ed ipotesi conformi ai Metadati (o schema).

I dati sono organizzati in **insiemi omogenei** (entità con stessa struttura) e sono collegabili con altri insiemi.

Il **modello dei dati** è un insieme di astrazioni che descrivono le associazioni tra i dati stessi. Si hanno principalmente i modelli di dati **relazionale o ad oggetti**.

Le risorse digitali aggregate

Tipi di data base

Le basi di dati possono avere **varie strutture**, tipicamente, in ordine cronologico:

1. **gerarchica** (rappresentabile tramite un albero - anni sessanta),
2. **reticolare** (rappresentabile tramite un grafo - anni sessanta),
3. **relazionale** (attualmente **il più diffuso**, rappresentabile mediante tabelle e relazioni tra esse - anni settanta),
4. **ad oggetti** (estensione alle basi di dati del paradigma "Object Oriented", tipico della programmazione a oggetti - anni ottanta),
5. **semantica** (rappresentabile con un grafo relazionale - inizio anni duemila).

Un requisito importante di una buona base dati consiste nel **non duplicare inutilmente le informazioni in essa contenute**: questo è reso possibile dai gestori di database relazionali, che consentono di salvare i dati in tabelle che possono essere collegate.

Le risorse digitali aggregate

data base relazionale

L'assunto fondamentale del modello relazionale è che **tutti i dati siano rappresentati come relazioni**; i dati sono manipolati con gli operatori dell'algebra relazionale. Il modello relazionale consente al progettista di database di creare una rappresentazione **consistente e logica** dell'informazione.

La **consistenza** viene ottenuta inserendo nel progetto del database appropriati **vincoli**, normalmente chiamati **schema logico**.

La **struttura base** del modello relazionale è il **dominio o tipo di dato**, definito come l'insieme dei valori che può assumere un determinato attributo.

Un **attributo** è una coppia ordinata di "nome di attributo" e "nome di tipo", mentre un valore di attributo è un valore specifico valido per quel tipo di dato. Una **relazione** consiste di una testata e di un corpo, dove la testata è un insieme di attributi e il corpo è un insieme di n valori possibili. La **tabella** è la rappresentazione grafica normalmente accettata per rappresentare la relazione.

le risorse digitali aggregate per il WWW

Ipertesto

Agostino Ramelli, ingegnere italiano nato nel 1531, ideò la **ruota dei libri**, leggio multiplo rotante, ideato per consentire l'agevole lettura contemporanea di più testi e che si può considerare una prima forma di ipertesto.

C'è chi fa risalire il concetto di ipertesto a **Vannevar Bush** che nel 1945 scrisse un articolo intitolato *As We May Think* nel quale descrive un sistema di informazione interconnesso chiamato **Memex**, mai realizzato ma considerato il nonno dei PC.

<http://www.theatlantic.com/doc/194507/bush>.

le risorse digitali e il WWW

Iper testo - storia

I due americani universalmente riconosciuti come gli inventori dell'ipertesto sono **Ted Nelson** e **Douglas Engelbart**: il primo inventò il termine **hypertext** nel 1965, con un significato più ampio, coinvolgendo qualsiasi sistema di scrittura non lineare che utilizza l'informatica.

Bill Atkinson realizzò nel 1980 alla Apple *HyperCard*, un'applicazione software che gestiva in maniera semplice grandi quantità di informazioni sotto forma di testo o di immagini, dotato di un avanzato linguaggio di programmazione ipertestuale, *HyperTalk*.

(Ted Nelson, *Dream machines*, 1974)

Con “ipertesto” intendo...

... una scrittura non sequenziale.

La scrittura normale è sequenziale per due ragioni: è nata dal linguaggio e dall'oratoria, che devono essere sequenziali e i libri sono comodi da leggere solo in sequenza.

Ma **le strutture delle idee non sono sequenziali**. Hanno legami in ogni direzione, E quando scriviamo, cerchiamo sempre di creare legami non sequenziali. (...) Molti scrittori hanno tentato di liberarsi dalla sequenza. Mi vengono in mente *Fuoco pallido* di Nabokov, il *Tristram Shandy* di Sterne e *Il gioco del mondo* di Julio Cortàzar.

Io penso che **stia sorgendo una nuova era**. La memoria del computer e lo schermo significano che non siamo più obbligati ad avere le cose in sequenza; sono possibili strutture del tutto arbitrarie e credo che una volta che le avremo provate abbastanza molti si accorgeranno di quanto siano auspicabili.

le risorse digitali e il WWW

Ipertesto e ipermedia

L'ipertesto è **un insieme di testi o pagine leggibili** con l'ausilio di un'interfaccia elettronica, in maniera **non sequenziale**, per tramite di particolari parole chiamate **collegamenti ipertestuali** (link o hyperlink), che costituiscono un **rete raggiata o variamente incrociata di informazioni**, organizzate secondo diversi criteri, ad esempio **paritetici o gerarchici**, in modo da permettere più percorsi di lettura.

Il sistema d'ipertesto più conosciuto e più ampio è certamente il **World wide web** di Internet, che utilizza il linguaggio HTML (HyperText Markup Language) per definire all'interno del testo istruzioni codificate per i collegamenti. Con l'inserimento di contenuti da vari media diversi (**multimedia**) in un ipertesto, si è cominciato a utilizzare il termine **ipermedia**.

le risorse digitali e il WWW

Le risorse in rete (Internet e WWW)

Gli oggetti e/o servizi disponibili in Internet includono:

- **file** (testi, immagini, archivi numerici strutturati, audio o video digitali, programmi per computer)
- **servizi interattivi** (sessioni interattive di telnet, servizi di conferenza elettronica, streaming audio e video)
- **aggregazioni di informazione** (basi di dati, archivi elettronici accessibili via FTP, archivi di messaggi di conferenze elettroniche)