

1

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

UniMc – BBCC 12/13 – II semestre

Laboratorio di informatica
avanzata per i beni culturali

Pierluigi Feliciati
pierluigi.feliciati@unimc.it

cos'è l'informazione?

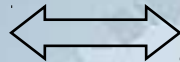
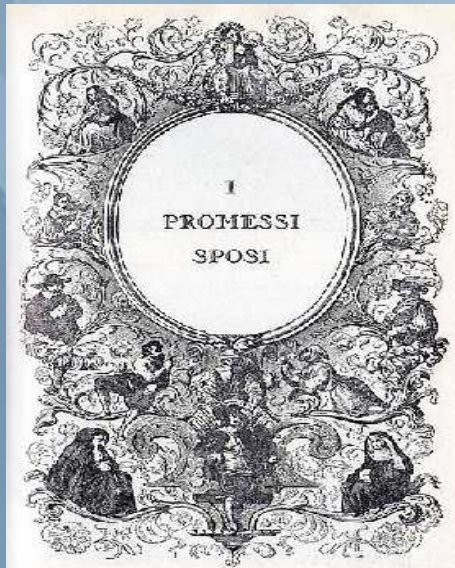
L'informazione è oggi oggetto di grande interesse: si parla comunemente, per esempio, di **società dell'informazione** ...

... ma quello di informazione è un **concetto diversificato e complesso**:

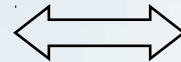
- qual'è l'unità minima di informazione?
- per i BBCC, cosa possiamo definire informazione?
- che rapporto ci può essere tra informazione, suoi autori e utenti?

informazione e mondi

informazione fisica, conoscenza e opera



Mondo 2



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*

informazione e informatica

sempre più l'informatica viene allargata
alle telecomunicazioni

ovvero alla

**Trasmissione/ricezione a distanza
dell'informazione**

IT Information Technology diventa allora

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.

Dati e metadati

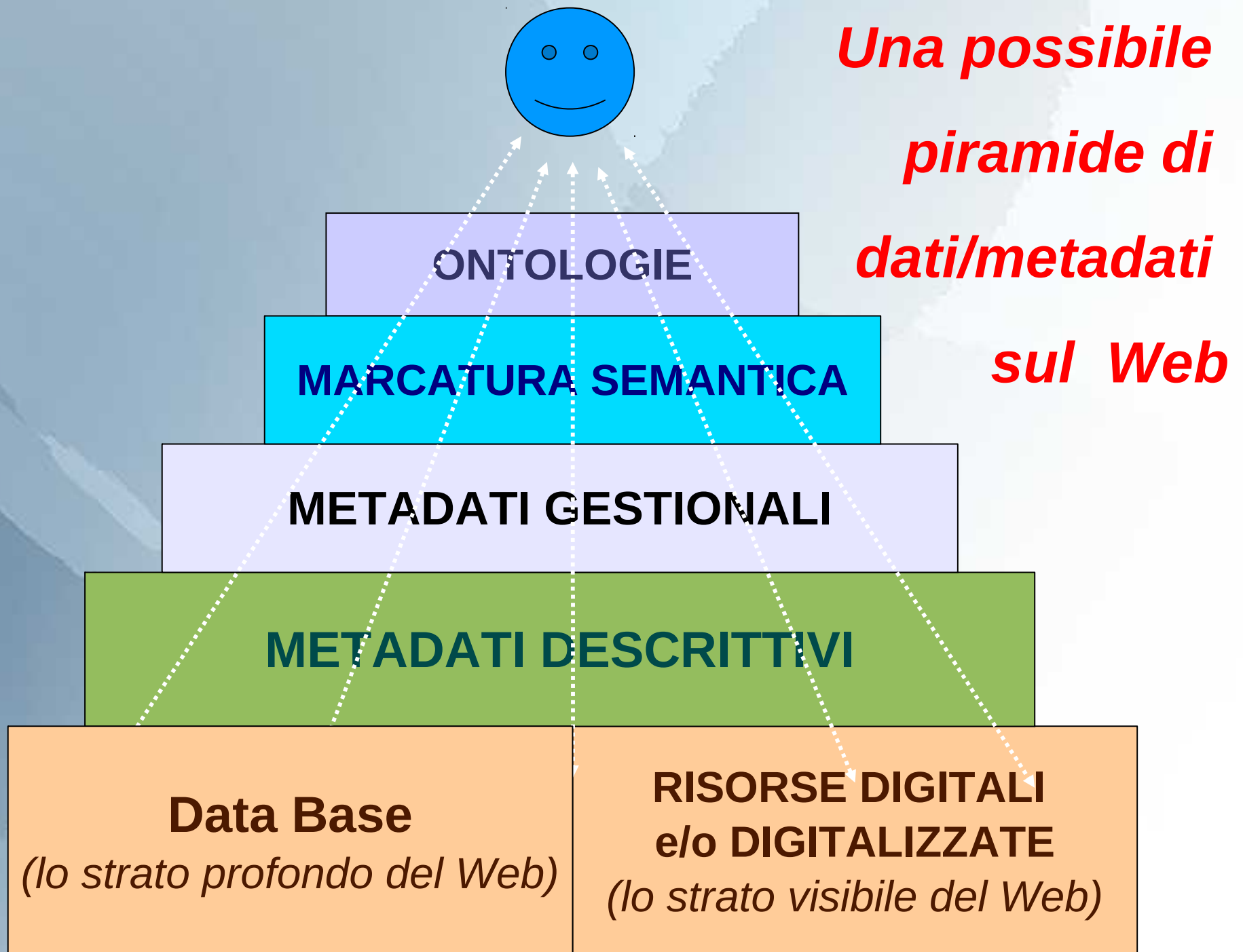
Da quello che si è accennato fin qui, risulta evidente che ogni oggetto/risorsa digitale deve “portarsi dietro” alcuni **dati che non fanno parte del suo contenuto informativo**, ma che sono indispensabili a renderlo rappresentabile tramite specifiche procedure di decodifica.

Le informazioni di base di un *file* (*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**.



Dati e metadati

Le funzioni dei metadati

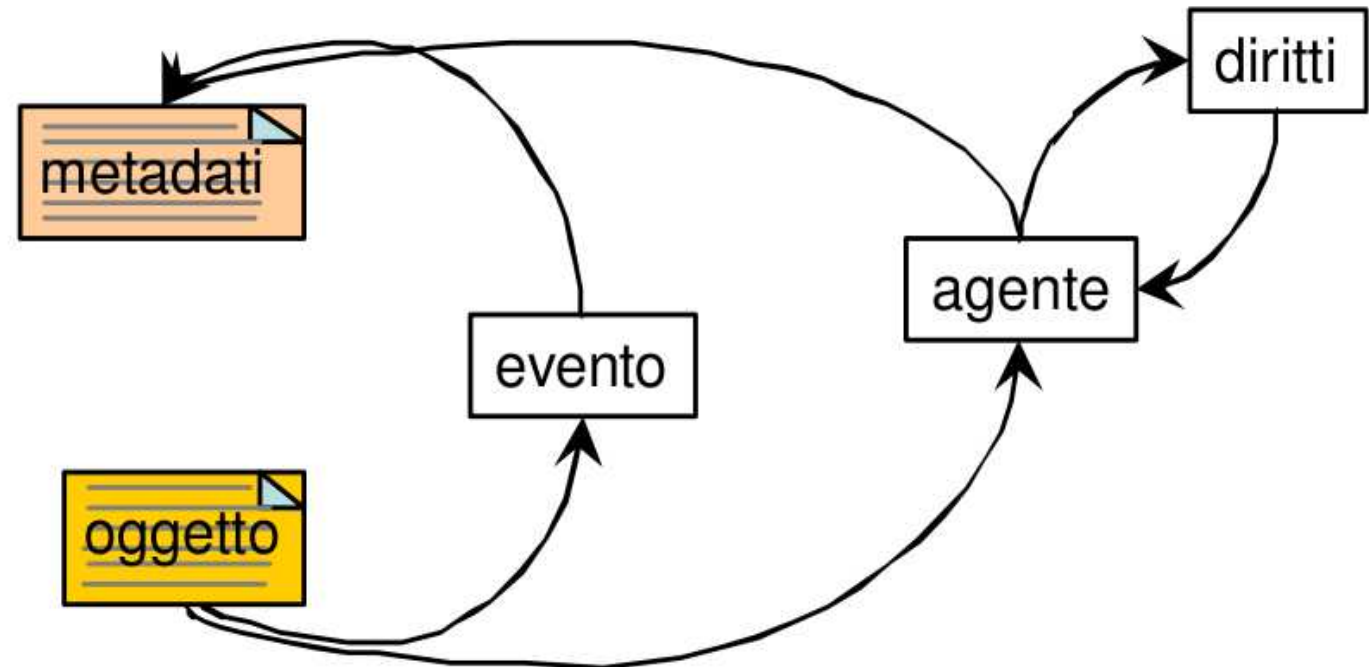
sono consentire il raggiungimento di alcuni obiettivi:

- **Ricerca**, individuare l'esistenza di un documento;
- **Localizzazione**, rintracciare una particolare occorrenza del documento;
- **Selezione**, analizzando, valutando e filtrando in una serie di documenti;
- **Gestione**, gestire le raccolte di documenti grazie all'intermediazione di banche dati, repository e cataloghi;
- **Disponibilità**, ottenere informazioni sull'effettiva disponibilità del documento
- **Conservazione**, gestire l'insieme complesso di attività che garantiscono nel tempo la fruizione

Uno schema del ciclo di vita di dati e metadati

Nel ciclo di vita di un oggetto digitale

Metadati = Biografia dell'oggetto digitale



Dati e metadati

Tipologie di metadati: 1 - Metadati descrittivi

servono per l'**identificazione ed il recupero** degli oggetti digitali; sono costituiti da **descrizioni** delle risorse, nate in formato digitale o digitalizzate; possono risiedere **all'esterno** dell'archivio digitale oppure essere **inseriti nelle risorse** o ancora **annidati in altri metadati**:

1. *Handle*, PURL (Persistent Uniform Resource Locator). DOI...
3. Dublin Core e MODS (metadati descrittivi)
4. I *collection description Schemas* (già citati)
5. I Meta Tag di HTML (descrittori all'interno della risorsa)
6. Encoded Archival Description (EAD) e EA Context (EAC)
7.

Identificazione e marcatura delle risorse digitali

per garantire le funzioni basilari (**trovare, identificare, selezionare, ottenere**) i *metadati descrittivi* 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 sono raramente così 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*).

Metadati descrittivi

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

Metadati descrittivi

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 strutturali

I metadati strutturali descrivono le **relazioni logiche o fisiche che collegano le parti di un oggetto composito**.

Un libro “fisico”, ad esempio, fa parte delle offerte con sconto, e a sua volta può consistere in una sequenza di capitoli, che sono composti da pagine...

Il processo di digitalizzazione/aggregazione di risorse può generare un certo numero di risorse digitali distinte, ad esempio una immagine per ogni pagina, ma il fatto che queste risorse formino una sequenza e che quella sequenza costituisca un oggetto composito o faccia parte di una struttura complessa multilivellare è evidentemente essenziale per il loro uso e la loro interpretazione.

I 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** (di cui parleremo)
2. **MAG** e **METS** (che tra le altre svolgono anche queste funzioni)

Metadati per gestire e conservare

i 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.

Dati e metadati: MAG

i Metadati Amministrativi-Gestionali (MAG)

MAG è un *profilo applicativo* (AP) espresso in XML che offre le specifiche formali per la fase di **raccolta e di trasferimento dei metadati e dei dati digitali** nei rispettivi archivi. Esso è indipendente da specifici hardware e software e comprende al suo interno alcuni elementi descrittivi.

I set MAG, più che per particolari tipologie di documenti fonte (ad es. periodici, musica a stampa, grafica, ecc.), viene definito per tipologie di oggetti digitali (es. immagini statiche, testi prodotti con tipologia OCR, suono, audiovisivi, ecc.).

Dati e metadati: MAG

Lo **schema dell'AP (metadigit.xsd)** è **composto da diverse sezioni**, alcune generali e altre specifiche del contenuto digitale:

<gen> informazioni generali sul progetto e sul tipo di digitalizzazione

<bib> metadati descrittivi sull'oggetto analogico

<stru> metadati strutturali

**** metadati specifici relativi alle immagini fisse

<ocr> metadati specifici relativi al riconoscimento ottico del testo

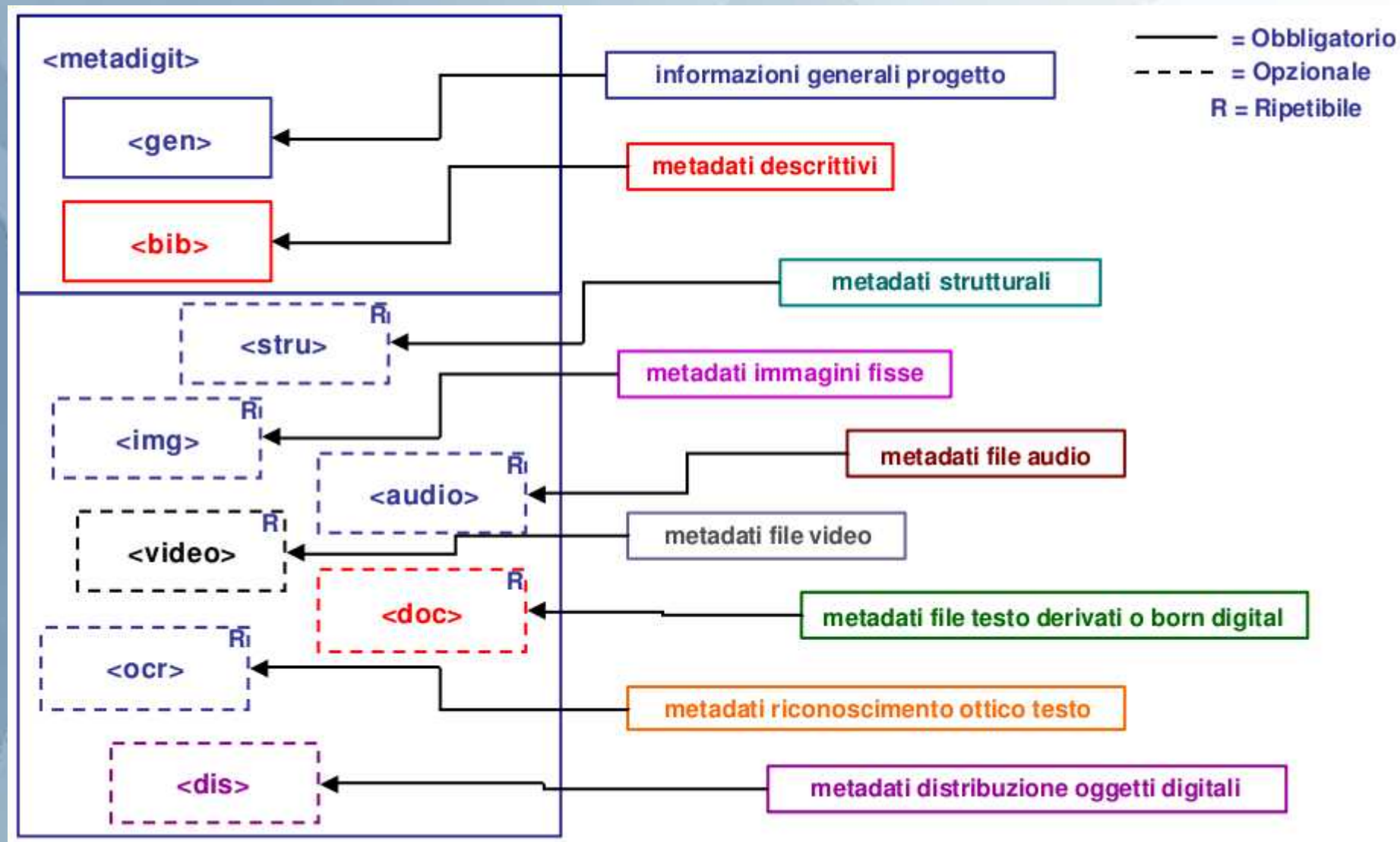
<doc> sezione utilizzata per descrivere un file testo (es. pdf o rtf)

<audio> metadati specifici per file audio

<video> metadati specifici per file video

<dis> metadati specifici per la distribuzione di oggetti bibliografici

Dati e metadati: MAG



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).

Le risorse digitali aggregate

le risorse **devono essere associate in insiemi/collezioni che siano organizzati / associati, per avere la garanzia di essere correttamente contestualizzate .**

Semplificando, i sistemi di risorse (collezioni digitali) possono essere:

- **Associati tramite dati esterni** (metadati strutturali)
- **Basarsi su collegamenti ipertestuali** (siti e applicazioni web, che approfondiremo più avanti)
- **Essere strutturati tramite Data Base**, quindi rispondere a una logica e sintassi definita, che può essere anche molto complessa e che comunque è la più rigida dei tre casi

Le risorse digitali aggregate

Un concetto importante: collezione digitale

Nei casi di **digitalizzazione** di materiali analogici, ma anche nella raccolta sistematica di risorse nate digitali si creano le cosiddette **collezioni digitali**, cioè aggregazioni ex novo di oggetti digitali.

Le associazioni possono essere fatte sulla base di **criteri diversi**: associazione **originaria** (serie di oggetti simili, collezioni artistiche, pagine di un libro, cataloghi di prodotti etc.), oppure per **tema**, **cronologia**, **obiettivi formativi** o **didattici**, **esposizioni**, **dimostrazioni**, etc.

Le risorse digitali aggregate

Descrizione delle collezioni

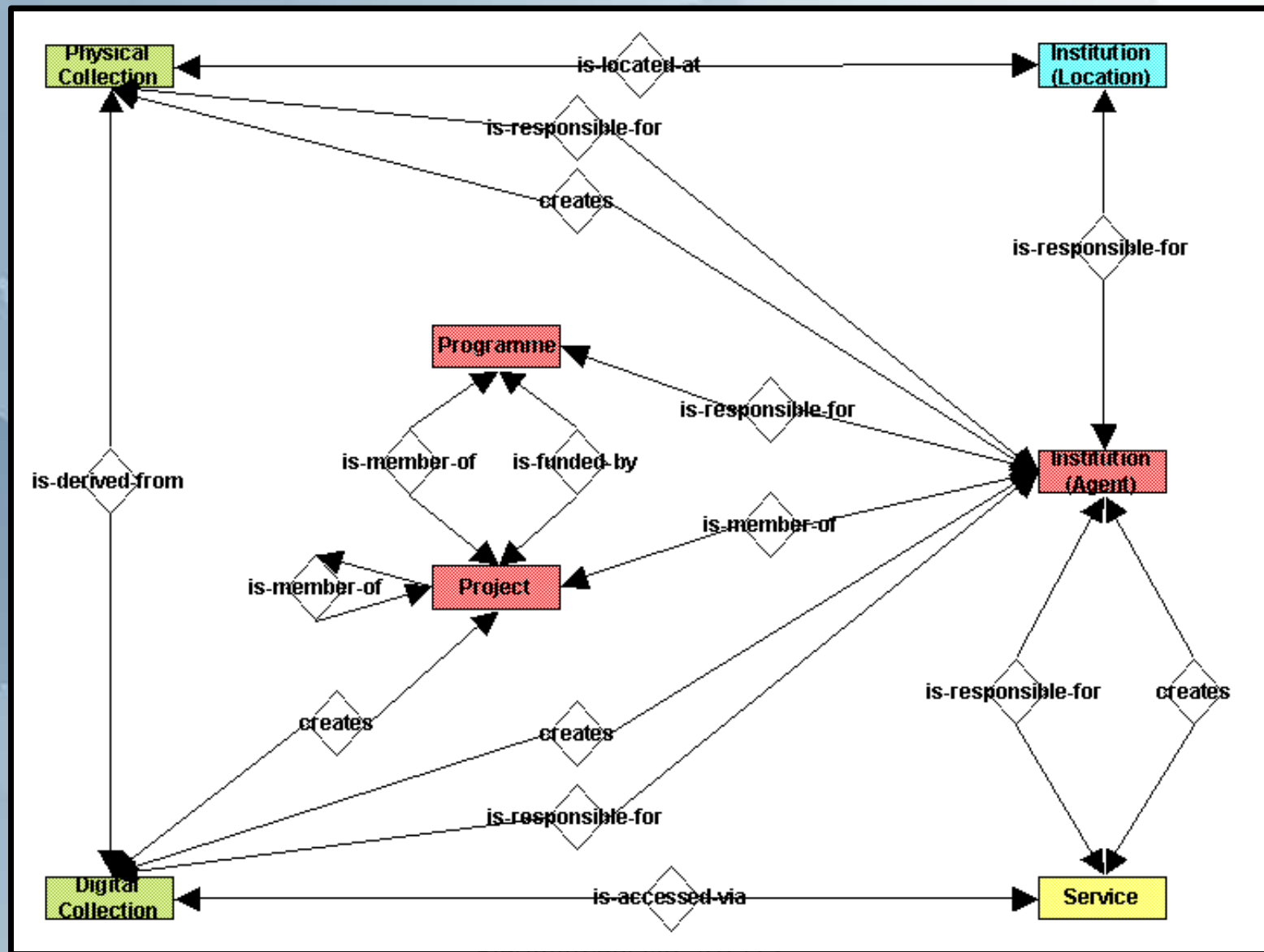
Per descrivere le collezioni su MICHAEL, ad esempio, si raccolgono informazioni su 5 entità distinte, in relazione tra loro:

- Collezione digitale
- Eventuale collezione/i fisica
- Organizzazione/i coinvolt/e
- Progetto/i all'origine della collezione
- Servizio/i che mettono la collezione a disposizione degli utenti



La descrizione delle collezioni

(schema relazionale)



Le risorse digitali aggregate: *data base*

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** (stessa struttura) e sono collegabili con altri insiemi. Il **modello dei dati** è un insieme di astrazioni che descrivono le associazioni tra i dati.

Le risorse digitali aggregate: *data base*

Una rappresentazione testuale di cosa significa informazione strutturata è quella di un file di testo che segue la sintassi XML (ne parleremo ancora):

```
<?xml version="1.0" encoding="UTF-8"?>
<utenti>
  <utente>
    <nome>Luca</nome>
    <cognome>Cicci</cognome>
    <indirizzo>Milano</indirizzo>
  </utente>
  <utente>
    <nome>Max</nome>
    <cognome>Rossi</cognome>
    <indirizzo>Roma</indirizzo>
  </utente>
</utenti>
```



Le risorse digitali aggregate

DBMS

Data Base Management System. Sistema centralizzato (programmi coordinati) o distribuito (rete) che permette di memorizzare, modificare ed estrarre informazioni da un database. **Garantisce l'integrità dei dati** (unica raccolta di dati anziché copie distinte scoordinate che potrebbero causare duplicazioni, ridondanze). **Organizza le informazioni** del database secondo la struttura di un database gerarchico, di un database di rete o di un database relazionale. **Garantisce l'accesso** alle informazioni solo a persone autorizzate, p.e. tramite una password. Un DBMS oltre a stabilire schemi organizzativi e di controllo, rende le informazioni accessibili agli utenti, tramite query.

Le risorse digitali aggregate

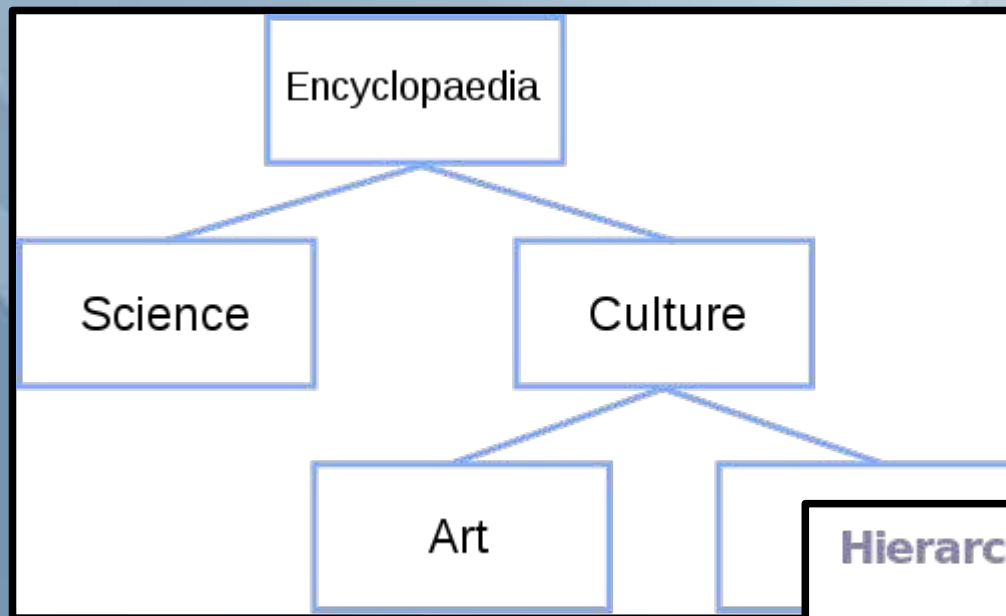
Tipi di database

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

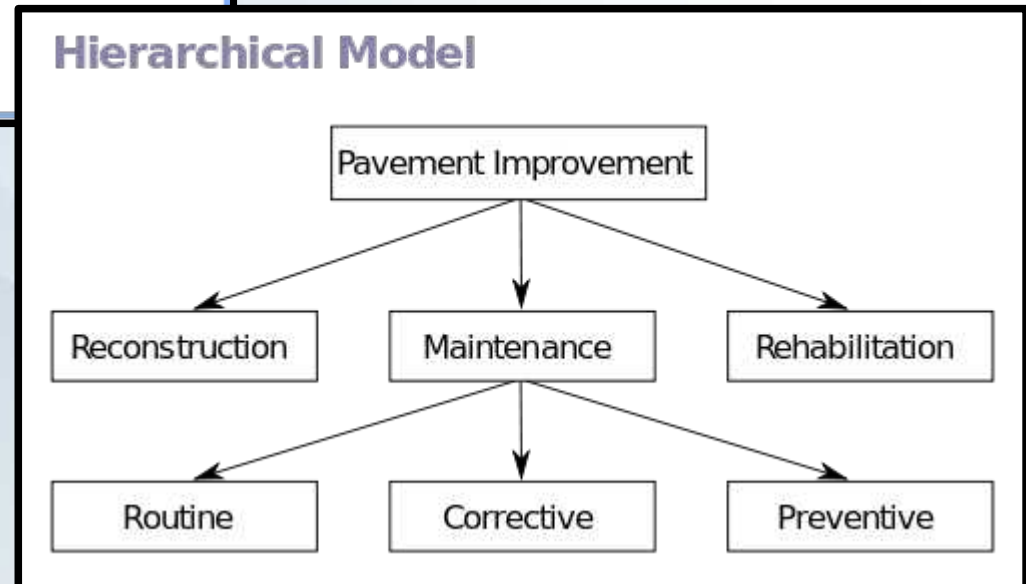
1. **gerarchica** (rappresentabile tramite un albero - anni sessanta),
2. **reticolare** (rappresentabile tramite un grafo - anni sessanta),
3. **relazionale** (attualmente **il più diffuso**, rappresentabile mediante tabelle e relazioni tra esse - anni settanta),
4. **ad oggetti** (estensione alle basi 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 supportato ad esempio dai DBMS relazionali, che consentono di salvare i dati in tabelle che possono essere collegate.

Le risorse digitali aggregate



Database
gerarchico



Le risorse digitali aggregate

data base relazionale

L'assunto fondamentale del modello relazionale è che **tutti i dati sono 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

data base
relazionale

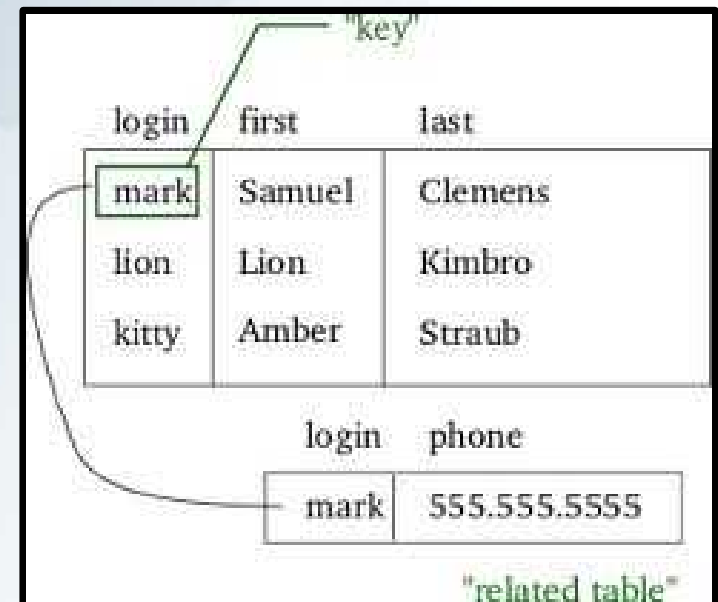
Relational Model

Activity Code	Activity Name
23	Patching
24	Overlay
25	Crack Sealing

Key = 24

Activity Code	Date	Route No.
24	01/12/01	I-95
24	02/08/01	I-66

Date	Activity Code	Route No.
01/12/01	24	I-95
01/15/01	23	I-495
02/08/01	24	I-66



Le risorse digitali aggregate

Possiamo distinguere, semplificando e solo a scopo didattico, **due tipi sostanziali di data base (banche dati)**, in base alla organizzazione delle macchine su cui sono conservati i dati e a dove si trova il software gestionale:

- con **dati organici** al DBMS
- con **dati indipendenti** dal DBMS

In generale, un **database distribuito** è un database che si trova sotto il controllo di un DBMS nel quale gli archivi *non sono memorizzati sullo stesso computer*.

Il database può essere dislocato in più computer situati nello stesso luogo, oppure distribuito in una rete di computer connessi tra loro.

Le risorse digitali aggregate

con dati organici al DBMS

in questo caso la struttura, il formato, le maschere di inserimento, le query e le modalità di aggregazione, ordinamento e output (anche Web) sono interamente gestite dal DBMS.

Questo se semplifica la fase di progettazione iniziale (modello “chiavi in mano”) può creare delle limitazioni e difficoltà per la portabilità e l’interoperabilità dei dati.

con dati indipendenti dal DBMS

in questo caso le strutture dei dati sono basate su formati leggeri e standard (XML prima di tutto)

le maschere di inserimento possono essere costruite con software di editing molto semplici

le query e le modalità di aggregazione, ordinamento, sicurezza, conservazione sono gestite dal DBMS

l’output e l’interoperabilità possono essere indipendenti dal DBMS, usando i dati originari

Le risorse digitali aggregate

i sistemi informativi

attualmente si cerca di avviare, nell'approccio alle risorse digitali per Archivi e Biblioteche, soprattutto, ma anche per le risorse culturali storico-artistiche, progetti che comprendano in un unico package le funzioni di descrizione, aggregazione, amministrazione, gestione, conservazione e comunicazione: i SI

“Il sistema informativo è costituito dall'insieme delle informazioni utilizzate, prodotte e trasformate da un'azienda durante l'esecuzione dei processi aziendali, dalle modalità in cui esse sono gestite e dalle risorse, sia umane, sia tecnologiche, coinvolte.”
(Wikipedia)

Sistemi informativi

Se dal punto di vista dei formati descrittivi associativi e gestionali valgono tutte le cose dette finora, vale la pena di accennare qualcosa sulle possibili **architetture dei sistemi informativi**.

Sinteticamente, questi possono essere **distribuiti o centralizzati**, ma i modelli sono in realtà ben più complessi, anche in relazione alla “dimensione”, cioè alla complessità e quantità dei dati e al contesto organizzativo.

Sistemi informativi

i **sistemi centralizzati** raccolgono i database (anche se magari non le funzioni gestionali e quelle di output Web) in un sistema centrale, alimentato tramite connessione di rete da n terminali di immissione periferici.

Non cambia molto se questa immissione avviene tramite poli intermedi, anche strutturati a più livelli.

Sistemi informativi

i pro e i contro dei **sistemi centralizzati**

- **i pro**: facilitano la omogeneità formale e la validazione dei dati e la creazione di liste d'autorità, evitando ripetizioni e sovrapposizioni (pensiamo al modello SBN)
- **i contro**: l'immissione dei dati è spesso lenta e faticosa, e non è sempre agevole valorizzare i contesti locali

Sistemi informativi

i sistemi distribuiti

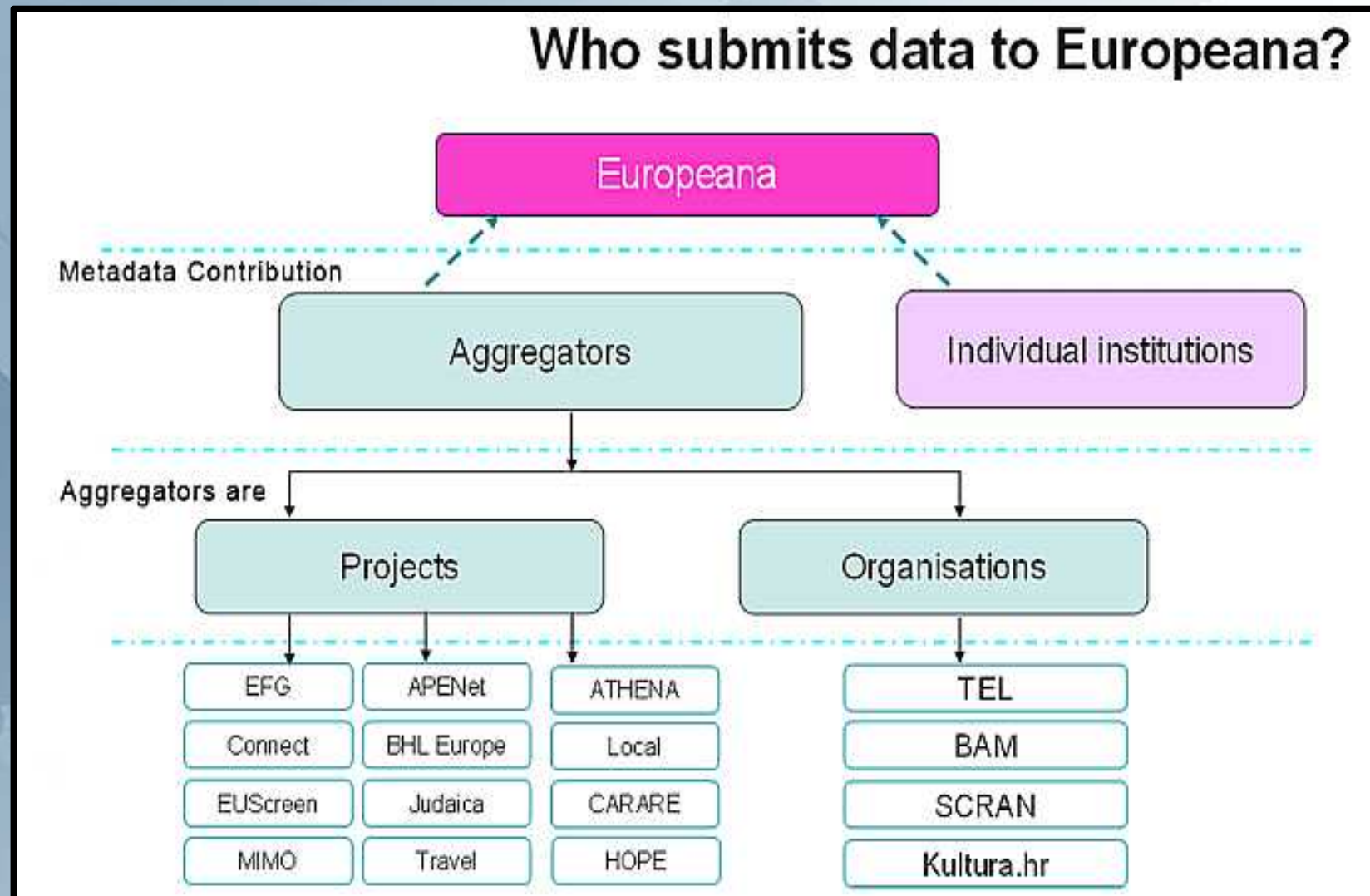
si basano sulla costruzione di n sistemi locali autonomi, che riversano tutti o parte dei loro dati su un sistema centrale deputato a funzioni particolari, tra cui quella Web. I sistemi locali sono autonomi ma nel loro rapporto col “centro” sono allineati attraverso procedure automatiche (ad es. OASIS) oppure no.

Sistemi informativi

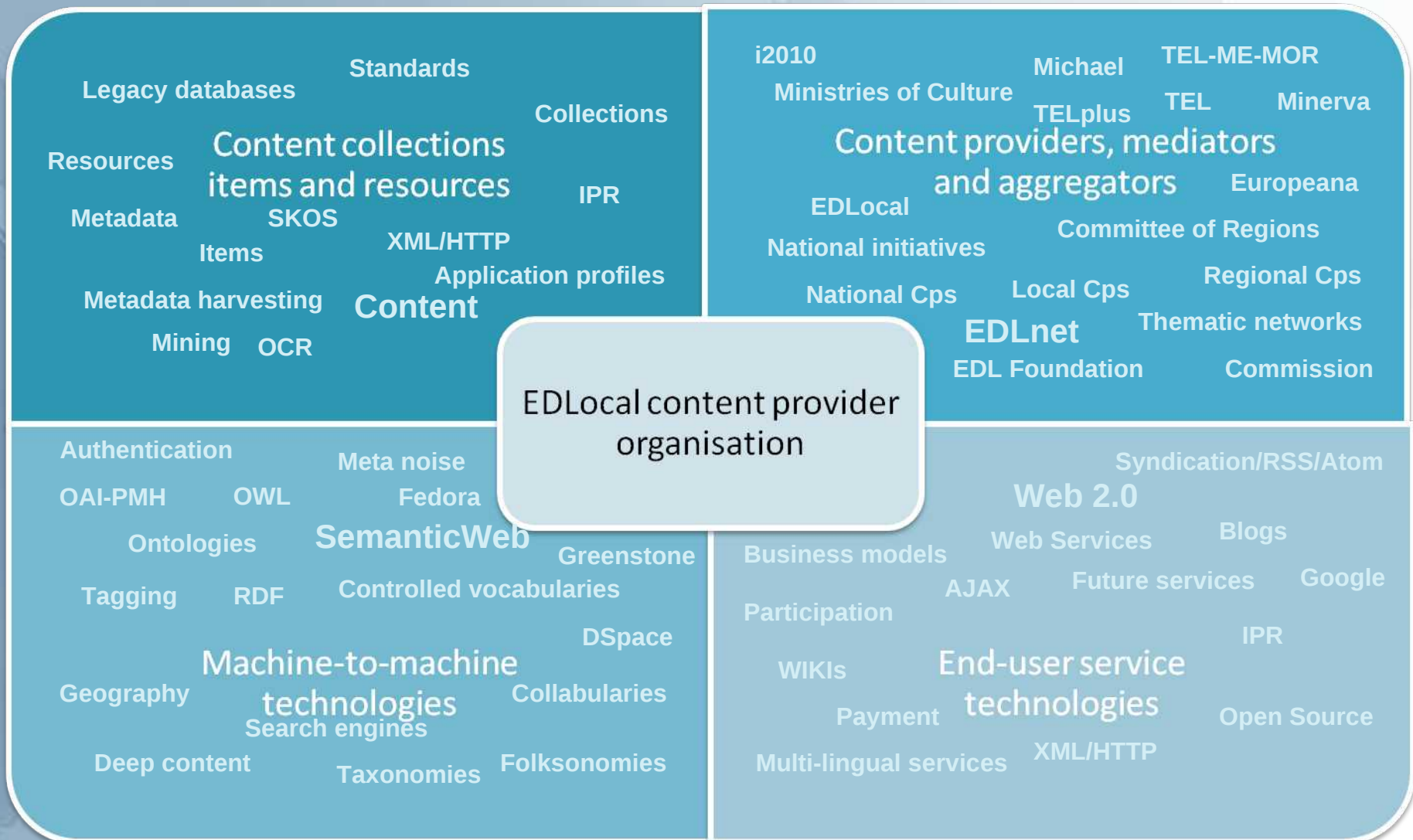
i pro e i contro dei **sistemi distribuiti**

- **i pro**: valorizzano i contesti locali permettendo personalizzazioni e applicazioni specifiche; velocizzano immissione e gestione dei dati
- **i contro**: le procedure di omogeneizzazione e validazione dei dati sono più delicate

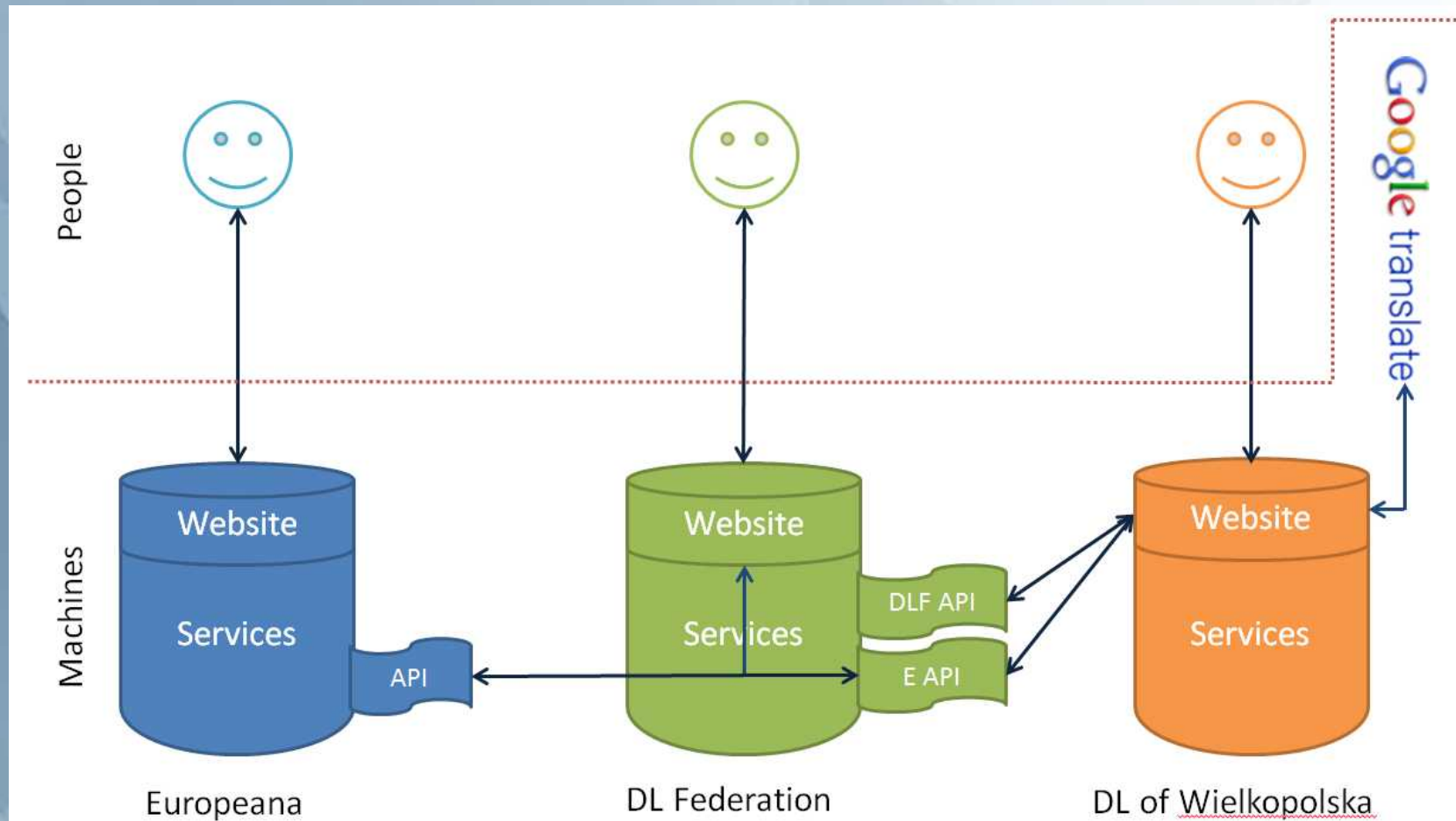
Sistemi informativi interoperabilità



Sistemi informativi fonti multiple



Sistemi informativi interoperabilità e utenti



Sistemi distribuiti rete di portali MICHAEL

