

Figura 2: Le astrazioni utilizzate nei modelli concettuali

Come si vede, si può arrivare ad uno stesso concetto (quello di PERSONA), per mezzo di tre procedimenti di astrazione diversi, ognuno dei quali, cioè, utilizza una diversa astrazione. Si può anche osservare che ognuna delle tre astrazioni coglie un aspetto della realtà che non può essere rappresentato per mezzo delle altre due.

1.1 Il modello Entità Relazione

In questa sezione descriviamo un particolare modello concettuale, che adotta in certa misura le astrazioni descritte nella sezione precedente. Una caratteristica interessante del modello è la adozione per i concetti di una rappresentazione grafica, che riportiamo in Figura 3.

Nel modello sono definiti cinque tipi di strutture di rappresentazione: entità, relazioni, attributi, sottoinsiemi, generalizzazioni. Sono anche definibili alcuni *vincoli di integrità*, che rappresentano proprietà esprimibili sui concetti presenti nello schema. Nel seguito esamineremo un particolare tipo di vincolo di integrità, le cardinalità.

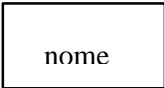
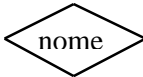

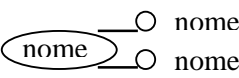

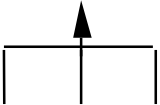
STRUTTURA DI CLASSIFICAZIONE	SIMBOLO
ENTITA'	
RELAZIONE	
ATTRIBUTO SEMPLICE	
ATTRIBUTO AGGREGATO	
SOTTOINSIEME	
GENERALIZZAZIONE	

Figura 3: Rappresentazione grafica dei concetti nel modello Entità Relazione

Entità

Le *entità* corrispondono a classi di oggetti del mondo reale (oggetti che chiameremo in seguito *istanze* della entità) che hanno proprietà omogenee ai fini della applicazione. Tra queste le proprietà elementari (cioè non strutturabili in proprietà più atomiche) sono dette *attributi semplici*. A un attributo semplice è associato un insieme di valori, detto anche *dominio*, che rappresentano l'insieme dei valori elementari che l'attributo può assumere. Così, ad esempio, PERSONA è una entità tipica di uno schema anagrafico, e NOME, COGNOME, ETÀ, SESSO, sono proprietà elementari che noi siamo interessati a descrivere di ogni persona, e quindi sono nel modello attributi della entità PERSONA. Accanto agli attributi semplici è comodo poter

definire *attributi composti*, che costituiscono aggregazioni di attributi. Ad esempio DATA può vedersi come attributo composto dagli attributi GIORNO, MESE, ANNO.

Relazioni

Le **relazioni** corrispondono a classi di fatti del mondo reale che sono significativi ai fini della applicazione; tali fatti mettono in relazione istanze di due o più entità. Ad esempio, in un censimento della popolazione possiamo essere interessati ad esprimere la relazione tra le persone e le città in cui sono nati, e chiamare tale relazione È-NATO; tale relazione può vedersi come astrazione di aggregazione applicata alle due entità PERSONA e CITTÀ. Accanto ad essa, possiamo essere interessati ad altre relazioni definite sulle stesse entità: ad esempio la relazione tra la persona e il luogo di residenza (relazione RISIEDE), la relazione tra persona e luoghi in cui ha avuto residenza (HA-RISIEDUTO).

Ancora riguardo alle relazioni, è bene osservare che accanto a relazioni definite tra due entità, come tutte le precedenti, ci possono essere relazioni definite su più di due entità. Un esempio è la relazione FORNITA, tra le entità FORNITORE, PARTE, PRODOTTO. Una parte può essere fornita da diversi fornitori, e per diversi prodotti. Se vogliamo cogliere la situazione più generale possibile, dobbiamo rappresentare questo tipo di fatti per mezzo di terne di informazioni (un fornitore, una parte, un prodotto), e dunque, nello schema, per mezzo di una relazione tra tre entità. In questo caso, la relazione ha un attributo, che rappresenta la generica QUANTITÀ che un fornitore fornisce per una certa parte e per un certo prodotto.

ESEMPIO

Lo schema di Figura 4 rappresenta la realtà informativa di un campionato di calcio. Nello schema compaiono:

1. La entità SQUADRA rappresenta l'insieme delle squadre che partecipano al campionato. Delle squadre vogliamo conoscere il NOME (ad esempio Juventus, Roma, Milan), la CITTÀ (nei tre casi precedenti Torino, Roma, Milano), ed il COGNOME dell'allenatore.
2. La entità GIOCATORE rappresenta i giocatori che giocano nelle squadre (in tutte le squadre). Per ogni giocatore vogliamo ricordare il NOME, il COGNOME, la DATA-DI-NASCITA (attributo composto dagli attributi GIORNO, MESE, ANNO) ed il RUOLO che ricopre nella squadra. Riguardo al RUOLO, facciamo per il momento la ipotesi che ogni giocatore abbia un unico ruolo, sempre fisso nel tempo. Facciamo anche la ipotesi che anche l'insieme dei ruoli sia fisso nel tempo (ad es. Portiere, Terzino sinistro, Ala destra, ecc.), ed in tutte le squadre si adottino gli stessi ruoli.
3. La entità PARTITA rappresenta l'insieme delle partite svolte nel campionato (ad esempio ROMA - MILAN della seconda giornata del girone di ritorno, JUVENTUS - VERONA della settima giornata del girone di andata, ecc.). Di ogni partita si vuole ricordare il GIRONE e la GIORNATA in cui si è svolta, la DATA (anche in questo caso distinta in GIORNO, MESE ed ANNO), ed il NUMERO della partita nell'ambito della giornata (ad esempio prima partita, seconda partita, ecc.).

Accanto alle precedenti entità sono definite nello schema le seguenti relazioni:

1. GIOCA-IN tra le entità GIOCATORE e SQUADRA, che rappresenta per ogni giocatore la squadra in cui gioca, ovvero, simmetricamente, per ogni squadra l'insieme dei giocatori che giocano nella squadra.
2. TRA, definita tra le due entità SQUADRA e PARTITA, che rappresenta per ogni partita le squadre coinvolte (ad es. nella partita ROMA - MILAN sono coinvolte le squadre ROMA e MILAN). La relazione TRA ha un attributo, chiamato GIOCA-IN-CASA, che per ognuna delle due squadre coinvolte nella partita dice se la squadra gioca o meno in casa. Il dominio di questo attributo è formato dai due valori [SI, NO]. È importante convincersi che l'attributo GIOCA-IN-CASA è una proprietà della relazione TRA, e non della entità PARTITA o della entità SQUADRA. Infatti per assegnargli un valore noi dobbiamo conoscere la partita e la squadra cui si riferisce: è perciò una proprietà della relazione tra le due entità.

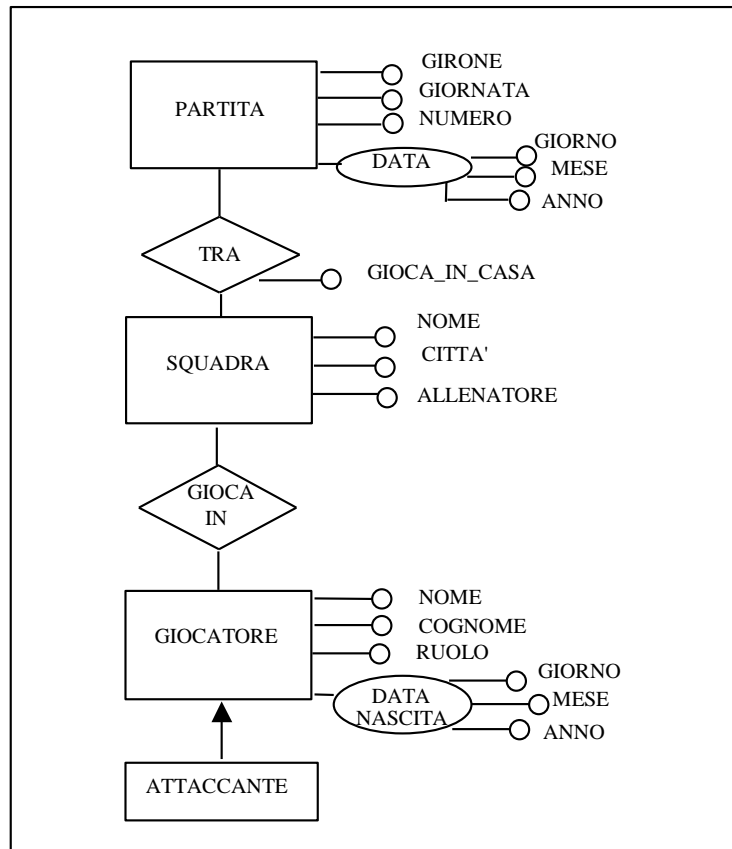


Fig. 4: Schema concettuale di un campionato di calcio

Generalizzazioni

Le Generalizzazioni mettono in relazione un insieme di entità (dette nel seguito entità figlie) con una nuova entità (detta entità padre) che di esse è astrazione appunto di generalizzazione. Possiamo ad esempio affermare che PERSONA è generalizzazione di UOMO e DONNA, ovvero che LUOGO è generalizzazione di COMUNE e STATO ESTERO. Nelle generalizzazioni potranno essere definiti attributi e relazioni che sono significativi solo per una delle entità figlie. È questo il caso dello schema di Figura 5 in cui l'attributo SITUAZIONE-MILITARE (che rappresenta lo stato della persona dal punto di vista del servizio militare: esente, arruolato, ecc.) ha senso solo per gli uomini.

Una fondamentale proprietà delle generalizzazioni è la seguente: in una astrazione di generalizzazione, ogni proprietà della entità padre è anche proprietà delle entità figlie. Per proprietà intendiamo gli attributi, le relazioni e le generalizzazioni cui partecipa la entità. Così, se ETÀ è attributo di PERSONA e PERSONA è generalizzazione di UOMO e DONNA, ETÀ è chiaramente attributo anche di tali due nuove entità.

Si noti che una entità può essere l'entità padre di più generalizzazioni: ad esempio, in una applicazione scolastica, la entità PERSONA può essere da una parte generalizzazione nelle entità PROFESSORE e STUDENTE, e dall'altra nelle entità UOMO e DONNA.

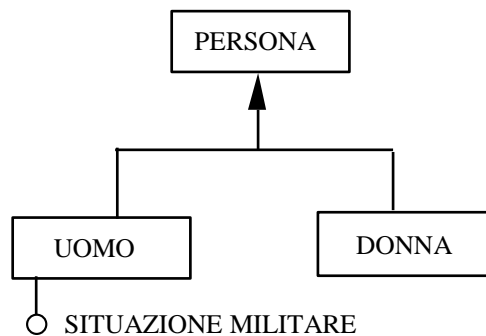


Figura 5: Un esempio di generalizzazione

Sottoinsiemi

Un sottoinsieme è un caso particolare di generalizzazione, quella che sussiste tra l'entità padre e una sola entità figlia. Ad esempio la entità ATTACCANTE è sottoinsieme di GIOCATORE nello schema del campionato, perché ogni giocatore attaccante è anche un giocatore, ma esistono giocatori che non sono attaccanti. Anche per i sottoinsiemi è definita la proprietà in precedenza definita per le generalizzazioni. Ad esempio, tutti gli attributi e le relazioni associati a GIOCATORE nello schema di Figura 4 vanno automaticamente associati anche alla entità ATTACCANTE.

Cardinalità

Le cardinalità sono proprietà di relazioni; la cardinalità minima di una entità definita in una relazione è il minimo numero di volte che ogni occorrenza della entità può essere coinvolta in una occorrenza della relazione. Il valore 0 significa che può esistere una occorrenza di entità non coinvolta in alcuna occorrenza di relazione. Il valore 1 (n) significa che non può esistere una occorrenza senza essere coinvolta in 1 (n) relazioni. Analoga definizione vale per la cardinalità massima, che definisce il numero massimo di istanze di relazione in cui può essere coinvolta una occorrenza di un'entità. Le cardinalità minime e massime n ed m di una entità in una relazione sono indicate con il simbolo (n,m) accanto alla entità. Ad esempio, nella relazione TRA dello schema di Figura 4, le cardinalità sono:

1. (2,2) per la entità PARTITA (una partita si svolge esattamente tra due squadre).
2. (1,1) per la entità SQUADRA (una squadra è coinvolta esattamente una volta in una partita).

Nella relazione GIOCA-IN le cardinalità sono:

1. (11, n) per la entità SQUADRA (una squadra deve avere come minimo 11 giocatori, ma in generale ne ha molti di più).
2. (1,1) per la entità GIOCATORE (ipotizzando che un giocatore, in un campionato, giochi solo in una squadra).

2 Possibili ulteriori utilizzi degli schemi dei dati nei sistemi informativi

In questo capitolo abbiamo visto l'attività di produzione di uno schema dei dati come strettamente legata alla attività di analisi. In effetti gli schemi dei dati possono essere prodotti e utilizzati anche in diversi altri contesti. Essi sono:

- La individuazione del contenuto informativo comune a due o più basi di dati;
- La individuazione delle ridondanze e delle incoerenze tra più basi di dati;

- La produzione dello schema dati della organizzazione;
- L'integrazione di due o più schemi di dati.

Per ciascuno di essi vediamo il significato, l'utilità nel ciclo di vita dei sistemi informativi, come possa essere svolta la attività, e qualche esempio di utilizzo

2.1 Individuazione del contenuto informativo comune a due o più basi di dati

Abbiamo detto che le basi di dati del sistema informativo di una grande organizzazione sono sempre sviluppate nel tempo in maniera disorganica. Per esempio la sola Pubblica Amministrazione centrale italiana gestisce circa 600 basi di dati di dimensione maggiore di un gigabyte, sviluppate negli ultimi 30 anni. Può essere utile per molte ragioni comprendere quale sia il contenuto informativo comune a due o più basi di dati, ed in particolare:

- per capire se sia opportuno unificare o fondere le due basi di dati, allo scopo di aumentare il numero di interrogazioni o elaborazioni che possono essere effettuate sull'insieme delle due basi di dati, e semplificarne la gestione e l'aggiornamento.
- per capire, anche mantenendo distinte le due basi di dati, quali interrogazioni o elaborazioni ulteriori potrebbero essere fatte collegando le due basi di dati in un sistema distribuito.
- Per capire quali ulteriori informazioni dovrebbero essere descritte in una o in entrambe le basi di dati per accrescere il contenuto informativo complessivo delle basi di dati.
- Per individuare le eventuali ridondanze, intendendo per ridondanza la presenza delle stesse tipologie di informazioni nelle diverse basi di dati. In un sistema distribuito, la presenza di ridondanze è una scelta che va fatta con molta cura, perché ogni ridondanza obbliga a gestire gli aggiornamenti delle diverse copie della stessa informazione attraverso elaborazioni che possono essere anche molto complesse, quanto più l'informazione è duplicata nel sistema.

Avendo a disposizione gli schemi concettuali di partenza, il contenuto informativo comune può essere individuato analizzando le entità e le relazioni dei due schemi e selezionando le entità e relazioni con gli stessi nomi, e quelle che pur non avendo gli stessi nomi risultano corrispondere allo stesso concetto del mondo reale (sinonimi).

2.2 Individuazione delle incoerenze tra più basi di dati

In questo caso vogliamo comprendere quanto le scelte progettuali effettuate in termini di informazioni rappresentate negli schemi abbiano portato ad introdurre scelte concettuali diverse nella rappresentazione della stessa tipologie di informazione. Per esempio, in una base dati anagrafica è possibile modellare l'indirizzo di residenza con diverse modalità concettuali:

- a. con un unico attributo **INDIRIZZO**, rappresentato con una stringa di caratteri di lunghezza fissata (esempio, 50 caratteri alfabetici)
- b. con quattro attributi **VIA**, **NUMERO CIVICO**, **CODICE AVVIAMENTO POSTALE**, **CITTÀ**, rappresentati rispettivamente con una stringa di caratteri, un numero, un numero, una stringa di caratteri.
- c. Con opportune varianti delle scelte precedenti.

Spesso, nel passato, al fine di ottimizzare lo spazio occupato in memoria, sono state effettuate scelte di tipo a, che portano ad una riduzione del numero degli attributi. Ma anche in questo caso, se ci si trovava a realizzare diverse basi di dati in cui compariva lo stesso attributo Indirizzo, potevano essere scelte per le stringhe di caratteri diverse lunghezze. Nella Pubblica amministrazione centrale esistono decine di basi di dati in cui è rappresentato un attributo indirizzo: le basi di dati fiscali, quelle catastali, quelle previdenziali, e così via.

Una diversa rappresentazione delle stesse informazioni in basi di dati diverse porta a diversi effetti indesiderati: prima di tutto non è più possibile, o, meglio, diventa molto più difficile, collegare le basi di dati per il tramite della informazione rappresentata in modo incoerente, perché essa può assumere valori diversi. In

secondo luogo, diventa molto oneroso l'aggiornamento delle diverse copie. In terzo luogo diventa oneroso qualunque progetto di integrazione delle basi di dati, perché è necessario ricostruire la versione corretta della informazione.

2.3 Integrazione di due o più schemi di dati

Avere a disposizione gli schemi concettuali delle basi di dati diventa importante anche in qualunque attività di integrazione delle stesse (vedi Figura 6).

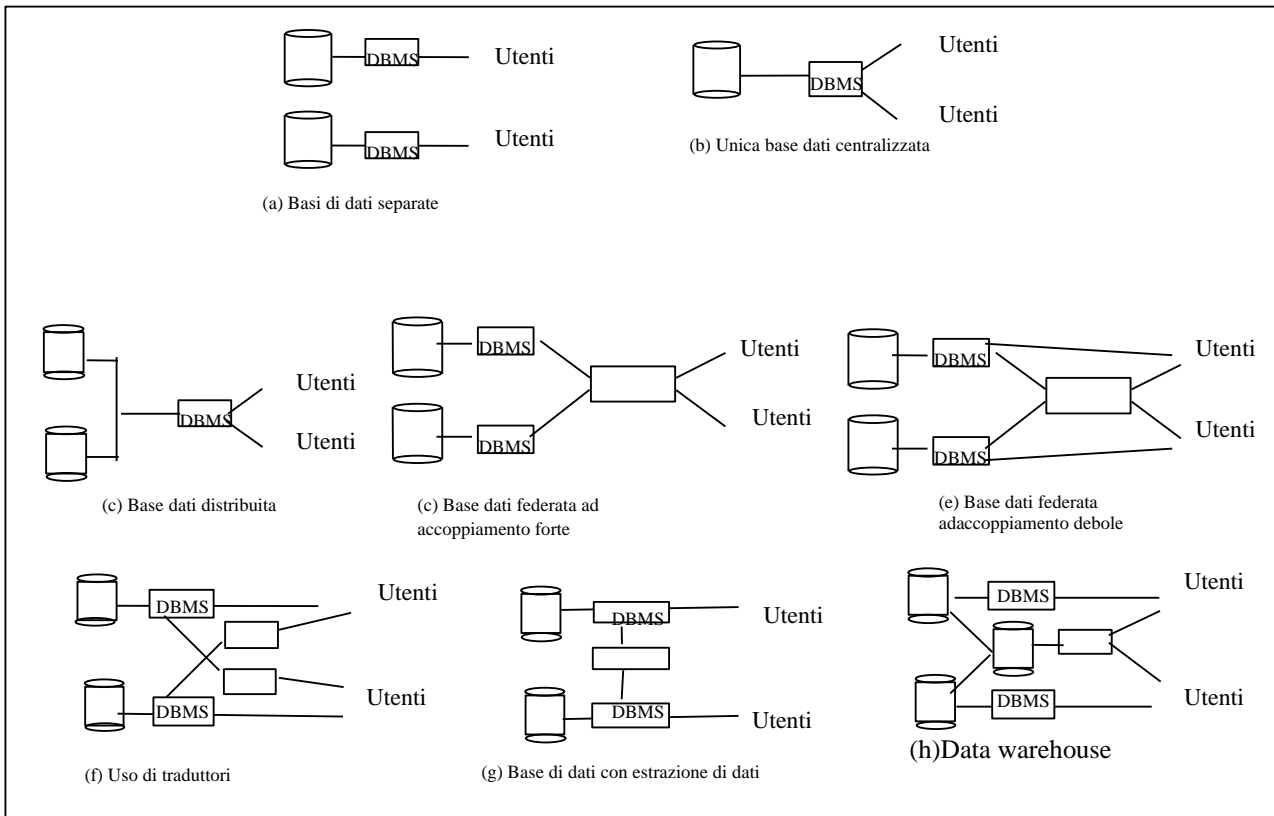


Figura 6: Forme di integrazione tra basi di dati

La forma più stretta di integrazione tra basi di dati si ottiene mediante la integrazione delle diverse basi di dati in una unica base dati (Figura 6.b), cioè il ritorno alla precedente situazione di totale centralizzazione. In questo caso, avere a disposizione gli schemi concettuali delle basi di dati da integrare permette di arrivare ad una integrazione di natura semantica tra le diverse basi di dati, che tiene cioè conto del significato delle informazioni presenti, e facilita perciò la comprensione tra gli utenti e la correttezza delle elaborazioni. Per fare un esempio, se nelle attuali basi di dati compaiono due archivi in uno dei quali sono riportati in distinte tabelle i fatturati di un insieme di aziende nei dodici distinti mesi dell'anno, espressi in migliaia di lire e nell'altro tale insieme di informazioni è rappresentato in una unica tabella, con valori espressi in milioni, nella base di dati integrata tali *differenze di rappresentazione* vengono rimosse, dando luogo ad una unica rappresentazione omogenea. Questa forma di integrazione è in realtà molto difficile da attuare, per la grande rapidità con cui dati e basi di dati evolvono nelle organizzazioni, e spesso inopportuna o non realizzabile, per la autonomia che utenti e gruppi di utenti intendono mantenere sulle basi di dati da essi create e aggiornate.

Una variante della precedente scelta è quella delle *basi di dati distribuite*, in cui la base di dati è ancora gestita da un unico DBMS, ma può essere frammentata in tante basi di dati locali. Ciò in genere migliora l'efficienza per accessi locali, ma rende più complessi problemi di aggiornamento su diverse basi di dati. In questa soluzione, usualmente, lo schema concettuale dell'insieme delle basi di dati è l'elemento di partenza per compiere le scelte razionali sulle modalità di distribuzione.

Tutte le soluzioni successivamente esposte indeboliscono progressivamente il tipo di integrazione tra le diverse basi di dati. In una *base dati federata ad accoppiamento forte* le diverse basi di dati vengono gestite autonomamente da diversi DBMS, ma esiste un nuovo componente software che ha la possibilità di accedere omogeneamente alle diverse basi di dati, coordinandone le operazioni. In questo caso lo schema concettuale è indispensabile per permettere a tale componente software di comprendere dove siano allocate nella federazione le diverse tipologie di informazione, e in quale formato concettuale e logico. In una *base di dati federata ad accoppiamento debole*, rispetto alla precedente, il nuovo componente garantisce soltanto una visione virtuale integrata, ma non ha potere di intervento gestionale sui sistemi che rimangono fortemente autonomi.

Le tecnologie a *gateway* (traduttori o filtri tecnologici) permettono di tradurre richieste formulate in un linguaggio per un DBMS A in termini di richieste per il linguaggio adottato dal DBMS B. Hanno il vantaggio che non richiedono nessuno o limitato sviluppo di software aggiuntivo, ma richiedono componenti ad hoc per ogni singola relazione tra sistemi distinti, e fanno perdere la trasparenza resa possibile dalle soluzioni precedenti.

Il meccanismo più debole di integrazione tra basi di dati è quello della *estrazione e trasferimento di dati*, con aggiornamento periodico della base di dati locale a partire dalla base di dati remota.

In tutte le precedenti forme di integrazione, questa viene instaurata operando sulle basi di esistenti con nuovi componenti architetturali. Nei datawarehouse (magazzini o depositi di dati) viene creata (periodicamente) una nuova base di dati, che nasce dalla integrazione stretta delle basi di dati esistenti, e tale nuova base di dati viene resa accessibile con potenti linguaggi per l'analisi e la estrazione di informazioni rilevanti a fini decisionali. La produzione dello schema concettuale integrato delle basi di dati è il passo preliminare fondamentale a tutte le attività di produzione del datawarehouse .