Enciclopedia dei dati digitali

Carlo Batini

Libro Secondo

I modelli dei dati ci aiutano a rappresentare e comprendere il mondo

Versione 1
2 febbraio 2021

Indice

Capitolo 1
I dati come rappresentazioni del mondo

Il mondo in cui viviamo è ricco di infinite sfumature, fatte di eventi, luoghi, persone, oggetti, panorami, montagne, esperienze, emozioni, sentimenti. Quando vogliamo fare in modo che tutta questa ricchezza di percezioni sia rappresentabile e elaborabile da un calcolatore digitale noi dobbiamo sempre compiere un'operazione di trasformazione della realtà percepita in una rappresentazione digitale comprensibile e elaborabile. Chiameremo nel seguito questa operazione con il termine di *modellazione*.

Se per esempio noi viviamo a Milano, e vogliamo sapere quante persone sono residenti nella nostra città, quale percentuale è minorenne, quanto tempo si impiega per andare da casa nostra al Duomo a piedi, quando arriverà la prossima vettura della metropolitana, e tutto questo lo vogliamo sapere senza fare una personale ricerca, ma per mezzo del nostro telefono cellulare, noi dobbiamo prima rappresentare i dati che ci servono in una memoria digitale. Poi, qualcuno deve scrivere un programma per fornire una risposta alla nostra esigenza.

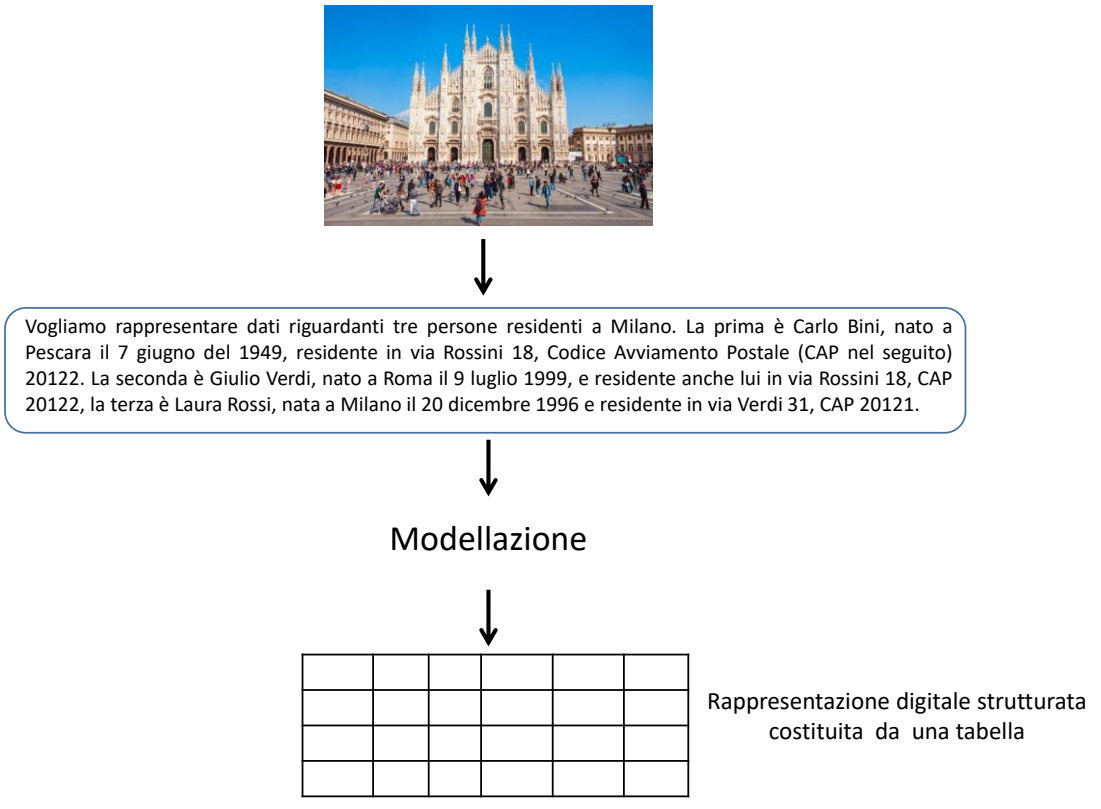


Figura 1: Il processo di modellazione di dati strutturati

Tutto questo processo è rappresentato nella Figura 1; per poter dare risposte a domande, in questo caso, sul Duomo di Milano (ad es. quanto è alto, quante persone sono in questo momento sul sagrato o in coda per entrare) noi dobbiamo anzitutto effettuare una operazione di ***modellazione***, che produce una rappresentazione digitale costituita, nella più

semplice delle modellazioni, da un insieme di *dati dotati di una struttura*, ad esempio una tabella o un insieme di tabelle, che nel mondo dell'informatica sono chiamate con il termine di ***base di dati***.

Nella Figura 1 ho mostrato inizialmente un'immagine di Milano come riferimento generale, ma ho anche immaginato di descrivere i dati che ci interessa rappresentare per mezzo di un testo in lingua italiana. E' questo testo che, dopo opportune trasformazioni, verrè rappresentato di qui a poco con un insieme di dati strutturati.

Attenzione: non tutti i fenomeni del mondo possono essere rappresentati con un insieme di dati strutturati. In Figura 2 mostriamo un altro testo, famoso, che non ci sogneremmo mai di tentare di rappresentare per mezzo di una tabella, tanto è privo di qualsiasi struttura. Riuscite a riconoscere del libro in cui compare questo testo, e del suo autore? Se non vi viene in mente, guardate la frase finale……

La soluzione all'inizio della prossima pagina.

……I saw them not long ago I love flowers Id love to have the whole place swimming in roses God of heaven theres nothing like nature the wild mountains then the sea and the waves rushing then the beautiful country with the fields of oats and wheat and all kinds of things and all the fine cattle going about that would do your heart good to see rivers and lakes and flowers all sorts of shapes and smells and colours springing up even out of the ditches primroses and violets nature it is as for them saying theres no God I wouldnt give a snap of my two fingers for all their learning why dont they go and create something I often asked him atheists or whatever they call themselves go and wash the cobbles off themselves first then they go howling for the priest and they dying and why why because theyre afraid of hell on account of their bad conscience ah yes I know them well who was the first person in the universe before there was anybody that made it all who ah that they dont know neither do I so there you are they might as well try to stop the sun from rising tomorrow the sun shines for you he said the day we were lying among the rhododendrons on Howth head in the grey tweed suit and his straw hat the day I got him to propose to me yes first I gave him the bit of seedcake out of my mouth and it was leapyear like now yes 16 years ago my God after that long kiss I near lost my breath yes he said I was a flower of the mountain yes so we are flowers all a womans body yes that was one true thing he said in his life and the sun shines for you today yes that was why I liked him because I saw he understood or felt what a woman is and I knew I could always get round him and I gave him all the pleasure I could leading him on till he asked me to say yes and I wouldnt answer first only looked out over the sea and the sky I was thinking of so many things he didnt know of Mulvey and Mr Stanhope and Hester and father and old captain Groves and the sailors playing all birds fly and I say stoop and washing up dishes they called it on the pier and the sentry in front of the governors house with the thing round his white helmet poor devil half roasted and the Spanish girls laughing in their shawls and their tall combs and the auctions in the morning the Greeks and the jews and the Arabs and the devil knows who else from all the ends of Europe and Duke street and the fowl market all clucking outside Larby Sharons and the poor donkeys slipping half asleep and the vague fellows in the cloaks asleep in the shade on the steps and the big wheels of the carts of the bulls and the old castle thousands of years old yes and those handsome Moors all in white and turbans like kings asking you to sit down in their little bit of a shop and Ronda with the old windows of the posadas 2 glancing eyes a lattice hid for her lover to kiss the iron and the wineshops half open at night and the castanets and the night we missed the boat at Algeciras the watchman going about serene with his lamp and O that awful deepdown torrent O and the sea the sea crimson sometimes like fire and the glorious sunsets and the figtrees in the Alameda gardens yes and all the queer little streets and the pink and blue and yellow houses and the rosegardens and the jessamine and geraniums and cactuses and Gibraltar as a girl where I was a Flower of the mountain yes when I put the rose in my hair like the Andalusian girls used or shall I wear a red yes and how he kissed me under the Moorish wall and I thought well as well him as another and then I asked him with my eyes to ask again yes and then he asked me would I yes to say yes my mountain flower and first I put my arms around him yes and drew him down to me so he could feel my breasts all perfume yes and his heart was going like mad and yes I said yes I will Yes. Trieste-Zurich-Paris 1914-1921

Figura 2 – Un testo totalmente privo di qualunque struttura

Stiamo parlando dell'ultima parte dell'Ulisse di Joyce, in cui la protagonista Molly Bloom, mentre si sta addormentando, si perde in un meraviglioso flusso di coscienza fatto di pensieri, sensazioni e sogni. Se non l'avete letto …. nessun problema!

I dati strutturati, come vedremo tra poco, sono rappresentati nei modelli per basi di dati mediante uno ***schema*** e un insieme di ***valori***.

Nelle prossime sezioni vedremo insieme cosa si intenda con questi due concetti, schema e valori, mostrando alcuni dei principali modelli usati in informatica per rappresentare un frammento di mondo attorno a noi. E vedremo anche come questi modelli siano diventati sempre più ricchi nell'esprimere il significato dei dati, nel rappresentare il mondo.

Chiaramente, il mondo è troppo complesso e ricco di sfumature per sperare di rappresentarlo in maniera completa, gli esempi che faremo sono veramente elementari, e sono il primo a rendermene conto. Ma non dobbiamo neanche essere troppo critici, il bagaglio di concetti che i modelli dei dati ci offrono, e che ci apprestiamo a conoscere, è già adeguato per migliorare il modo con cui possiamo osservare il mondo e con cui possiamo comprenderlo meglio.

I concetti più importanti

I concetti più importanti che condivideremo sono i seguenti:

- Modello dei dati
- Tabella
- Modello relazionale dei dati
- Linguaggio di interrogazione
- Modello Entità Relazione
- Modello a grafo semantico

Un lettore che voglia compiere un percorso più breve di apprendimento, può saltare in prima lettura gli aspetti riguardanti i seguenti concetti

- Linguaggio di interrogazione
- Modello Entità Relazione.

Le parole crociate senza schema e con schema

Avete mai passato il tempo a fare le parole crociate? Se non avete mai avuto questo come passatempo, allora saltate questa sezione che vi fa solo perdere tempo. Se invece avete un pò di esperienza di parole crociate, sapete certamente che ce ne sono almeno due tipi:

- *senza schema*, anche dette a schema libero, e
- *con schema*, o a schema fisso.

Esempi dei due tipi sono in Figura 3. In entrambi I casi noi dobbiamo inserire in una figura formata da diverse righe e colonne alcune parole, per le quali vengono fornite definizioni o

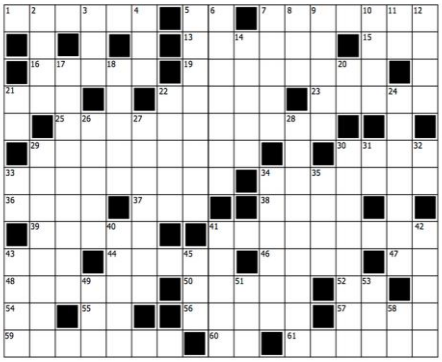
frasi che le richiamano. Le parole devono essere precedute e seguite da celle di colore nero, che segnalano da dove iniziano e dove terminano le parole, disposte in orizzontale o in verticale. Le parole da inserire in orizzontale e verticale sono definite più in basso.
Nelle *parole crociate senza schema*, non abbiamo nessun aiuto nel capire dove disporre le celle di colore nero. Nelle *parole crociate con schema*, le celle nere, che separano le parole e ne segnalano l'inizio, sono rappresentate esplicitamente, e costituiscono una guida per il riempimento delle restanti celle della matrice.

CASELLE NERE::28

ORIZZONTALI
[1] xeres ···agli estremi-- la provincia di ostuni [2] citta' dell'inghilterra celebre per le porcellane [3] istitutrice inglese--- si rende parlando [4] delfini d'acqua dolce-- dovuto per gli inglesi, irrinunciabile [5] nome dell'artista von schiele-- importante porto giapponese [6] con disegno a mosaico [7] nel 1929 entro-' a far parte della general motors-- l'ex ginnasta korbut [8] nome di urzi-- il burrasca di vamba [9] jan hendrik astronomo olandese-- figlio di eracle e deianira [10] brandina,giaciglio [11] latine della propaganda-- in geometri'a c'e' quello "greco"

VERTICALI
[1] eccessivamente dolce o lezioso [2] un movimento rotatorio nel golf-- nome della perego in tv [3] suffisso da carboidrato-- la-sedicesima dell'alfabeto [4] lavorano nei cantieri militari marittimi [6] periodo di prosperita' spazzato via con la prima guerra-mondiale [8] relativo all'origine delle parole [9] antichi celti-- la marlen di una nota canzone [10] citta' e porto della francia-- un antico ballo [11] capo cosacco

b. Parole crociate senza schema



ORIZZONTALI
1. Le profondità oceaniche 5. Savona sulle targhe 7. XXX 13. Dottrina dell'ebraismo medievale
15. Può imporlo il vigile urbano 16. Uno strumento nella torre di controllo 19. Valutano e interpretano dati e tendenze 21. Il massimo della gamma 22. La Calamai diva del cinema 23. XXX 25. XXX 29. XXX
30. Città di un famoso Jacopone 33. Quello cinese è detto anche Kumquat 34. La scienza applicata alla coltivazione della terra. 36. Li comandava Attila 37. L'isola natale di Ippocrate 38. XXX 39. L'allenatore del Milan 41. XXX 43. Il Fleming romanziere. 44. XXX 46. Curva di fiume. 47. Simbolo dell'arsenico
48. Seccante, palloso 50. Sono famose quelle senesi 52. Iniziali di Galvani 54. In mezzo alla mela
55. Sono pari nella mucca 56. Lo strumento di Ravi Shankar 57. Ammasso stellare nella costellazione del Toro 59. Fare una pausa durante la marcia 60. Le vocali di moda 61. Fa sempre piacere riceverlo

VERTICALI
2. XXX 3. Un punto cardinale 4. La banca vaticana 5. È affine alla cipolla 6. XXX 7. Saluto arabo
8. Parti laterali di un edificio 9. Enorme pietra 10. Ardita incursione in campo nemico 11. Articolo determinativo
12. Il numero che può essere volante 14. XXX 17. La catena con i monti Cimone e Pollino 18. Città portuale della Giordania 20. Topo senza coda 21. Lo si dà agli amici 22. XXX 24. Confina con Michigan e Illinois
26. XXX 27. Un fan di Renato Zero 28. Si è laureato al Politecnico 29. Metropoli brasiliana 30. XXX
31. La fine di Belfagor 32. Un dittongo di Biagio 33. Iniziali di Utrillo 34. Un elemento distintivo dei vini
35. Diana cantante di black-music 40. Resta dopo avere mangiato il branzino 41. XXX 42. XXX
43. Il nome della Sastre modella e attrice 45. Incognita matematica 49. Una chiamata del giudice di linea
51. Aumenta col passare degli anni 53. La trovata del comico 58. La prima preposizione semplice

a. Parole crociate con schema

Figura 3 – Le parole crociate senza schema (o a schema libero) e con schema

Il concetto di schema, che ritroveremo tra poco, esprime dunque una guida, una struttura che aiuta a comprendere l'organizzazione e il significato della parte della matrice che va completata con le parole orizzontali e verticali. Chiunque abbia una volta riempito delle parole crociate, sa bene quanto questa guida aiuti nell'arrivare alla soluzione finale e alla matrice completata.

Cerchiamo ora di capire come tutto ciò riguardi i dati.

Capitolo 2

Come si fa a organizzare un insieme di dati in una struttura, e a utilizzarli per i nostri scopi – Il modello relazionale

Ogni giorno noi entriamo in contatto con tanti dati in forma di numeri, parole, testi scritti, immagini, suoni. Per poter utilizzare i dati per i nostri scopi, ad esempio per prenotare un posto in treno o fare una ricerca bibliografica sul Web, abbiamo *prima* bisogno di organizzarli in una struttura che ci permetta di comprenderne il significato e i legami. Fornendo I dati di una struttura, noi riduciamo anzitutto la confusione nella rappresentazione del mondo che percepiamo attorno a noi.

Torniamo al nostro esempio iniziale di Milano e supponiamo, riprendendo l'esempio del Capitolo 5 del Primo volume della Enciclopedia, che qualcuno ci mandi un messaggio, rappresentato in Figura 4, in cui ci descrive alcune caratteristiche di tre milanesi. Provate a leggere con attenzione il testo e poi rispondere a queste domande:

1. Dove sono nate le persone (citate nel messaggio) che vivono in un indirizzo che ha CAP 20127?
2. Quando è nato Giulio Verdi?
3. Quante persone sono nate dopo il 1950?

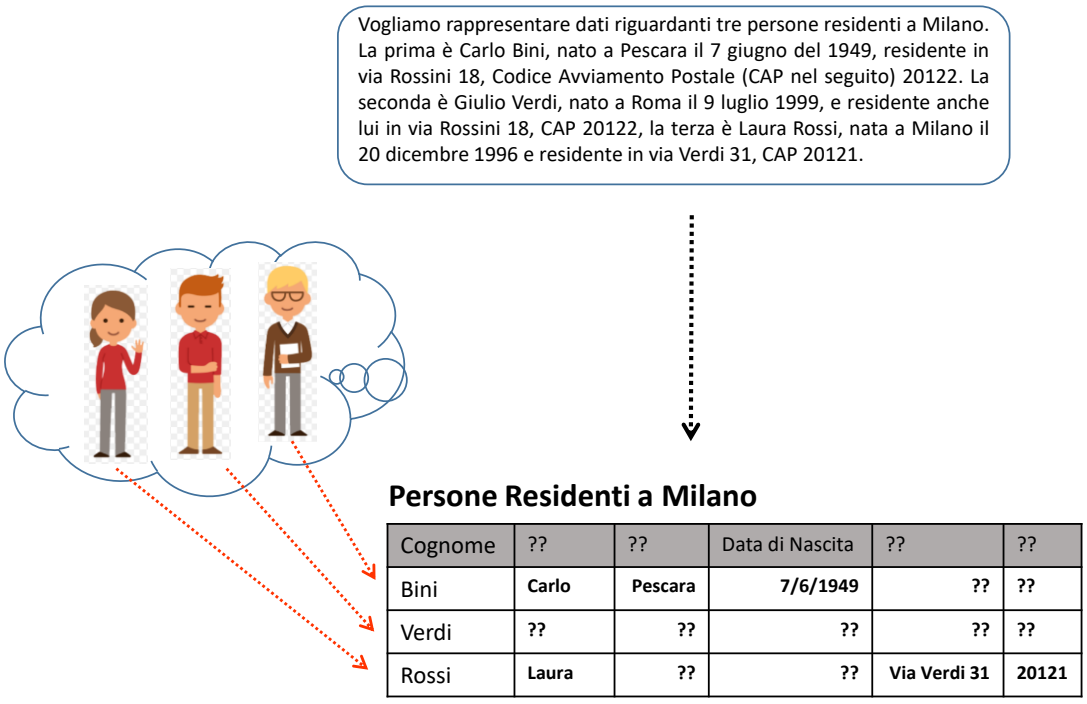


Figura 4 – Come associare ai dati una struttura

La risposta a queste domande è un primo esempio di utilizzo dei dati; per rispondere, è necessario leggere e talvolta rileggere il testo, e ogni volta individuare i dati da utilizzare per rispondere alla domanda. In questo siamo facilitati dal fatto che I dati nel testo sono descritti in forma molto ordinata, prima tutti quelli di Carlo Bini, poi tutti quelli di Giulia Verdi e infine quelli di Laura Rossi.

Per esempio, per rispondere alla prima domanda dobbiamo individuare nel testo i vari codici di avviamento postale e, quando troviamo il valore "20127" tornare un pò indietro e individuare l'indirizzo. Il problema che abbiamo nel cercare i dati in un testo è che quanto più il testo è lungo, tanto più diventa faticoso estrarre i dati utili alla risposta. Gli informatici dicono che lo sforzo per rispondere alla domanda non è *scalabile*, insomma, rispondere diventa via via più complicato all'aumentare della lunghezza del testo.

Proviamo allora prima ad affrontare il problema posto in Figura 4, cioè a organizzare i dati nel testo nell'ambito della tabella, costituita da quattro righe e sei colonne. Per facilitarvi il compito, ho inserito alcuni dati in un certo numero di celle della tabella. Attenzione che la prima riga è in grigio mentre le altre sono bianche, comprenderemo tra poco la ragione di questa diversità della prima riga rispetto alle altre.

Concentriamoci sulla prima colonna. Cosa notate? Nella prima colonna sono riportati i dati

- Cognome
- Bini
- Verdi
- Rossi

Se ragioniamo sui quattro dati, arriviamo facilmente alla conclusione che c'è una fondamentale differenza tra Cognome e tutti gli altri dati: Bini, Verdi e Rossi sono tutti *dati elementari* (che cioè non possiamo scomporre in dati più semplici) che costituiscono esempi di cognomi, mentre Cognome è una classe di dati, di cui e Bini, Verdi e Rossi sono elementi, elementi che chiameremo nel seguito ***valori*** della classe. La classe di valori Cognome è una *proprietà* delle persone rappresentate nella Tabella *Persone residenti a Milano*, proprietà che chiameremo ***attributo*** della tabella.
Ora è più chiaro il ruolo della prima riga grigia nella Figura 4, essa rappresenta le classi di valori, mentre le righe successive rappresentano gruppi di valori. Le classi rappresentate sono due, Cognome e Data di Nascita, e i valori rappresentati sono nove.

Facciamo ora un esercizio. Tenendo presente la Figura 4, guardate la Figura 5 e concentratevi sul testo più in basso. L'esercizio consiste nel sottolineare le parti elementari del testo che corrispondono a valori e circondare con una cornice le parti del testo che corrispondono secondo voi a classi.

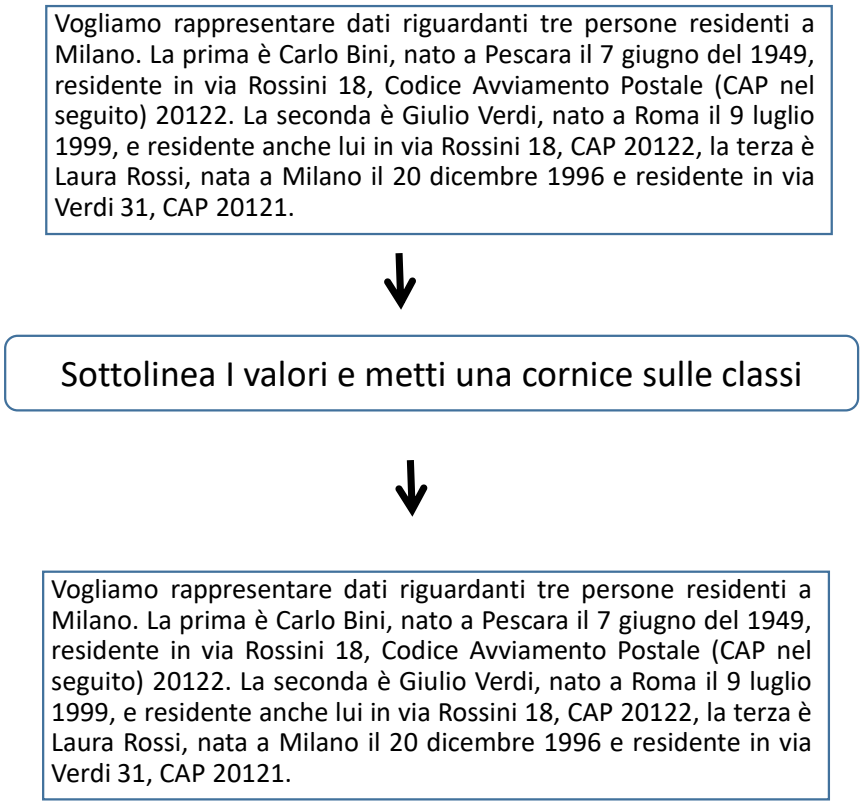


Figura 5 – Dati come valori e insiemi di valori

Nella prossima pagina trovate la soluzione.

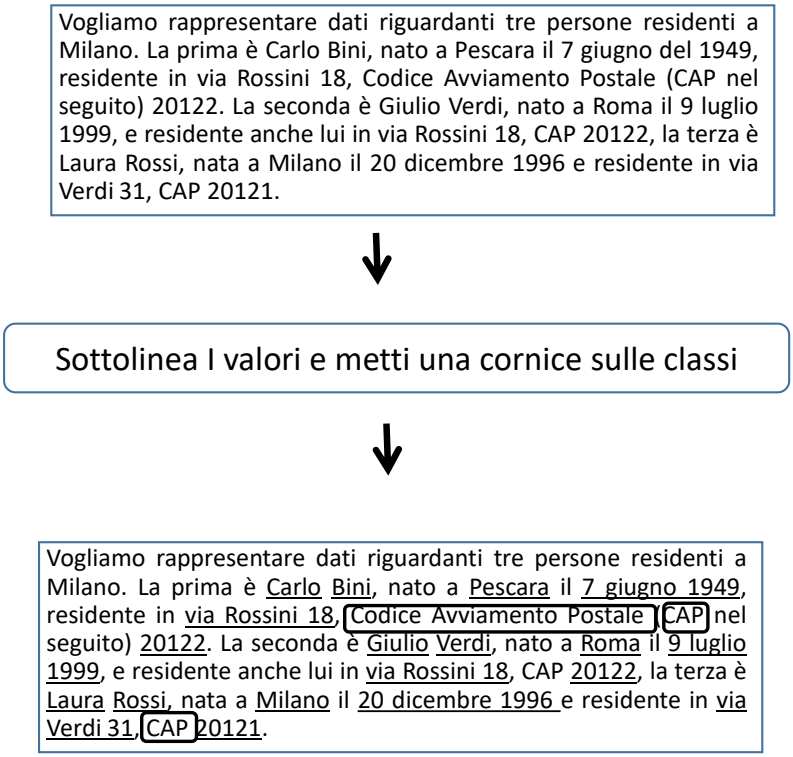


Figura 6 – Soluzione dell'esercizio

Nella parte inferiore di Figura 6 abbiamo sottolineato 18 valori elementari (ad esempio "Carlo" e "Bini" sono due valori elementari) e una sola classe, che ha due nomi diversi, "Codice Avviamento Postale" e "CAP".

Ricapitoliamo. Nella tabella di Figura 4 ci sono due classi (o attributi) Cognome e Data di Nascita, e nove valori, e nella Figura 6 abbiamo individuato una classe, CAP, e 18 valori elementari. E' chiaro fin qui?

Sì, tutto chiaro, ma dove vuoi arrivare?

Ancora un attimo di riflessione, poi spero che sarà tutto chiaro. Per capire ancora meglio la differenza tra valori e classi, guarda la Figura 7.

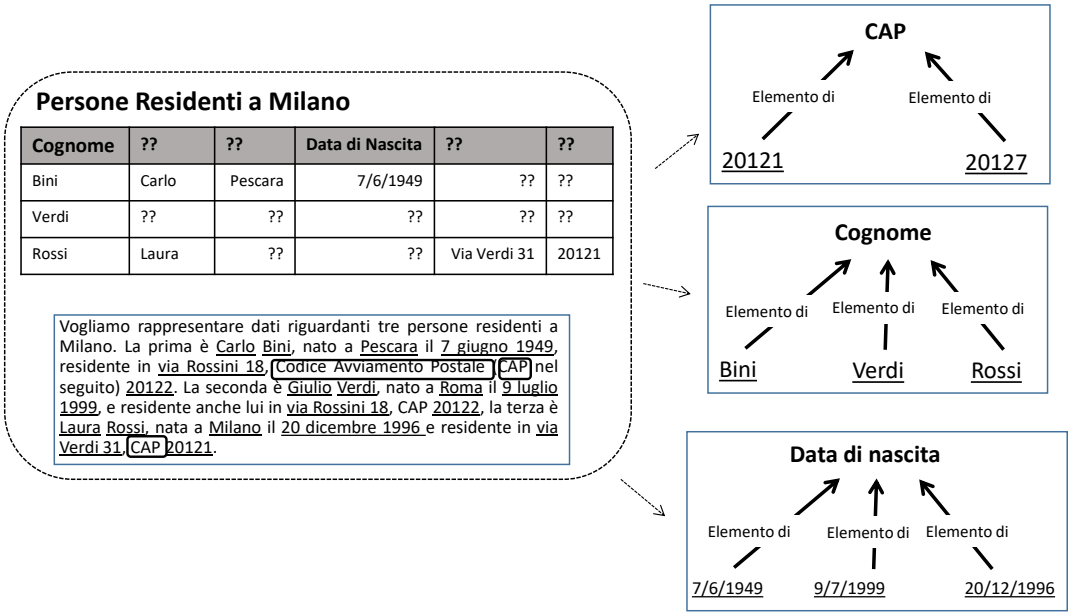


Figura 7 – Partendo dalle tre classi Cognome, Data di Nascita e CAP
ricostruzione della relazione tra classi e valori

Nella figura le due classi Cognome e Data di Nascita della Tabella "Persone residenti a Milano" e la classe CAP di Figura 6 sono messe in relazione con i loro valori, che risultano essere elementi delle classi. Per essere completi, a rischio di essere pignolo, nota che riguardo ai valori di Data di Nascita ho effettuato una trasformazione tra il modo in cui la data è rappresentata nel testo (es. 7 giugno 1949) e come è rappresentata nella tabella (7/6/1949).

Lettore – Scusa se interrompo le tue spiegazioni, ma ho una domanda. Invece di rappresentare Data di Nascita con una sola colonna, che tu chiami attributo, potevamo rappresentarla mediante tre attributi, Giorno di Nascita, Mese di Nascita e Anno di Nascita?

Certo, in questo caso però avremmo avuto bisogno di tre colonne invece che una.

Quindi ci sono diversi modi di rappresentare lo stesso testo in linguaggio naturale mediante dati strutturati?

E' così, e questo, forse lo stai iniziando a percepire, rende la modellazione una attività creativa e non automatica o meccanica.

Provate ora a completare la tabella con i restanti valori e attributi nella tabella di Figura 6. Quando avete finito, potete passare alla prossima pagina per una soluzione.

Ecco la soluzione.

Vogliamo rappresentare dati riguardanti tre persone residenti a Milano. La prima è Carlo Bini, nato a Pescara il 7 giugno del 1949, Residente in via Rossini 18, Codice Avviamento Postale (CAP nel seguito) 20122. La seconda è Giulio Verdi, nato a Roma il 9 luglio 1999, e residente anche lui in via Rossini 18, CAP 20122, la terza è Laura Rossi, nata a Milano il 20 dicembre 1996 e residente in via Verdi 31, CAP 20121.

Persone Residenti a Milano

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 8 – La tabella completa

La Tabella completa di tutti I dati e gli attributi è mostrata in Figura 8.

Il modello relazionale dei dati

L'insieme delle due strutture costituite da
- tabelle e
- attributi

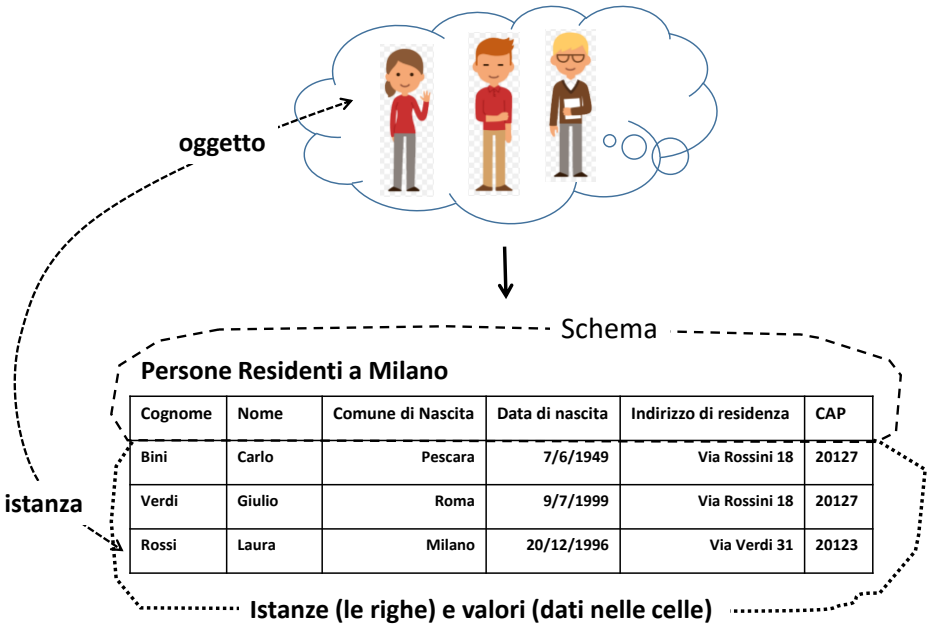
è ciò che in informatica viene chiamato ***modello relazionale dei dati***. Il modello relazionale ci dice che qualunque insieme di dati può essere rappresentato per mezzo di un insieme di tabelle, ciascuna delle quali è definite in termini di un insieme di attributi. In Figura 8, la tabella ha nome "Persone Residenti a Milano" e gli attributi, in numero di sei, sono Cognome, Nome, Comune di Nascita, Data di Nascita, Indirizzo di Residenza e CAP.

Ricapitoliamo quanto detto finora, e introduciamo un po' di terminologia.

Guardate la Figura 9. La tabella nella parte inferiore è una struttura di rappresentazione tipica del modello relazionale con cui noi rappresentiamo un frammento del mondo reale, alcune caratteristiche di tre persone che vivono a Milano. Le tre persone sono esempi di oggetti del mondo reale; ognuna e rappresentata per mezzo di una riga che chiameremo ***istanza*** (le istanze nel modello sono dunque rappresentazioni di oggetti del mondo reale). Ogni singolo dato che compare in una istanza è chiamato ***valore***, o anche valore elementare perché non è ulteriormente scomponibile in valori più semplici.

Il nome della tabella (nel nostro caso "Persone residenti a Milano") più l'insieme dei nomi degli attributi sono anche chiamati ***schema di dati***. La tabella rappresenta dunque un insieme o classe di istanze, mentre i singoli valori degli attributi rappresentano le caratteristiche reali

delle persone, il nome Carlo, il Cognome Bini, ecc. Notate che nel tempo accadrà in generale che nuove persone e quindi nuove righe verranno aggiunte alla tabella, come accade per esempio nella anagrafe di un comune, mentre lo schema resterà in genere più stabile nel tempo.



Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 9 – Schema e valori in una tabella

In Figura 10 riporto una breve sintesi dei concetti introdotti fin qui per i modelli dei dati e per il modello relazionale in particolare.

Concetto	Significato
Modello dei dati	Insieme di strutture per la rappresentazione di un insieme di oggetti del mondo
Modello relazionale	Insieme di due strutture di rappresentazione, la tabella e l'attributo
Tabella	Classe di istanze associate a singoli oggetti del mondo
Attributo	Classe di valori, che costituiscono proprietà delle istanze
Istanza	Un oggetto del mondo reale rappresentato in una riga di una tabella
Valore	Un dato elementare di una istanza che rappresenta una sua proprietà o attributo

Figura 10 – I concetti introdotti finora sui modelli dei dati

Il modello relazionale è il modello più utilizzato nei sistemi informativi in tutte le organizzazioni nel mondo, ma non è, come vedremo, l'unico modello possibile per organizzare un insieme dei dati in una struttura a tabella; il modello relazionale si è imposto nel tempo per la sua grande semplicità.

Ma.. adesso mi puoi dire cosa c'entrano le parole crociate con lo schema e i valori di una tabella?

Eh, c'entrano un pò indirettamente, nel senso che nelle parole crociate lo schema è costituito da un certo numero di celle nere che ci dicono quando cominciano le parole in verticale e in orizzontale e quindi ci dicono anche le lunghezze di queste parole; insomma, lo schema ci fornisce alcune proprietà delle parole che dobbiamo scoprire. Questo accade anche in una schema nel modello relazionale, dove nella prima riga delle tabelle noi riportiamo i nomi degli attributi, che esprimono le proprietà comuni dei dati nella rispettiva colonna.

Estrarre informazioni dai dati

Sono un pò deluso: fino ad ora negli esempi che hai fatto, I dati sono una cosa inanimata, "fredda". Io vorrei lavorarci con I dati, estrarre ciò che mi serve, insomma, vederli come una cosa viva! Mi fai vedere come posso fare a utilizzare I dati per una mia esigenza?

Certo, inizio subito, osservando che organizzare un insieme di dati per mezzo di una tabella ci permette, ad esempio, di rispondere in modo più efficiente alle tre domande che abbiamo visto in precedenza, e che in informatica prendono il nome di *interrogazioni*. Te le ripropongo nella prossima cornice.

1. Dove sono nate le persone (citate nel messaggio) che vivono in un indirizzo che ha CAP 20127?
2. Quando è nato Giulio Verdi?
3. Quante persone sono nate dopo il 1950?

Anzitutto, prova a rispondere alle tre domande usando gli occhi per muoverti tra I dati e la mente per fare i calcoli o memorizzare I risultati. Prova con la prima domanda e poi guarda a pagina successiva.

Bene, dovresti aver mosso gli occhi come nella prossima Figura 11.

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 11. Come si muovono gli occhi nell'eseguire una interrogazione

Dapprima dovresti aver guardato il primo CAP, e siccome è quello fornito nella domanda, dovresti aver mosso lo sguardo sulla città di nascita. La stessa cosa per il secondo CAP, mentre per quanto riguarda il terzo, siccome non corrisponde a quello cercato ed è l'ultimo, hai terminato l'esame della tabella.

E' così…..

Bene, ti faccio osservare che consapevolmente o meno hai usato una sequenza di passi elementari nell'esaminare la tabella e rispondere alla domanda, che nella figura precedente sono espressi con linee e frecce. Prova ora a esprimerli in linguaggio naturale cercando di generalizzare le scelte per una tabella non di tre righe, ma di un numero imprecisato di righe, diciamo 100. Dovresti aver prodotto un testo come il seguente in Figura 12.

1. Leggi il CAP della prima riga
2. Se è 20127 allora leggi la città di nascita e forniscila in risposta, altrimenti non fare niente.
3. Spostati sulla riga successiva e esegui:
   Se il CAP è 20127, allora leggi la città di nascita e forniscila in risposta, altrimenti non fare niente.
   Quando sei arrivato alla riga 100, termina, altrimenti torna al passo 3.

Figura 12 – Un primo linguaggio per esprimere interrogazioni

Questa sequenza di passi elementari, o un'altra simile che puoi aver prodotto, è un *algoritmo*, un metodo risolutivo che esprime operativamente le operazioni elementari che dobbiamo eseguire per, in questo caso, rispondere a una interrogazione. Non preoccuparti se il tuo testo differisce da quello proposto sopra, è importante che ogni frase esprima una azione chiara e *non ambigua*.

Mi fai un esempio di descrizione non chiara e ambigua?

Eccolo.

1. Leggi il CAP di una generica riga
2. Se è 20127 allora leggi la città di nascita e forniscila in risposta, altrimenti non fare niente.
3. Spostati su un'altra riga a caso finchè non sei arrivato alla riga 100 e esegui:
   Se il CAP è 10127, allora leggi la città di nascita e forniscila in risposta, altrimenti non fare niente.
   Fine

Cosa significa "generica riga"? Cosa significa "a caso"? Cosa significa qui "finchè non sei arrivato alla riga 100"? E' tutto molto impreciso: no, questo non è un algoritmo!

Ora, la grande sfida è questa: se noi riusciamo a concepire un linguaggio "artificiale" che possa essere tradotto in linguaggio macchina da un altro programma (chiamato in informatica compilatore), noi possiamo scrivere la interrogazione in questo linguaggio, far tradurre la interrogazione dal compilatore, e far eseguire la interrogazione al calcolatore!

Il linguaggio SQL - cenni

Il linguaggio artificiale predisposto fin dagli anni 70' del secolo scorso per esprimere interrogazioni è l' **SQL** (Structured Query Lanugage, Linguaggio di Interrogazione Strutturato). Vediamo in Figura 13 come è possibile esprimere le interrogazioni precedenti in un linguaggio ispirato all'SQL, e che differisce dall'SQL perchè le istruzioni sono espresse in italiano invece che in inglese, e per l'uso di costrutti un pò più intuitivi di quelli utilizzati nel linguaggio più tecnico costituito dall'SQL.

1. Dove sono nate le persone (citate nel messaggio) che vivono in un indirizzo che ha CAP 20127?
TROVA Comune di Nascita
NELLA TABELLA Persone residenti a Milano
SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE CAP = "20127"

2. Quando è nato Giulio Verdi?
TROVA Data di Nascita
NELLA TABELLA Persone residenti a Milano
SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE NOME = "Giulio" **E** COGNOME = "Verdi"

3. Quante persone sono nate dopo il 1950?
TROVA Conta (Data di Nascita)
NELLA TABELLA Persone residenti a Milano
SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE: Data di Nascita **DOPO** 1/1/1950

Figura 13 - Come si esprimono le interrogazioni
in un linguaggio "comprensibile" a un calcolatore

Nota che in SQL l'espressione della interrogazione è un pò diversa che nel linguaggio di comandi usato in precedenza. Ad esempio nella interrogazione
1. Dove sono nate le persone (citate nel messaggio) che vivono in un indirizzo che ha CAP 20127?
TROVA Comune di Nascita
NELLA TABELLA Persone residenti a Milano
SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE CAP = "20127"

si afferma semplicemente che vanno cercate nella tabella *Persone residenti a Milano* tutte le righe che rispettano la condizione CAP = "20127"; per tutte queste righe, va individuato il Comune di Nascita.

Nel linguaggio usato in Figura 12 eravamo stati più precisi nell'indicare la sequenza di azioni da svolgere: iniziare dalla prima riga, spostarci sulla successiva, finchè non viene raggiunta l'ultima. Si suole dire che l'SQL è un ***linguaggio dichiarativo***: questo significa che nell'SQL io dichiaro semplicemente *cosa* voglio ottenere come risultato della interrogazione, non *come* va prodotto. Si dice anche che il linguaggio introdotto in Figura 12 è ***operazionale***, nel senso che esprime in maniera più precisa e diretta le azioni da compiere.

Che differenza c'è tra usare un linguaggio dichiarativo e un linguaggio operazionale, è preferibile l'uno o l'altro?

Diciamo che nei linguaggi dichiarativi fai meno fatica a dire cosa vuoi, perchè non devi entrare in tanti particolari. D'altra parte proprio perchè sono più sintetici e potenti, sono più difficili da imparare, è un pò come andare da Milano a Roma a piedi, con una automobile o con un aereo di cui siamo il pilota. A piedi è facile, ma ci si mette un tempo infinito, se usiamo un'auto o un aereo dobbiamo avere la patente, e guidare un'aereo richiede un lungo processo di apprendimento……

Vorrei a questo punto invitarvi a riflettere su una questione importante. Per la prima volta in questo libro tocchiamo con mano un aspetto che influenza profondamente la nostra vita e se ci pensiamo bene anche la nostra società: le istruzioni nel linguaggio SQL che abbiamo scritto vanno bene sia per tabelle con tre righe che per tabelle con un milione di righe. Certo, un calcolatore impiega un tempo diverso per eseguirle nei due casi, ma si tratta di poche centinaia di microsecondi, dove un microsecondo è un milionesimo di secondo. L'essere umano, al contrario, impiegherebbe per eseguire l'operazione un tempo che cresce moltissimo tra il primo e secondo caso.

Quindi, se vogliamo porre una domanda su una tabella, e far eseguire automaticamente l'algoritmo per trovare la risposta, basta generare una sola volta la interrogazione SQL, e da questo momento in poi non dobbiamo più perdere tempo per eseguire le azioni necessarie per rispondere alla domanda. Ciò pone due questioni molto rilevanti:

1. tutte le operazioni sui dati che possono essere espresse in termini di programmi eseguibili con un calcolatore digitale, trasferiscono lavoro dagli esseri umani alle macchine. Questa è una questione di capitale importanza, che sta profondamente influenzando la distribuzione del lavoro tra le persone e le tecnologie digitali.
2. Chi ci assicura che la interrogazione SQL esprime proprio la domanda che avevamo in mente? Potremmo esserci sbagliati nel progettarla! Questo tema, che viene chiamato della correttezza dei programmi l'ho affrontato in un altro libro, "Le basi dell'informatica", scritto oltre 35 anni fa e ormai esaurito, che però è scaricabile gratuitamente dal sito http://hdl.handle.net/10281/97703. Se siete curiosi, leggete il capitolo xx di quel libro.

Torniamo alle tabelle. Eh sì, le tabelle non le hanno inventate gli informatici; le tabelle sono utilizzate fin dall'inizio della civiltà per organizzare dati. Torniamo alla tabella introdotta nel Capitolo 4 del Primo libro della Enciclopedia, utilizzata dal medico John Snow per capire le ragioni della diffusione del colera nel 1845 a Londra, che riproduco in versione semplificata in Figura 14. Osservatela alla luce dei concetti che ho introdotto in questa sezione; secondo voi è ben fatta o ha dei difetti? Le risposte a pagina successiva.

TABLE IX.

	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company . . .	26,107	98	37
Rest of London	256,423	1,422	59

Figura 14 – La Tabella di Snow con le compagnie erogatrici di acqua e il numero di morti nei quartieri serviti

Anzitutto, il nome della tabella, “Table IX”, non è molto descrittivo del suo significato, è estremamente generico. Possiamo sostituirlo (in italiano) con il nome “Morti per colera suddivisi per Società che eroga l’acqua”. Secondariamente, l’attributo più importante, quello nella prima colonna, non ha un nome, e questo rende meno leggibile la tabella; possiamo dare all’attributo il nome “Società che eroga l’acqua”, lo stesso usato nel nome della tabella. Per il resto, c’è solo da capire meglio il significato di house, se cioè denoti un caseggiato o un appartamento, a occhio e croce sembra denotare un caseggiato. E infine, nel comprendere il significato dei valori, c’è da ricordare che spesso nei paesi anglosassoni, la virgola (,) denota quello che per noi è il separatore delle migliaia, il punto (.)

Il problema della identità

In questa sezione affrontiamo il problema della identità, che occupa le menti dei filosofi da millenni, filosofi che cercano di rispondere alla domanda: cosa esprime la mia identità, il fatto che io sia un essere unico e irripetibile, e distinguibile da tutti gli altri esseri umani?

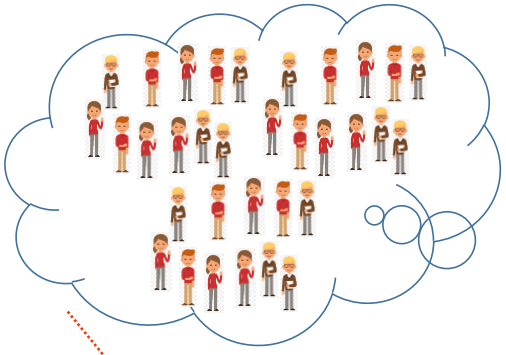
Da quando esistono le basi di dati il problema della identità è anche oggetto di interesse dei tecnologi dei dati. Partiamo da un esempio. Se tornate alla Figura 8 e vi chiedo di dirmi la data di nascita di Carlo Bini, senza esitazione mi rispondete 7/6/1949. Se riflettiamo un attimo, la ragione della pronta risposta è che c’è un solo Carlo Bini nella tabella. Ma se vi chiedo il nome e cognome della persona che abita in un indirizzo con CAP 20127, subito mi dite:

Ci sono due persone che abitano a un indirizzo con CAP uguale a 20127, quale delle due vuoi?

E hai ragione! Ricapitoliamo; la tabella rappresenta tre persone, a ciascuna delle quali corrisponde una riga; e caratteristiche delle tre persone rappresentate nella tabella sono il Nome, il Cognome, il Comune di Nascita, la Data di Nascita, l’Indirizzo di residenza, il CAP. Per distinguere una persona dall’altra, per associare a ogni persona una identità unica, non possiamo usare il CAP, mentre possiamo usare il Nome e Cognome, perchè per ogni valore di nome e cognome abbiamo una sola persona per cui quei valori sono definiti.

Ora passiamo a un esempio diverso, cioè gli studenti di un corso di laurea (per esempio Fisica o Scienza dei dati) della Università di Milano Bicocca, vedi Figura 15.

Figura 15 - Gli studenti di un corso di laurea universitario



Vogliamo rappresentare dati riguardanti gli studenti della Università di Milano Bicocca, di seguito i dati relativi a quattro di loro Il primo studente è Carlo Bini, nato a Pescara il 7 giugno del 1949, residente in via Rossini 18, CAP 20127. Il secondo è Giulio Verdi, nato a Roma il 9 luglio 1999, e residente anche lui in via Rossini 18, CAP 20122, la terza studentessa è Laura Rossi, nata a Milano il 20 dicembre 1996 e residente in via Verdi 31, CAP 20121. Il quarto studente è Carlo Bini, nato a Cantù il 3 luglio 1997, e residente in Via Duomo 13, CAP 20643

Studenti della Università di Milano Bicocca

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123
Bini	Carlo	Cantù	3/7/1997	Via Duomo 13	20138

Figura 16 - La tabella di partenza

La tabella in Figura 16 rappresenta una particolare categoria di persone, gli studenti (meglio, alcuni studenti) iscritti alla Università di Milano-Bicocca. Se vi chiedo di trovare la data di nascita di Carlo Bini nella tabella, in questo caso mi rispondete: quale Carlo Bini? Ce ne sono due! E io devo dire, ad esempio: quello che è nato a Pesaro, e se ce ne fossero due nati a Pesaro, devo aggiungere ulteriori valori di attributi, fino a poter identificare il Carlo Bini che mi interessa.

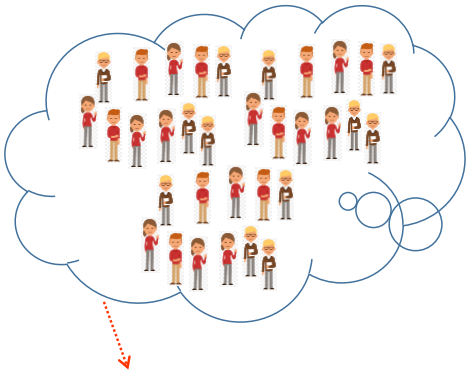
Perchè questo accade? Perché gli studenti della Università di Milano Bicocca sono molte migliaia, e quindi è frequente che ci siano più studenti con lo stesso nome e cognome.

E’ per questa ragione che nelle Università si assegna un *numero di matricola* agli studenti, per identificarli all’interno dei servizi offerti dalla Università.

C’è un secondo aspetto che va approfondito, con riferimento agli attributi e ai relativi valori. Provate a riflettere su questo punto: nella tabella precedente venivano rappresentati cittadini residenti a Milano, per cui era noto il comune di residenza; in questa tabella rappresentiamo studenti di una Università, che possono risiedere in molti comuni diversi. Vi dice qualcosa questa osservazione? Dobbiamo forse intervenire sulla tabella, aggiungendo un altro attributo?

Ragionate dunque su queste due questioni (matricola e residenza) e provate a produrre un nuovo schema di tabella.

In Figura 17 vediamo un nuovo schema, estensione del precedente di Figura 16, in cui sono stati



Vogliamo rappresentare dati riguardanti gli studenti della Università di Milano Bicocca, di seguito i dati relativi a quattro di loro Il primo studente è Carlo Bini, matricola 18167, nato a Pescara il 7 giugno del 1949, residente a Milano in via Rossini 18, CAP 20127. Il secondo è Giulio Verdi, matricola 21342, nato a Roma il 9 luglio 1999, e residente anche lui a Milano in via Rossini 18, CAP 20122, la terza studentessa è Laura Rossi, matricola 231\33, nata a Milano il 20 dicembre 1996 e residente a Milano in via Verdi 31, CAP 20121. Il quarto studente è Carlo Bini, matricola 12145, nato a Cantù il 3 luglio 1997, e residente a Seregno in Via Duomo 13, CAP 20643

Studenti della Università di Milano Bicocca

Matricola	Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	Comune di residenza	CAP
18167	Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	Milano	20127
21342	Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	Milano	20127
23133	Rossi	Laura	Milano	20/12/1996	Via Verdi 31	Milano	20123
12145	Bini	Carlo	Cantù	3/7/1997	Via Duomo 13	Cantù	20643

Figura 17 - La tabella arricchita con nuovi attributi e relativi valori

aggiunti gli attributi Matricola e Comune di residenza, e i relativi valori. Matricola è stato aggiunto per poter assegnare una precisa identità a ogni studente, e quindi una precisa riga nella tabella, e Comune di residenza perché ora l'indirizzo non è sufficiente per specificare in modo non ambiguo dove ogni studente è residente. L'attributo Matricola è anche chiamato ***chiave*** della tabella.

Perché viene usato questo strano nome: "chiave"?

Ottima domanda, il bello di poter dialogare con gli studenti a lezione e con i lettori un po' più virtualmente è che dopo tanto tempo che lavoro sui dati, ancora ci sono domande che non mi ero mai posto, come questa! La risposta è che non lo so. Pensandoci, la parola chiave viene usata per indicare il fatto che il valore di quell'attributo, in questo caso Matricola, è una chiave per entrare nella tabella e trovare ciò che ci serve. Se non sei soddisfatto e sei curioso su questi argomenti, prova a cercare nel Web altre informazioni su questa questione.

Va bene, ma effettivamente non mi sembra una questione importantissima, la prendo come una convenzione…

Una tabella o tante tabelle?
Come scegliere tra diverse rappresentazioni degli stessi dati

Senti ma non mi vorrai far credere che un frammento di mondo, come lo chiami tu, si possa sempre rappresentare con una sola tabella!?

Infatti, in generale una applicazione informatica come, ad esempio, rappresentare nel 2020 tutti i tipi di dati utili per analizzare e comprendere l'evoluzione della epidemia del virus Covid, richiede molte tabelle diverse, che rappresentano, per esempio, nel caso del Covid, i positivi nel tempo nel mondo, in Italia, in ciascuna delle venti regioni italiane, I guariti, ecc. ecc.

Poi, se pensiamo ai dati che utilizziamo nella nostra vita, noi siamo abituati a memorizzare i dati che ci servono su diversi supporti: i numeri di telefono, che ormai molti di noi conservano in un archivio digitale sul telefono cellulare, gli appuntamenti, che alcuni di noi conservano in una agenda cartacea, ed altri sul cellulare, e così via. Tutti questi insiemi di dati sono conservati in tabelle, alcune cartacee, altri digitali. Insomma, in modo naturale siamo abituati a conservare i dati in diverse tabelle, e, hai ragione tu, l'esempio della precedente sezione, con una sola tabella, è troppo semplicistico.

C'è una regola, un metodo, che possiamo applicare per decidere se rappresentare dei dati con una, due, tre o tante tabelle?

Dunque, nel risponderti ti ricordo il concetto di *base di dati*. Una base di dati è un insieme di tabelle che considerate globalmente rappresentano un frammento di mondo (ad es. i positivi Covid per le diverse regioni italiane, gli stipendi degli impiegati di una azienda, i cittadini residenti in un comune, ecc.).

Vediamo ora un esempio di frammento di mondo che può essere rappresentato mediante (almeno) due possibili basi di dati, una composta di una tabella, e un'altra composta di tre tabelle, vedi la Figura 18.

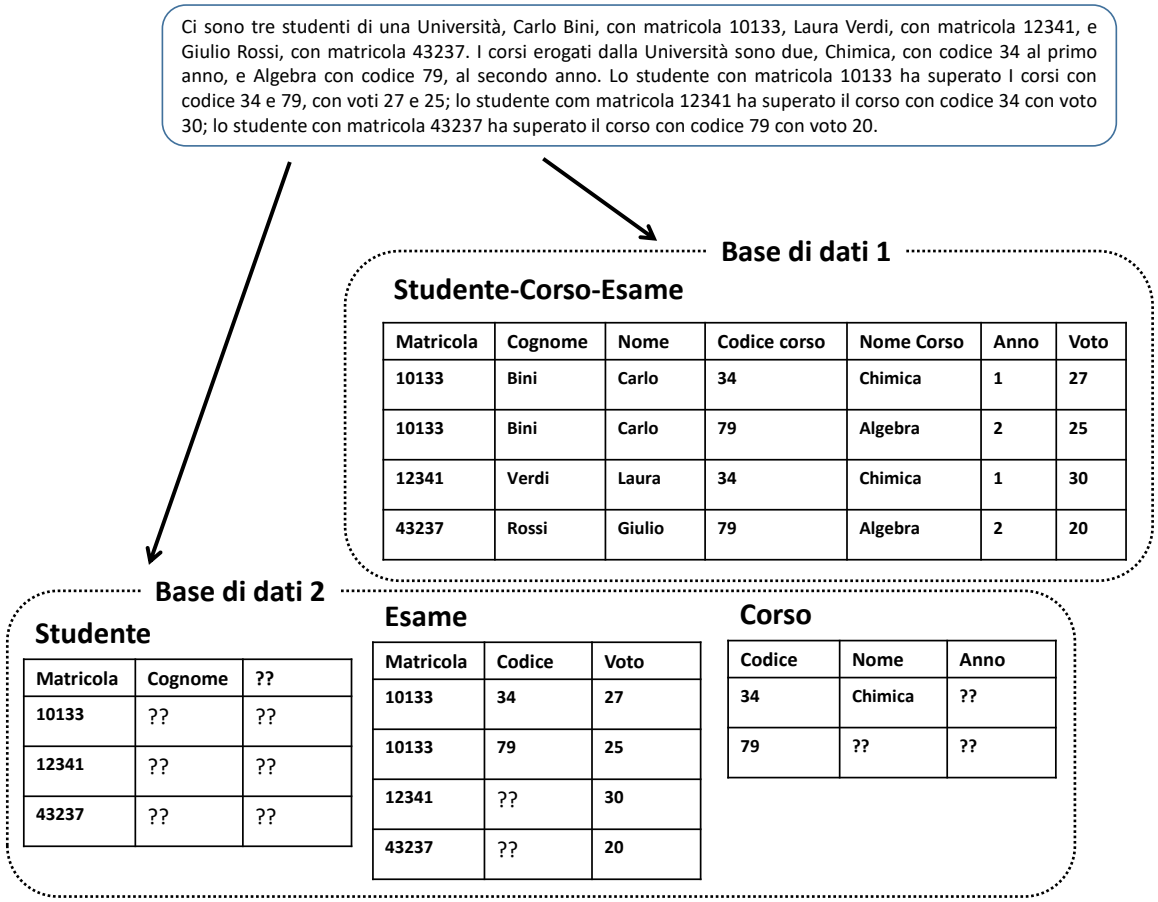


Figura 18 – Una tabella o tre tabelle?

Come ho fatto in precedenza, mentre la base di dati composta di un'unica tabella ha tutti i valori specificati, nel caso della base di dati composta da tre tabelle ho specificato tutti gli attributi delle tre tabelle, e solo parte dei valori. Provate ora a completare la base di dati composta dalle tre tabelle.

In Figura 19 trovate la soluzione. In questo caso dico “la soluzione” perché avendo io deciso la ripartizione degli attributi nelle tre tabelle, e avendo specificato alcuni valori, c’è un solo modo di completare le tabelle con i valori restanti.

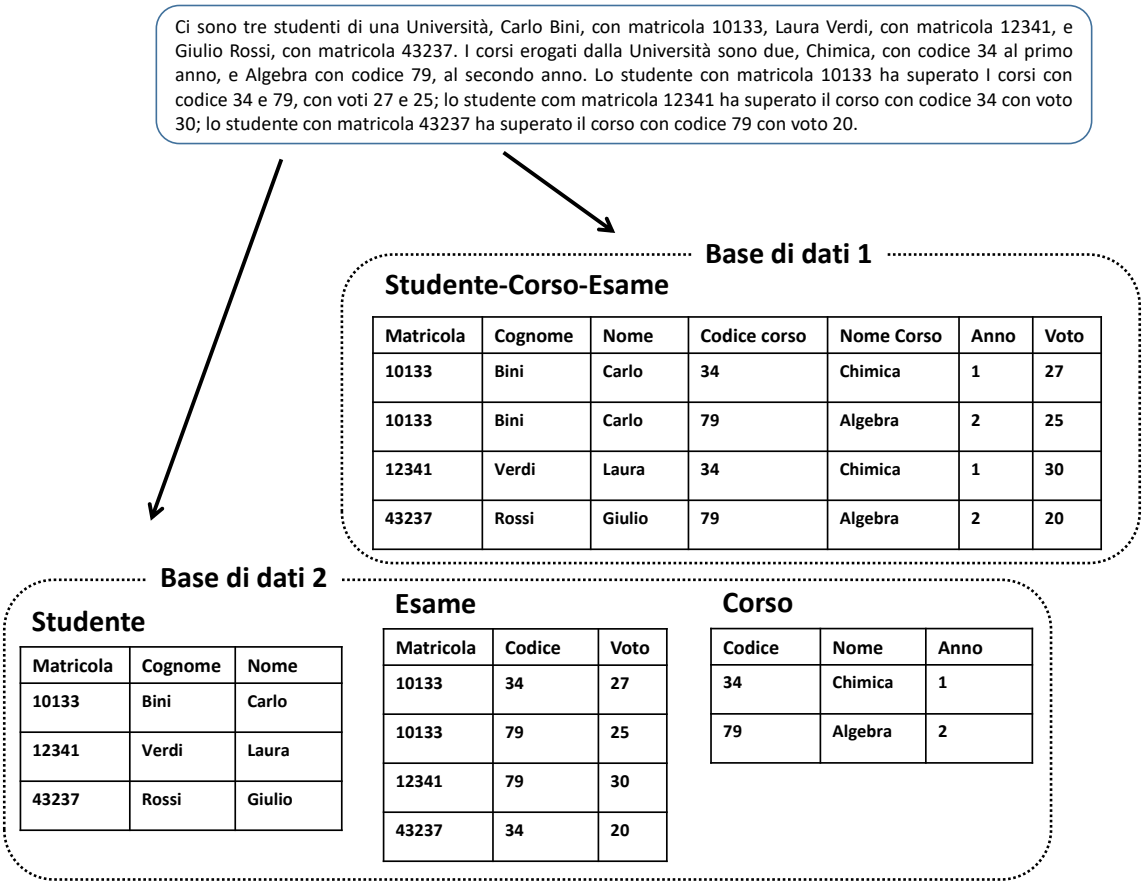


Figura 19 – Soluzione

Non andare troppo veloce. Quello che non capisco è perché hai rappresentato la Matricola dello Studente e il Codice del Corso due volte: nelle tabelle Studente e Corso, e fin qui va bene, ma anche nella Tabella Esame. Che bisogno c’era di rappresentare Matricola e Codice nella tabella Esame?

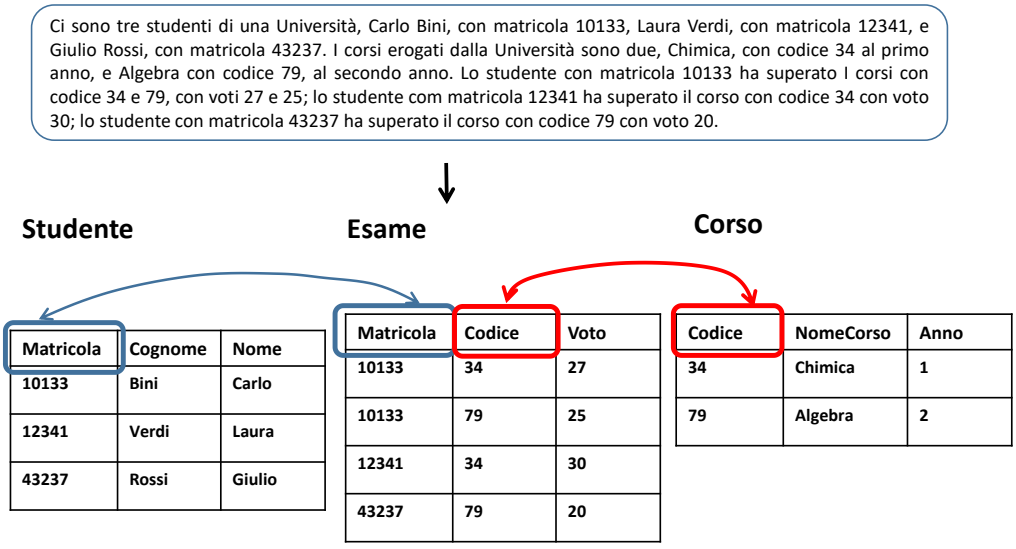


Figura 20 – Come sono collegati i dati tra di loro
quando sono rappresentati in diverse tabelle

Ottima domanda. Guarda la Figura 20 e prova mentalmente a trovare i voti ottenuti dallo studente con matricola “43237” negli esami che ha sostenuto. E’ chiaro che rispondi “20”, perché sei riuscito a collegare i dati della prima tabella con quelli della seconda facenti riferimento allo studente con matricola “43237”. Insomma soltanto rappresentando anche in Esame la matricola degli studenti io riesco a collegare logicamente le informazioni che legano gli studenti agli esami.

Ho capito. La stessa cosa vale per Corso ed Esame?

Certo, assolutamente la stessa cosa. Un esame è in sostanza definito, si dice anche in questo caso: identificato, dallo studente che lo svolge e dal corso a cui fa riferimento. Quindi, per identificarlo, per dire “questo è l’esame superato dallo studente con matricola 43237 per il corso con codice 79, devo ripetere “43237” e “79” nella tabella esame, e quindi devo far comparire nello schema della tabella gli attributi Matricola e Codice.

Abbiamo scoperto un aspetto importante dei dati digitali, la loro grande versatilità nella rappresentazione di un frammento di mondo. Volendo approfondire, si potrebbero rappresentare gli studenti, gli esami e i corsi anche con due tabelle, oppure con un numero maggiore di tre, separando ad esempio il cognome dal nome.

D’accordo, ma a me la rappresentazione con tre tabelle sembra molto chiara, ogni tabella ha un nome sintetico che esprime direttamente il significato dei dati nella tabella.

Hai perfettamente ragione, la rappresentazione con tre tabelle è molto più chiara di quella con una sola tabella, è una riprova di questo fatto è nel nome che ho dovuto dare alla tabella in Figura 19, un nome composto che non esiste nel nostro linguaggio comune. Quel nome composto è la riprova che nell’unica tabella abbiamo mescolato tre diversi concetti, Studente, Esame e Corso, che invece sono visti separatamente nelle tre tabelle.

Certo, però abbiamo dovuto pagare un prezzo nella rappresentazione con tre tabelle, perché abbiamo dovuto rappresentare due volte Matricola e due volte Codice…..

E' vero, abbiamo dovuto pagare un prezzo che porta ad un maggiore sforzo cognitivo, perché ora per collegare dati nelle tre tabelle dobbiamo fare uno sforzo in più nel rimetterli insieme. Come vedremo nella prossima sezione.

Ci sono tanti punti di vista nel mondo….

Dunque, uno stesso frammento di mondo si può descrivere in tanti modi, mediante tabelle relazionali. La Figura 21 mostra addirittura quattro modi diversi di rappresentare il testo che stiamo considerando, con una, due, tre e quattro tabelle! E per quanto riguarda la rappresentazione con due tabelle, è chiaro che accanto a quella che compare in Figura 18 possiamo prendere in considerazione la simmetrica, costituita dalle due tabelle:

1. Studente, definita sugli attributi Matricola, Cognome, Nome
2. Esame-Corso, definita sugli attributi Matricola, Codice Corso, Nome, Anno, Voto.

Se confrontiamo la soluzione con tre tabelle con tutte le altre, vediamo che solo in quella con tre tabelle riusciamo ad usare nomi che corrispondono ad aspetti del mondo reale, e questo è un importante segno che questa soluzione è la più chiara e comprensibile tra tutte. Per esempio, la soluzione in cui dividiamo in due la tabella Studente, collocando nella prima tabella Matricola e Nome e nella seconda Matricola e Cognome, ci appare un po' "forzata", che necessità c'è di dividere Nome e Cognome in due tabelle?

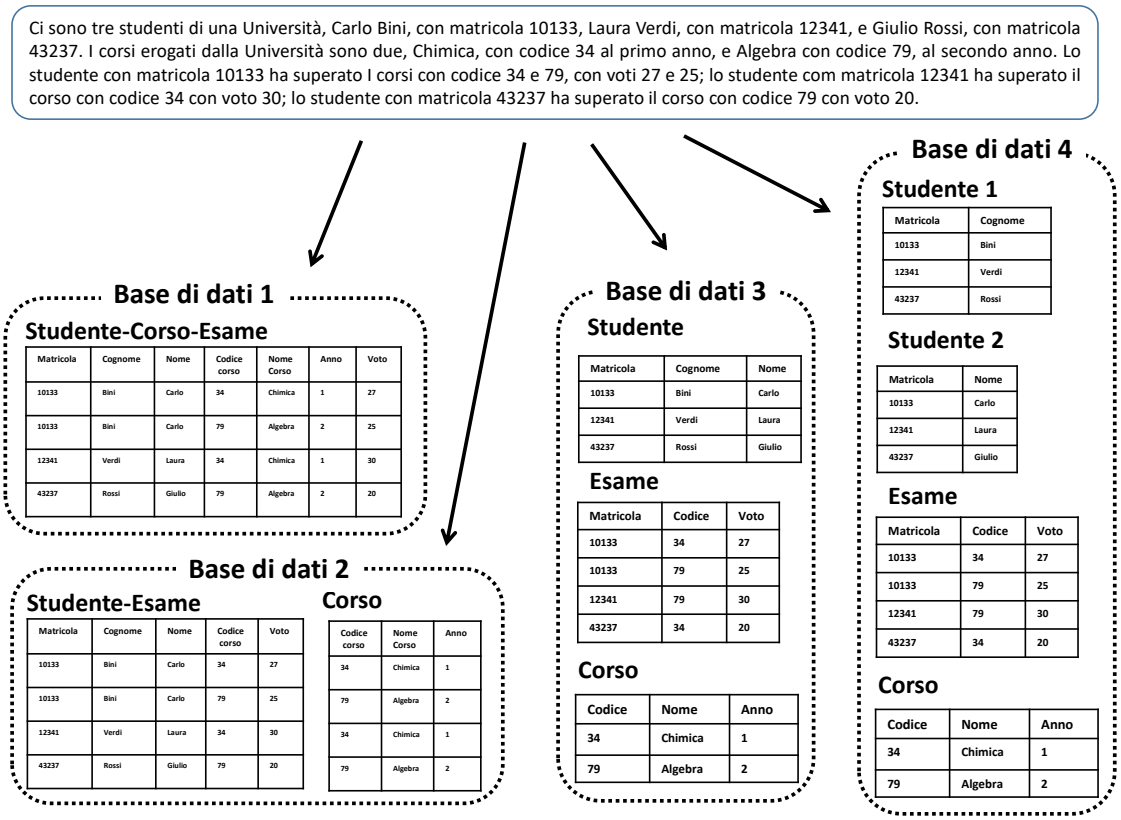


Figura 21 – Ci sono tanti modi diversi di rappresentare un frammento di mondo

Proprio l'ultimo esempio, quello delle due tabelle Studente1 e Studente2 In Figura 21, mi permette di introdurre un'altra questione della massima importanza quando si parla di dati e di come modellarli.

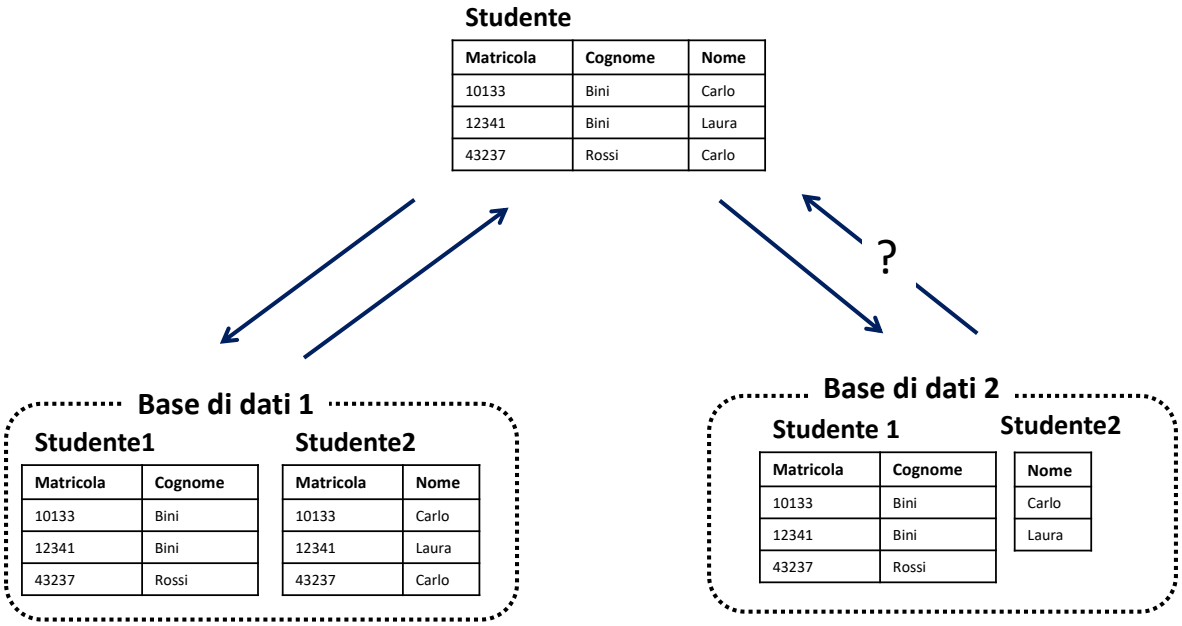


Figura 22 – Quando si può tornare sui propri passi?

Osservate la Figura 22. Nella Figura a sinistra noi decomponiamo due volte la tabella Studente, con due tabelle Studente1 e Studente2, in cui rappresentiamo Matricola in entrambi i casi e una volta Nome e l'altra Cognome; a destra rappresentando Matricola e Cognome in Studente1 il solo Nome nella tabella Studente2; notate in questo secondo caso che ho rappresentato il nome Carlo una volta sola, perché le nostre tabelle da un punto di vista matematico sono insiemi, e quindi non ha senso ripetere Carlo due volte (mentre ha senso ripetere Bini due volte perché associato a due matricole diverse). Poniamoci ora la seguente domanda: In quale o quali dei due casi partendo dalle coppie di tabelle è possibile ricostruire le righe della tabella iniziale? Tieni presente nel rispondere alla domanda che una volta ottenute le due tabelle, è come se noi non disponessimo più della tabella di partenza…..

Ci provo. Nel caso a sinistra, io ho rappresentato sia nella prima tabella che nella seconda l'attributo Matricola, quindi basta che collego le coppie di righe con lo stesso valore di Matricola, e ottengo le righe di partenza è corretto?

Ottimo! E nel secondo caso?

Eh, qui è più complicato, perché Matricola è scomparsa dalla seconda tabella, e i cognomi nella prima tabella e i nomi nella seconda, se non so la matricola, come li accoppio? Non so proprio come si possa fare…

E infatti non si può fare, perché i nomi non sono più associati alla matricola e quindi io non so a chi appartengono… Dunque soltanto in certi casi è possibile ricostruire i dati di partenza.

Questo è un aspetto importante dei dati che utilizziamo: essi sono sì una rappresentazione del mondo, ma attenzione a non frammentarli troppo, e a commettere così delle azioni irreparabili!

Mi potresti dare una regola generale da seguire per capire se posso ricostruire o no la tabella di partenza?

Si, la regola è che gli attributi in comune nella prima e seconda tabella devono essere la chiave o della prima o della seconda tabella, o di tutte e due. Questo accade nella decomposizione a sinistra, e non, ovviamente nella decomposizione a destra, in cui non ci sono attributi in comune. La proprietà ha un significato intuitivo; devo sempre essere in grado di identificare le righe, per poterle ricollegare.

E veniamo ora alla Figura 23. In questo caso noi assumiamo che la tabella SCE e le tre tabelle siano legate dalla solita decomposizione, ma in un contesto in cui

1. le due persone che hanno progettato la tabella non decomposta e le tre tabelle si siano presi delle libertà nello scegliere i nomi delle tabelle e degli attributi, e quindi siano venuti fuori queli nomi strani: SCE, Stud, Es, C;
2. i dati siano stati aggiunti nella tabella SCE e nelle tre tabelle indipendentemente, da persone diverse, che possono anche aver commesso degli errori di digitazione.

Beh, insomma, non credo che nelle applicazioni reali le tabelle e gli attributi siano chiamate con questi strani nomi! E credo anche che quando uno inserisca dei dati, stia attento!

Purtroppo non è così! Le organizzazioni pubbliche e private che usano le basi di dati hanno nei loro sistemi informativi centinaia, spesso migliaia di tabelle diverse, e i progettisti, talvolta per non perdere tempo, e talvolta per essere gli unici che capiscono il significato dei dati, tendono spesso a usare questi nomi “strani”; inoltre, ti sarà capitato anche a te di sbagliare tasto talvolta.
Quindi una carente progettazione e errori di digitazione, quando questa avvenga manualmente e non ci siano controlli, sono da mettere sempre in conto.

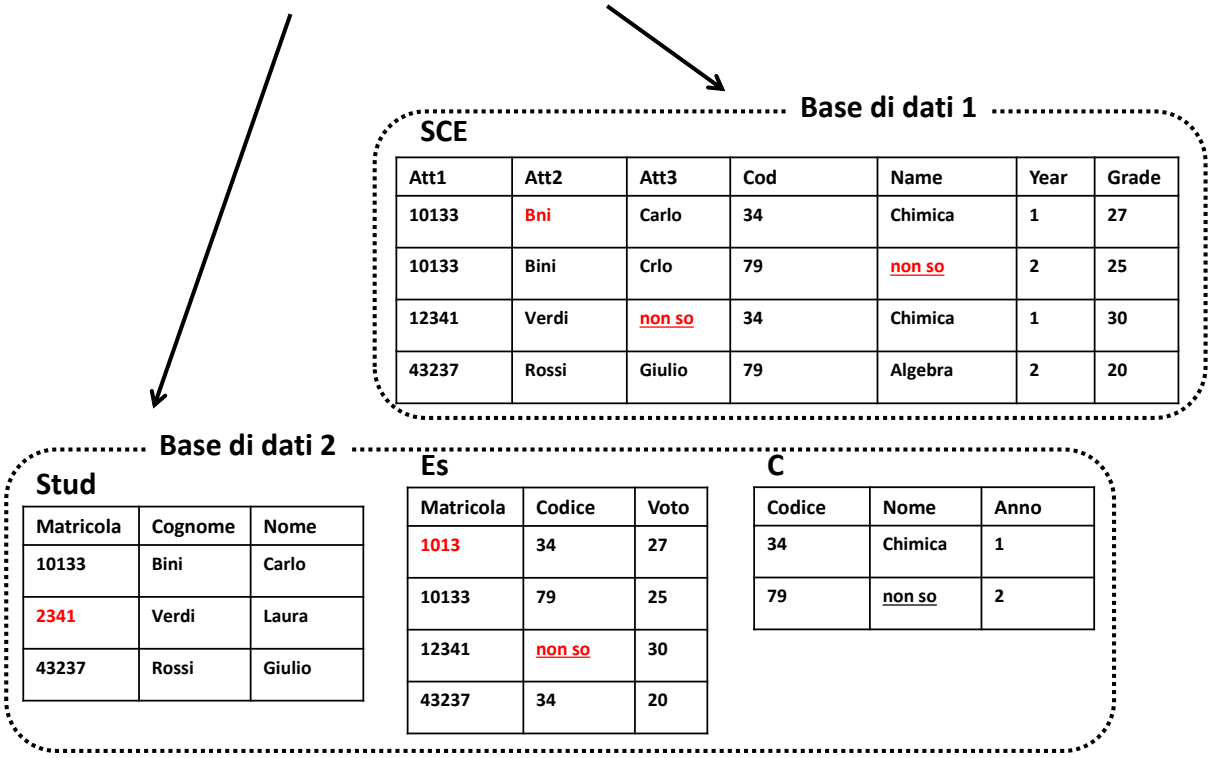


Figura 23 – Ci sono tanti modi diversi di rappresentare *male* un frammento di mondo….

Se infatti ora guardiamo con attenzione i valori dei dati nelle tabelle, facciamo delle brutte scoperte. Bini in un caso è rappresentato con un carattere in meno; inoltre ci sono due matricole anche esse con errori. In un certo senso questi errori su Matricola sono più gravi dell'errore su Bini, perché ci impediscono di ricollegare le righe relative.

La realtà in cui viviamo è piena di errori e di imperfezioni e queste imperfezioni vengono ereditate anche nei dati digitali che la realtà rappresentano; con il risultato che ci fanno vedere la realtà con una lente deformata, e che i dati vanno curati, aspetto questo che discuterò nel Capitolo 5.

Come facciamo a rimettere insieme i dati

In questa sezione parliamo delle operazioni che rimettono insieme dati presenti in diverse tabelle. Osservate le tre tabelle di Figura 24 e la interrogazione espressa a parole. Provate prima a capire, indipendentemente dagli aspetti di dettaglio, quali sono le tabelle coinvolte dalla esecuzione della interrogazione.

Trova I voti ottenuti agli esami dallo studente con Cognome Bini

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	NomeCorso	Anno
34	Chimica	1
79	Algebra	2

Figura 24 - Come eseguire una interrogazione su tre tabelle

Dopo aver deciso, guardate la Figura 25.

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	NomeCorso	Anno
34	Chimica	1
79	Algebra	2

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

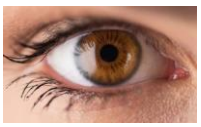
Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Figura 25 - Primo passo: *Quali* tabelle devo usare?

Vi torna? Spero di sì: noi stiamo cercando i voti, che sono contenuti nella tabella Esame; ci interessano i voti ottenuti da Bini, valore che è contenuto nella tabella Studente. Per collegare le due tabelle abbiamo bisogno della matricola di Bini.

Adesso cerchiamo di capire come si muovono gli occhi per poter rispondere alla interrogazione con i nostri sensi. Vedi la Figura 26.



Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Figura 26 - Secondo passo: *come* trovo i dati nelle due tabelle?
Per trovare i voti dobbiamo navigare

In pratica, prima troviamo Bini nella tabella Studente, poi la sua Matricola, a questo punto ci spostiamo sulla tabella Esame e per tutte le righe di Esame che contengono la matricola di Bini troviamo il voto.

Prima di proseguire, volevo capire una cosa. Questo metodo di spostare gli occhi è molto intuitivo. Perché voi informatici non avete inventato un linguaggio visivo, insomma un linguaggio che invece di esprimere le interrogazioni tramite comandi in linguaggio naturale, usa comandi visivi?

Accipicchia che domande sempre più interessanti che mi fai! Gli informatici hanno investigato linguaggi visivi, e alcuni sono stati trasfromati in prodotti. Ma bisogna dire che le interrogazioni in genere sono formulate da esperti, che preferiscono di gran lunga linguaggi programmativi, come quello che ho iniziato a usare nella sezione 5.

Adesso, come abbiamo fatto in precedenza, proviamo a descrivere l'interrogazione in un linguaggio programmativo. In questo caso dobbiamo collegare, "rimettere insieme" dati presenti in tabelle diverse.

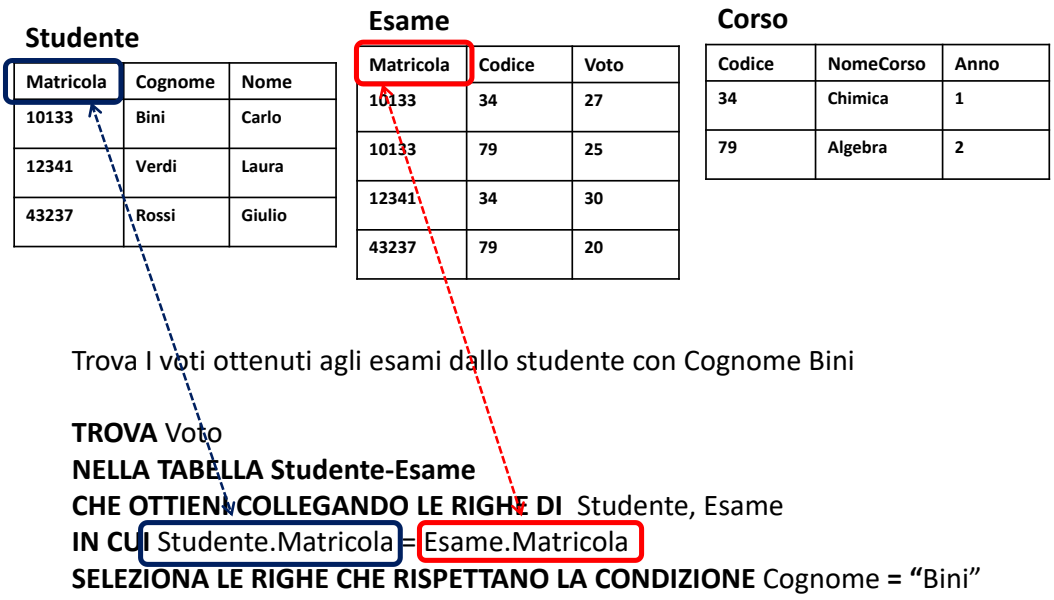


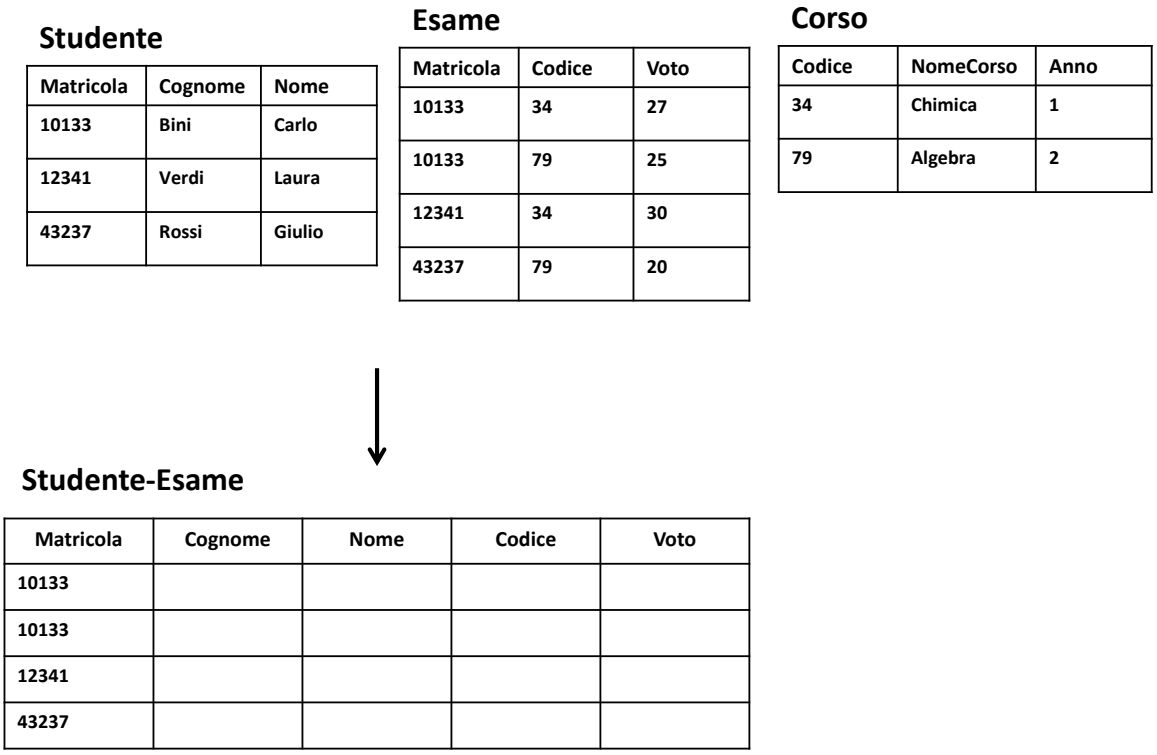
Figura 27 – L'istruzione COLLEGA

Questo si può fare con un nuovo comando, che chiamiamo COLLEGA, che ha proprio il compito di produrre una nuova tabella, chiamata nella interrogazione Studente-Esame, le cui righe sono ottenute concatenando righe delle tabelle Studente e Esame che coincidono nella Matricola. Vedi in Figura 27 la forma che assume la interrogazione.

Proviamo a capire insieme l'effetto della istruzione COLLEGA. In Figura 28 ho prodotto una prima versione della tabella Studente-Esame, che ora vi invito a completare. In pratica, vi ho già detto che la tabella ha tutti gli attributi della tabella Studente più quelli della tabella Esame.

Inoltre, siccome in ogni riga devo inserire una coppia di righe delle due tabelle che hanno lo stesso valore per Matricola, devo prima capire quante sono queste righe. La prima riga di Studente ha la stessa matricola delle prime due righe della tabella Esame; questo è un modo un po' contorto di dire che lo studente con Matricola "10133" ha superato due esami! La riga con matricola "12341" si accoppia con una riga di Esame, e la riga con matricola "43237" si accoppia con una riga della tabella Esame. Insomma, le righe di Studente-Esame sono quattro.

A questo punto, provate a completare la tabella Studente-Esame con tutti gli altri valori.



Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

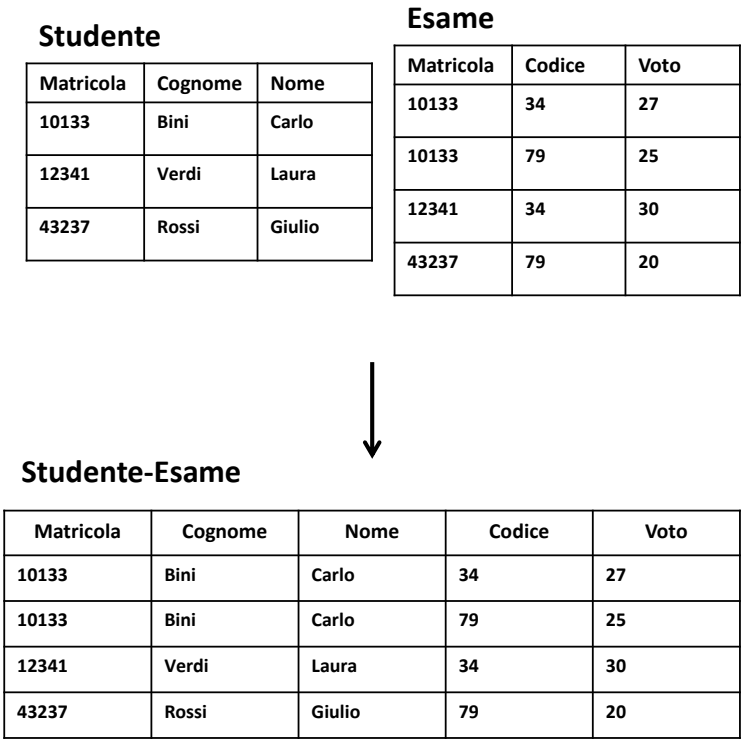
Codice	NomeCorso	Anno
34	Chimica	1
79	Algebra	2

Studente-Esame

Matricola	Cognome	Nome	Codice	Voto
10133				
10133				
12341				
43237				

Figura 28 – Un primo suggerimento per gli attributi, le righe e i valori della Tabella Studente-Esame

In Figura 29 compare la forma finale della tabella Studente-Esame. In ogni riga devo riportare tutti i valori degli attributi di Studente, e poi uno dei valori dei codici della tabella Esame, più il voto.



Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Studente-Esame

Matricola	Cognome	Nome	Codice	Voto
10133	Bini	Carlo	34	27
10133	Bini	Carlo	79	25
12341	Verdi	Laura	34	30
43237	Rossi	Giulio	79	20

Figura 29 – L'effetto della esecuzione della istruzione COLLEGA

A questo punto dobbiamo eseguire l'istruzione

SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE Cognome **= "**Bini"

Provate a costruire la nuova tabella.

Studente-Esame

Matricola	Cognome	Nome	Codice	Voto
10133	Bini	Carlo	34	27
10133	Bini	Carlo	79	25
12341	Verdi	Laura	34	30
43237	Rossi	Giulio	79	20

Studente-Esame

Matricola	Cognome	Nome	Codice	Voto
10133	Bini	Carlo	34	27
10133	Bini	Carlo	79	25

Figura 30 – La tabella risultato della selezione

In Figura 30 vediamo la tabella, che contiene le due sole righe riguardanti lo studente “Bini”.

A questo punto viene eseguita l’istruzione

TROVA Voto

Il cui risultato finale è la coppia di dati

27
25

Quindi, esprimendo la interrogazione nel tuo linguaggio, viene prima costruita una nuova tabella, e poi vengono esaminate le righe di questa tabella? Ma non è un po' complicato, non è più inefficiente del metodo che esplora le due tabelle una di seguito all’altra, come hai fatto con il linguaggio visivo?

Hai ragione, ma abbiamo già visto che quel metodo visivo corrisponde a un linguaggio *procedurale*, in cui ogni azione è esplicitata nel metodo risolutivo, mentre un linguaggio programmativo simile all’SQL che sto proponendo adotta un metodo risolutivo di tipo *dichiarativo*, dice in termini generali cosa si deve fare, lasciando scegliere le specifiche azioni o istruzioni a un altro programma che traduce la interrogazione nel linguaggio macchina.

Ora ridiventate voi i protagonisti. Provate a esprimere nel linguaggio che ho introdotto in precedenza la interrogazione di Figura 31.

Trova I Cognomi degli studenti che hanno sostenuto l'esame relative al corso con Codice 79

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	Nome	Anno
34	Chimica	1
79	Algebra	2

Figura 31 – Esercizio

Vi propongo di procedere in due passi. Nel primo passo concentratevi sui comandi TROVA, NELLA TABELLA e CHE OTTIENI COLLEGANDO LE RIGHE DI, e disinteressatevi dei comandi IN CUI e SELEZIONA LE RIGHE… Ti ricordo i ruoli dei tre comandi:

1. TROVA esprime l'attributo o gli attributi i cui valori sono richiesti dalla interrogazione.
2. NELLA TABELLA (o NELLE TABELLE) fornisce il nome della tabella (o tabelle) dove ritrovare il precedente attributo o attributi. In genere questa è una nuova tabella, ottenuta concatenando righe di tabelle presenti nella base di dati (vedi comando successivo).
3. CHE OTTIENI COLLEGANDO LE RIGHE DI individua la o le tabelle coinvolte dalla interrogazione.

Prova a esprimere la interrogazione nel linguaggio che usa questi tre comandi.

Nella Figura 32 la soluzione. Noi vogliamo trovare valori dell'attributo Cognome (TROVA), in una nuova tabella (NELLA TABELLA), che dichiariamo di costruire concatenando righe (CHE OTTIENI COLLEGANDO LE RIGHE DI) delle *tre* tabelle Studente, Esame, Corso. Eh sì, qui servono tutte e tre le tabelle, perché per collegare il Nome del corso (Chimica) con gli attributi i cui valori stiamo cercando (Cognome), occorre "passare per" la relazione Esame.

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	NomeCorso	Anno
34	Chimica	1
79	Algebra	2

Trova I Cognomi degli studenti che hanno sostenuto l'esame relativi al corso con Nome Chimica

TROVA Cognome
NELLA TABELLA Studente-Esame-Corso
CHE OTTIENI COLLEGANDO LE RIGHE DI Studente, Esame, Corso

Figura 32 – Una prima parte della interrogazione….

Provate ora a costruire la interrogazione completa. Ciò che resta da fare è specificare la condizione di concatenamento espressa dal comando IN CUI seguito dalla condizione di uguaglianza tra i valori di matricola nelle due tabelle Studente e Esame, e (comando E) tra i valori di Codice nelle tabelle Esame e Corso. Provate a scrivere la interrogazione e poi guardate la Figura 33.

Studente

Matricola	Cognome	Nome
10133	Bini	Carlo
12341	Verdi	Laura
43237	Rossi	Giulio

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	NomeCorso	Anno
34	Chimica	1
79	Algebra	2

Trova I Cognomi degli studenti che hanno sostenuto l'esame relativi al corso con Nome Chimica

TROVA Cognome
NELLA TABELLA Studente-Esame-Corso
CHE OTTIENI COLLEGANDO LE RIGHE DI Studente, Esame, Corso
IN CUI Studente.Matricola = Esame.Matricola **E** Esame.Codice = Corso.Codice
SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE NomeCorso **= "**Chimica"

Figura 33 – Soluzione

Come vedete, le condizioni che permettono di collegare le tre tabelle sono espresse nella istruzione IN CUI. Qui sono espresse *due* condizioni collegate dall'operatore logico E (AND in inglese e nel calcolo proposizionale); le singole condizioni sono espresse da uguaglianze tra coppie di attributi, gli attributi sono indicati con il loro nome, preceduto dal nome della tabella in cui compaiono.

L'ultima condizione (SELEZIONA LE RIGHE CHE RISPETTANO LA CONDIZIONE) esprime l'esigenza di selezionare le sole righe della tabella Studente-Esame-Corso relative al Corso di Chimica.

Se come spero siete arrivati fin qui, vi faccio i complimenti! Avete appreso i primi concetti di un linguaggio programmativo, e più in generale, un primo metodo per rimettere insieme dati separati in differenti fonti, attività tipica nel nostro mondo in cui le fonti dati sono così tante e così rapidamente mutevoli nel tempo.

Un accenno a un modello "più gerarchico" del modello relazionale

Il modello relazionale rappresenta i dati mediante tabelle, ma non è l'unico modello che usa tabelle.

Infatti, le tabelle nel modello relazionale sono, come diceva il suo inventore, Ted Codd, di spartana semplicità; ogni riga di una tabella è composta di celle in ognuna delle quali compare un valore elementare. Come conseguenza tutte le righe hanno la stessa struttura, dalla prima all'ultima.

Questo è il pregio del modello relazionale, pregio che però ha una controparte, nel senso che alcuni valori possono comparire più volte in una tabella, e quindi si può avere una sensazione di spreco di spazio.

Consideriamo ad esempio la tabella a sinistra nella Figura 34. Siccome lo studente con matricola "10133" ha dato *due* esami, il numero di matricola compare *due* volte nella tabella. Lo stesso vale per i corsi con codice "34" e "79". Questa ripetizione dei valori è proprio il prezzo che dobbiamo pagare per garantire la struttura "piatta" delle tabelle.

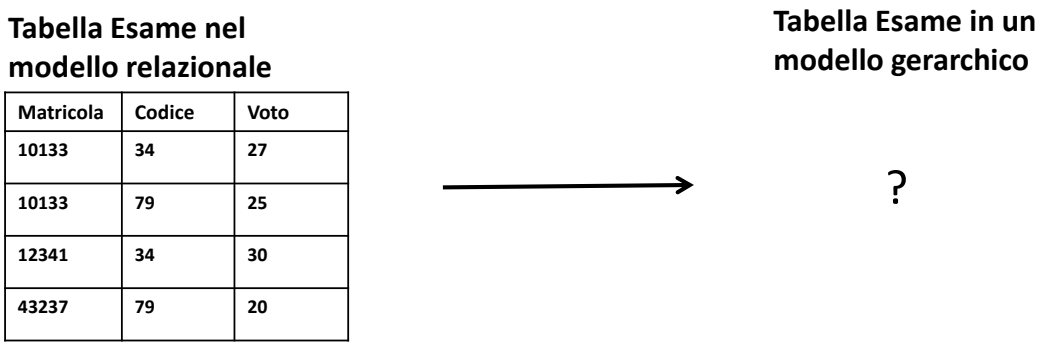


Figura 34 – Come possiamo rappresentare la tabella Esame
senza ripetere due volte la matricola 10133?

Come potremmo riorganizzare le celle della tabella con una rappresentazione che mantiene lo stesso contenuto informativo, e, allo stesso tempo, è più compatta?

Ecco, questo è il senso delle due domande che faccio ora a voi:

1. come possiamo riorganizzare la tabella riducendo da due a uno i valori della matricola "10133"?
2. Come possiamo riorganizzarla riducendo da due a uno i valori del codice pari a "34", e a "79"?

Provate a risolvere i due problemi e poi trovate le soluzioni all'ìnizio della prossima pagina.

Per la prima domanda, ho qualche idea, ma per la seconda sono prioprio al buio. Mi puoi dare qualche consiglio?

Non so se vado nella direzione giusta, ma tieni presente che in una tabella puoi modificare l'ordine delle righe e l'ordine delle colonne, senza alterare il contenuto informativo. Ti è d'aiuto questo mio consiglio?

Mah, ci penso…..

Ecco la risposta alle due domande, vedi Figura 35. Accorpare i valori di Matricola uguali significa sostituire le due celle che contengono il valore “10133” con una sola, e far corrispondere a questa sola cella i due esami sostenuti.

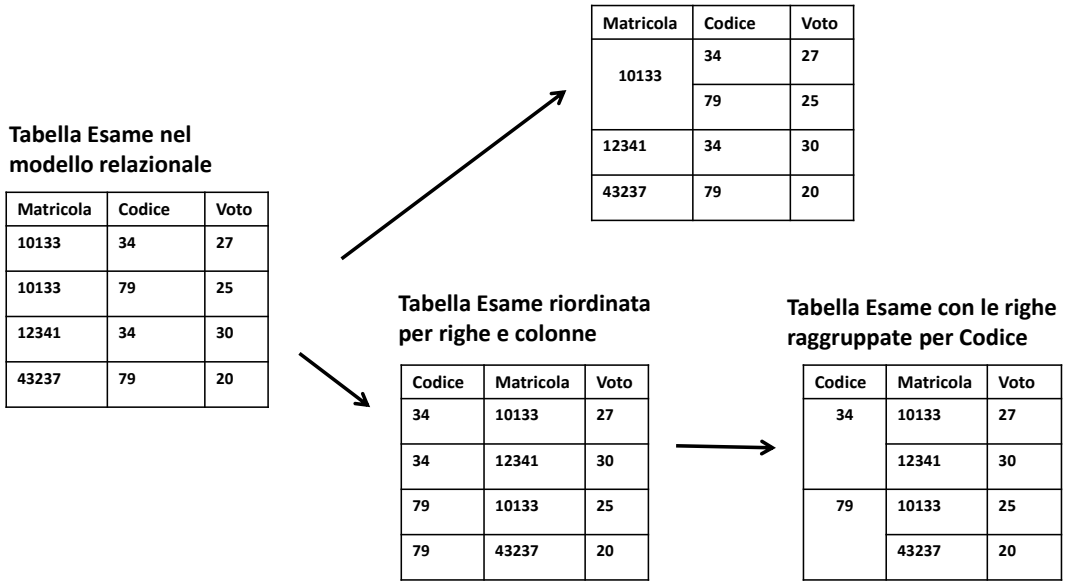


Figura 35 – Soluzione: ora il modello raggruppa le righe che hanno uguali valori di Matricola ovvero che hanno uguali valori di Codice

Per Codice il discorso è un pochino più complicato; bisogna infatti prima spostare l’attributo Codice in prima colonna, e poi riordinare le righe con uguale valore di Codice. A questo punto procediamo come prima per Matricola.

Il modello così definito prende il nome di *modello gerarchico*, perché in entrambe le soluzioni mostrate in Figura 35 i dati sono disposti in gerarchia; nel primo caso sono disposti per Matricola (attenzione, non ordinati, solo disposti) e dopo ogni matricola riportando gli esami svolti dalla matricola, nel secondo caso sono disposti per Codice.

Capitolo 3

Il modello relazionale è troppo semplice per descrivere il mondo, ci serve un modello più espressivo
Il modello Entità Relazione

Torniamo al modello relazionale. Il titolo di questa sezione ci fa già capire il suo punto di arrivo; il modello relazionale è stato proposito fin dagli anni 70 del secolo scorso da Ted Codd come modello di spartana semplicità. Fino ad allora si usavano modelli piuttosto complicati per descrivere il mondo, su cui non mi soffermerò; Codd concepì un modello così semplice che più semplice non si può, con due sole strutture di rappresentazione, la tabella, formata a righe tutte uguali, e l'attributo, che può solo assumere valori elementari.

Mi piace il modello relazionale! Essendo così semplice, tabelle + attributi, immagino che sia semplice rappresentare con il modello anche "frammenti di mondo" molto complessi….

Purtroppo non è così… Pensa al linguaggio naturale, e supponi di poter usare solo un numero limitato di vocaboli per descrivere un sentimento o una sensazione complessa, o un sogno. Non ti senti limitato? Quando si vuole rappresentare un frammento di mondo solo con le tabelle e gli attributi, ci si trova in una condizione simile a quella in cui dei bambini in una spiaggia hanno solo due formine per fare castelli di sabbia, una quadrata e una rotonda. La semplicità ha i suoi pregi, ma mostra dei limiti nella attività di modellazione di una base di dati, quando la realtà che si vuole rappresentare è ricca di significati.

Le tabelle non sono tutte uguali

In particolare, nel modello relazionale le tabelle non sono tutte uguali, non hanno tutte lo stesso "ruolo modellistico" nell'esprimere il significato dei dati. Osserviamo nuovamente la base di dati costituita dalle tabelle Studente, Esame e Corso. Le tabelle Studente e Corso rappresentano due classi di elementi dal significato molto chiaro e molto "primitivo":

- gli Studenti che frequentano una Università e
- i Corsi erogati dalla Università.

Se ci pensiamo un attimo, gli esami rappresentano classi di oggetti che nascono dalla composizione di altre classi, appunto le classi degli studenti e degli esami. In Figura 36 vediamo un esempio di statino d'esame cartaceo come ancora si usano in alcune Università. Su uno statino cartaceo, i professori devono riportare i dati relativi allo studente e al corso associato all'esame, più il voto assegnato allo studente.

Esame di _____ CFU n. _____

Modulo | | | | | | | Codice | | | | | | |

sostenuto dal sig./sig.ra _____

matricola | | | | | | | altra identificazione _____

nato/a il _____ a _____

Corso di Laurea _____

Indirizzo _____

Anno Accademico _____ sessione _____

Il sottoscritto dichiara che alla data d'esame:

4) è in regola con il pagamento di tasse e contributi, avendo versato la I rata (entro il 30 novembre dell'a.a. di riferimento) il _____; la II rata (entro il 31 marzo) il _____; la III rata (entro il 30 giugno) il _____;
5) ha legittimamente preso l'iscrizione al su indicato corso d'insegnamento e di averne ottenuto l'attestazione di frequenza, ove previsto;
6) ha rispettato la norme di propedeuticità.

Firma _____

RISERVATO ALLA COMMISSIONE ESAMINATRICE

Lo studente risulta: RESPINTO ☐ IDONEO ☐ PROMOSSO ☐

con voto (in lettere) _____ e lode SI NO

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

FIRMA DEGLI ESAMINATORI

_____ _____ _____

Presidente

Numero d'ordine del verbale _____ data _____ / _____ / _____

RISERVATO ALLA SEGRETERIA STUDENTI

Codice del Corso di Laurea dell'insegnamento _____

_____

(il segretario)

Figura 36 – Uno statino d'esame cartaceo

In Figura 37 riporto un insieme di statini "stilizzati", che forniscono i valori che compongono la tabella Esame nella parte destra.

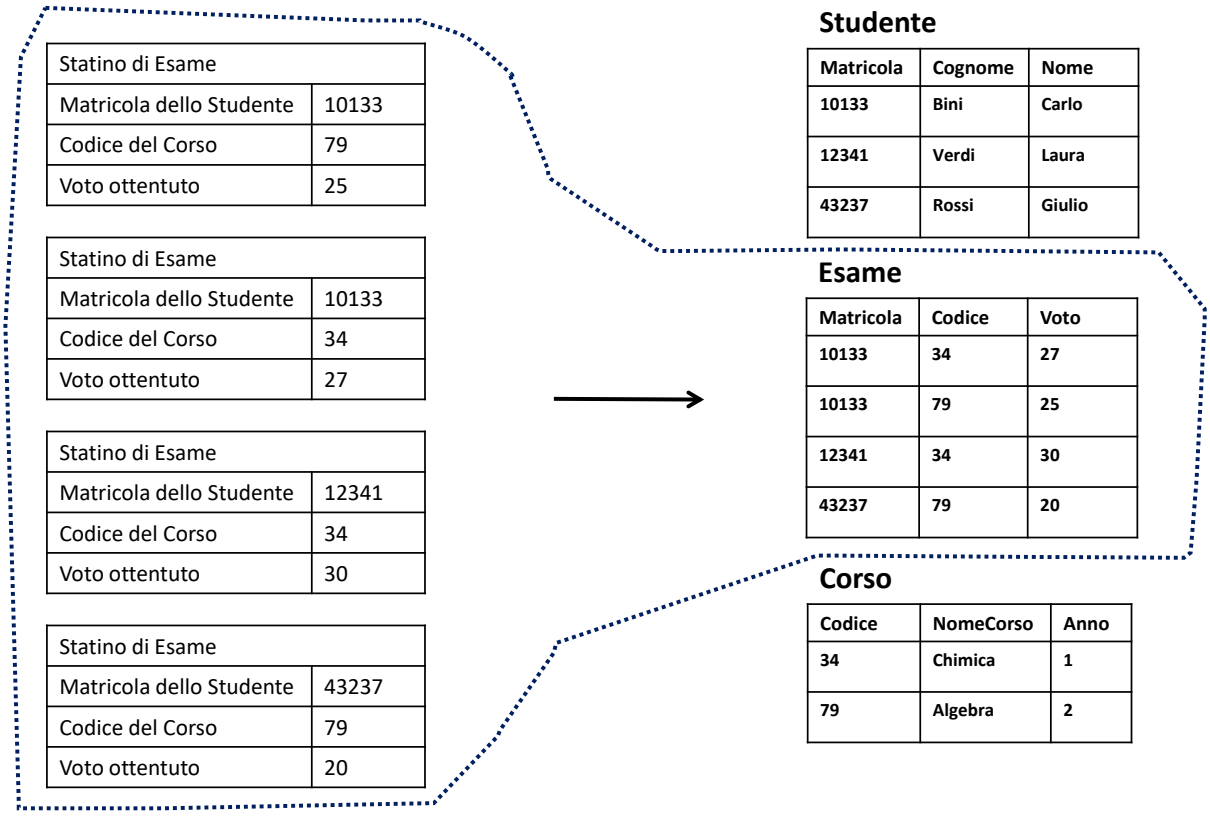


Figura 37 - Gli statini di esame e la tabella che li rappresenta

Insomma, mi sembra chiaro che Esame è un concetto che lega Studenti e Corsi; non rappresenta dunque oggetti "primitivi" della realtà, come Studente ed Esame, ma *un legame tra oggetti della realtà*, legame che spesso è chiamato *fatto*, quando si fa riferimento al legame tra due valori (Bini e Chimica), o *relazione,* quando fa riferimento al legame tra due classi (gli Studenti e i Corsi).

Si può capire dunque perché a partire dagli anni '70 del secolo scorso i ricercatori hanno cominciato a investigare modelli di dati più espressivi del modello relazionale. E fu proposto il modello Entità Relazione; quala è la differenza principale tra il modello relazionale e il modello Entità Relazione?

Il modello Entità Relazione

Nel modello relazionale, per rappresentare dati tutti caratterizzati dallo stesso insieme di attributi (ad esempio Studenti con Matricola, Nome e Cognome, oppure Esame, con Matricola, Codice e Voto) abbiamo a disposizione la sola struttura di *tabella*, nel modello Entità Relazione si distingue tra due strutture di rappresentazione:

1. *strutture di rappresentazione primitive*, non scomponibili in concetti più semplici, chiamate *Entità*; nel nostro caso sono entità Studente e Corso, e
2. *strutture di rappresentazione composte,* che mettono in relazione due Entità, chiamate *Relazioni* tra Entità; nel nostro caso, Esame può essere visto come composto di Studente e Corso (uno specifico esame riguarda uno specifico studente e uno specifico corso, come si vede nelle Figure 36 e 37).

Per capire ancora meglio le differenze tra modello relazionale e modello Entità Relazione, guardate la Figura 38. Nel modello relazionale, il legame tra Studente e Corso stabilito dal concetto Esame, viene raqppresentato duplicando in Esame le chiavi di Studente (Matricola) e Corso (Codice): è questa idea che ci permette di collegare studenti, corsi ed esami, come abbiamo fatto nelle interrogazion iviste in precedenza.

Passando alla parte superiore della figura, notiamo anzitutto che il modello Entità Relazione è dotato di una rappresentazione grafica per lo schema, che sottolinea il diverso ruolo e significato delle Entità e delle Relazioni, rappresentando le prime con rettangoli e le seconde con rombi, collegati con le Entità coinvolte dalla Relazione. In questa rappresentazione grafica le tabelle Studente e corso sono rappresentate da Entità, con i rispettivi attributi, e la tabella Esame è rappresentata come Relazione tra Studente e Corso.

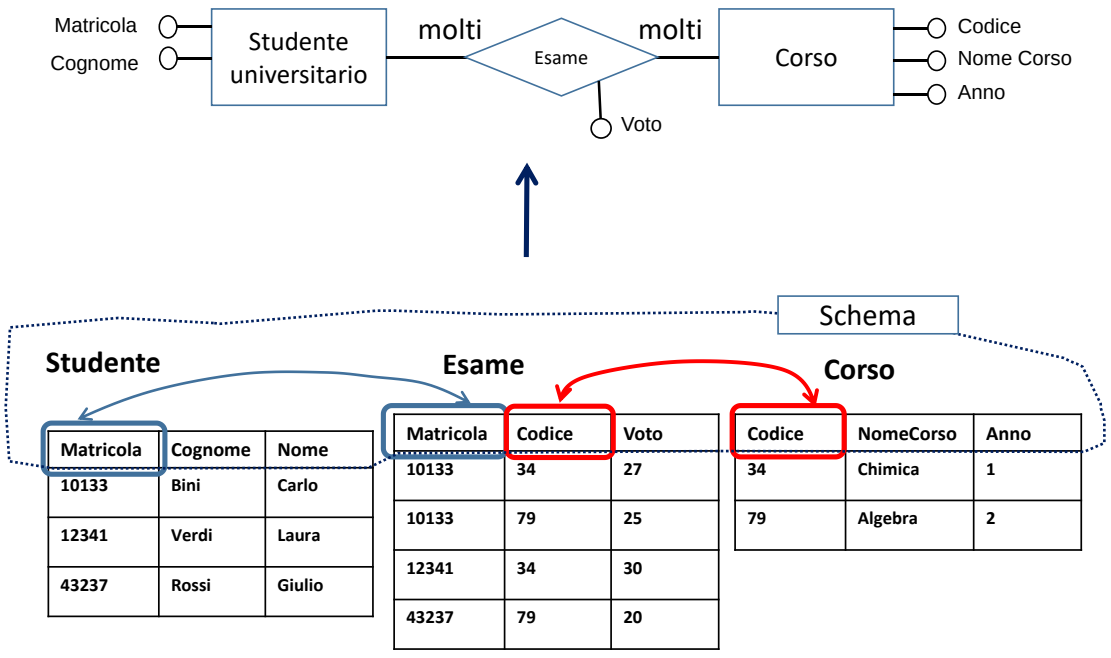


Figura 38 – Lo schema Studenti, Esami, Corsi rappresentato nel modello Entità Relazione

Nel modello Entità Relazione si rappresenta solo lo schema

Perché nell'esempio precedente hai rappresentato nel caso del modello relazionale sia i valori che lo schema, mentre nel modello Entità Relazione hai rappresentato solo lo schema?

Sei un osservatore attento! Avrei potuto rappresentare i valori anche nel modello Entità Relazione, ma la questione è che il modello Entità Relazione è stato proposto per *progettare basi di dati*, non *per usarle*. Mentre il modello relazionale è stato oggetto di tante realizzazioni tecnologiche, il modello Entità Relazione è stato proposto come modello più vicino all'utente nel modo con cui descrive la realtà, come vedremo ancora meglio tra poco. E lo si può usare anche se non si vuole realizzare una base di dati, semplicemente per comprendere meglio una realtà complessa. Quindi del modello ha senso prendere in considerazione lo schema, non i valori.

Entità, Relazioni, Attributi come strutture del modello Entità Relazione

Ricapitolando, nel modello Entità Relazione sono definiti i seguenti concetti.

1. Le Entità, che corrispondono a *insiemi di oggetti*, artefatti, persone, cose del mondo reale, che sono rappresentate con un *rettangolo*. In Figura 38 abbiamo due entità, Studente Universitario e Corso.
2. Le Relazioni, che corrispondono a *insiemi di coppie di oggetti*, artefatti, persone, cose, sono rappresentate con un rombo e due linee che le collegano con le Entità su cui sono definite.
3. Gli Attributi, che corrispondono a proprietà di Entità o Relazioni, che sono rappresentate con un pallino. In Figura 38 abbiamo sei attributi, rispettivamente Matricola e Cognome, attributi della Entità Studente Universitario, Codice, Nome Corso e Anno, attributi della Entità Corso, e infine Voto, attributo della Relazione Esame.

Perché Voto è attributo di Esame?

Eh, questa è la domanda più frequente quando insegno questi argomenti. Così come Cognome è una proprietà di Studente, perché ogni studente ha un cognome, Voto è attributo di Esame perché ogni esame ha un voto; detto in negativo, Voto non è proprietà di Studente, perché uno studente prende tanti voti, non è una proprietà di Corso perché un corso può essere superato da tanti studenti in tanti esami. Un voto (ad esempio 27), insomma, è associato ad una coppia <studente, corso>, ed è quindi proprietà di Esame.

Vediamo se hai capito. Se volessi aggiungere la Data in cui si è svolto l'esame, e la Data di Nascita dello Studente, a quali concetti nello schema li devo aggiungere?

Mmh… mi sembra semplice, Data dell'esame all'Esame, e Data di Nascita a Studente, però mi sembra una risposta troppo semplice….

Non esistono risposte semplici o complicate, esistono risposte esatte e sbagliate, e la tua risposta è esatta!

Le cardinalità delle Relazioni

Se tornate alla Figura 38, noterete che ho associato ai due rami della relazione Esame una etichetta, in entrambi i casi la etichetta "molti". Cosa significa questa etichetta? Significa che ogni studente può sostenere in generale molti esami, più di un esame, perché nel suo piano di studi ha, a seconda delle lauree, diversi corsi. Ugualmente, un corso in generale è stato superato con un esame da molti studenti, sicuramente più di uno studente. Si dice anche che la relazione Esame è *molti a molti*.

Tutte le relazioni sono molti a molti?

No, ne esistono di altri tipi.

Pensa alla relazione *Nato a* tra Persona e Comune di Nascita, assumendo che Persona e Comune di Nascita siano entità in uno schema. Ogni persona è nata in un comune, in ogni comune sono nate più persone. Si dice che la relazione Nato a è *uno a molti*.

Pensa ora alla relazione *Possiede* tra Persona e Carta di Identità. Ogni Persona possiede una carta di identità e ogni carta di identità è associata a una persona. Si dice che la relazione Possiede è uno a uno.

Credo di aver capito, ho però un dubbio. Tornando alla relazione Esame, ci sono studenti, iscritti al primo anno di Università, che non hanno superato nessun esame, in che senso associamo il valore molti alla relazione, "lato" Studente?

E' vero quello che dici, e infatti potremmo invece che dire *molti*, ad indicare un *valore massimo*, potremmo scrivere [0, molti] ad indicare sia un ***valore minimo*** (0) che un valore massimo, ma non mi va di entrare in tanti aspetti del modello, non voglio annoiarti con tanti concetti, nelle conclusioni fornirò alcuni riferimenti se volessi studiare un po' più a fondo i modelli dei dati. Basta solo dire che uno e molti vengono chiamate nel modello Entità Relazione *cardinalità*.

Per fissare bene questo concetto, guardate ora la Figura 39.

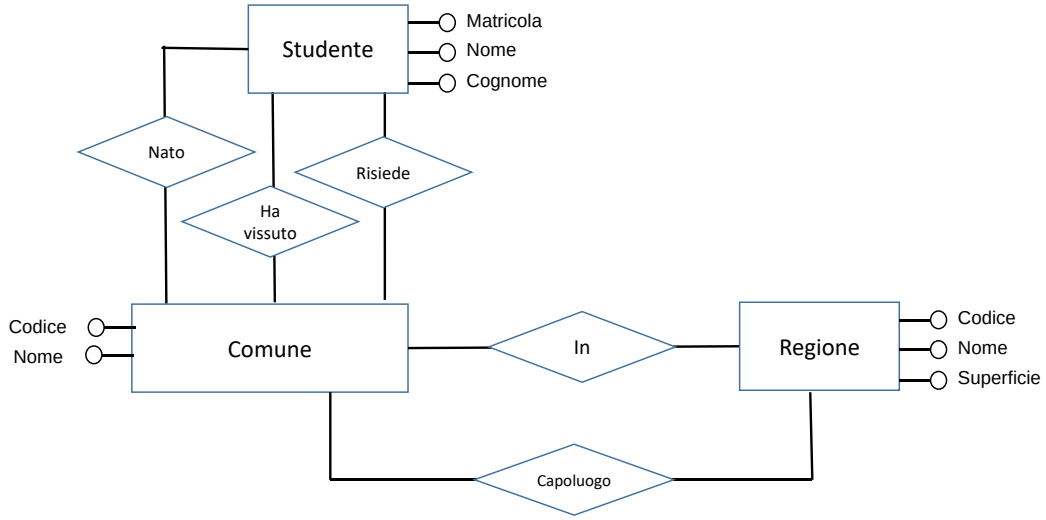


Figura 39 – Esercizio che ha lo scopo di trovare le cardinalità

Il significato di questo schema è nelle sue linee generali facilmente esprimibile. Stiamo rappresentando gli Studenti, i Comuni dove sono nati, dove risiedono e dove hanno vissuto eventualmente nel passato, e riguardo ai Comuni intendiamo rappresentare la Regione in cui sono collocati territorialmente. Infine, per le Regioni vogliamo anche rappresentare il Comune capoluogo.

Provate ora a definire le cardinalità per le diverse Entità e Relazioni nello schema. La soluzione è mostrata in Figura 40.

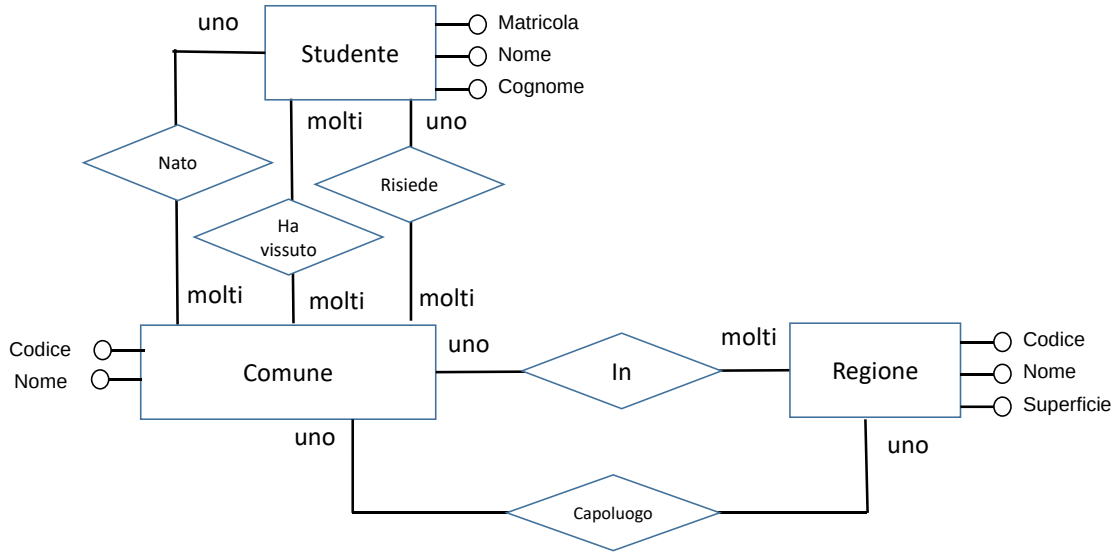


Figura 40 – Soluzione

Non è così facile decidere tra uno e molti, mi puoi suggerire un metodo per scegliere le cardinalità?

Certo, puoi farti domande del tipo: Dato uno studente, in quanti comuni è nato? Cosa rispondi?

E' evidente, in un solo comune!

E dato un Comune, di quante regioni può essere capoluogo?

Una!

E dato un comune, quanti studenti possono risiedervi?

Tanti!

Ecco, vedi, non è difficile.

Il concetto di aggregazione

Mi preme ora dire che le Relazioni nel modello Entità Relazione esprimono un costrutto che noi usiamo ogni giorno per organizzare la nostra conoscenza, il costrutto di ***aggregazione***. Abbiamo già parlato della aggregazione nel Capitolo 9 del primo libro della Enciclopedia. Pensate a una data; può essere vista come composta di un giorno, un mese, un anno. Possiamo dire che una data si compone, è (il risultato di una) operazione di *aggregazione* di un giorno, un mese e un anno. Possiamo anche dire che la relazione Esame è una aggregazione di uno Studente e un Corso.

Le aggregazioni sono dunque strutture modellistiche che possiamo usare per rappresentare realtà complesse. Le aggregazioni sono concetti composti di parti; un giorno è parte di una data, uno studente è parte di un esame.

Il concetto di generalizzazione

Guardate ora l'esempio di Figura 41, in cui nel testo ho introdotto due tipi di studenti, gli italiani e gli stranieri; per gli italiani ci interessa rappresentare il comune di nascita, per gli stranieri il paese di provenienza. L'esempio di Studente che si può specializzare in Studente italiano e Studente straniero mostra in maniera semplice una situazione che incontriamo ogni giorno; la realtà di fronte a noi evidenzia tante diversità, pensate agli esseri animati e inanimati, pensate alle classificazioni delle piante, degli animali, degli atomi, di tanti altri aspetti della natura.

Come facciamo dunque a rappresentare due diversi tipi di studenti in una tabella?

Ci sono tre studenti di una Università, Bini, italiano, con matricola 10133, nato a Milano, Verdi, italiano, con matricola 12341, nato a Roma, e Rossi, peruviano, com matricola 43237. I corsi erogati dalla Università sono due, Chimica, con codice 34 al primo anno, e Algebra con codice 79, al secondo anno. Lo studente con matricola 10133 ha superato I corsi con codice 34 e 79, con voti 27 e 25; lo studente com matricola 12341 ha superato il corso con codice 34 con voto 30; lo studente con matricola 43237 ha superato il corso con codice 79 con voto 20.

Studente

Matricola	Cognome	Italiano?	Comune Nascita	Stato
10133	Bini	sì	Milano	-
12341	Verdi	sì	Roma	-
43237	Rossi	no	-	Perù

Esame

Matricola	Codice	Voto
10133	34	27
10133	79	25
12341	34	30
43237	79	20

Corso

Codice	Nome	Anno
34	Chimica	1
79	Algebra	2

Figura 41 – Rappresentare due tipi di studenti diversi

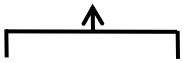
In Figura 41 è mostrata una tabella Studente in cui abbiamo usato un attributo, "Italiano?" per distinguere con un "sì" o con un "no" la nazionalità degli studenti, e due altri attributi, Comune Nascita e Stato, che presentano valori significativi nel primo caso (si) per gli italiani e nel secondo (no) per gli stranieri.

Un attimo, un attimo! Ciò che hai fatto è scorretto! Le parole sono importanti! Tu hai chiamato "Italiano?" un attributo della tabella Studente. Ora, capisco che non c'era nessuna cattiva intenzione, ma perché non lo hai chiamato "Straniero?"? In fondo i due attributi hanno lo stesso contenuto informativo, solo che quando "Italiano?" ha valore ""si", "Straniero?" ha valore "no" e viceversa; ma "Italiano?" mettere in enfasi la italianità; non possiamo usare un nome più neutro di "Italiano?" o "Straniero?", che metta italiani e stranieri sullo stesso piano?

Accipicchia, ti stai creando una sensibilità che ti porta a fare osservazioni molto giuste e molto profonde, come questa! Insomma, mi pare che stai cominciando a sperimentare il fatto che i

modelli dei dati non sono asettiche rappresentazioni della realtà, ma *strumenti per descrivere il mondo*, e quindi vanno usati con cura, così come va usato con cura il linguaggio naturale. Anzi, forse, siccome gli schemi dati sono molto più sintetici di un testo in linguaggio naturale, e sono usati nelle applicazioni informatiche, ad esempio nella erogazione di servizi nelle pubbliche amministrazioni, vanno usati con ancora maggiore attenzione rispetto al linguaggio naturale! Riguardo alla tua obiezione, modifico il nome dell'attributo in Nazionalità, che riporterò nelle prossime figure.

Guardiamo ora la Figura 42, in cui riporto solo lo schema relazionale, senza i valori. Guardando lo schema, spero condividiate che non è immediato percepire dallo schema che ci troviamo di fronte a due tipologie di studenti, e che Comune fa riferimento solo agli studenti italiani e Stato agli studenti stranieri. Il modello relazionale svela qui come la sua semplicità crei grossi limiti di espressività. Avremmo bisogno di una nuova struttura che ci permetta di dire che gli studenti sono di due tipi, italiani e stranieri. Questo è ciò che offre il modello Entità Relazione, il cui schema è mostrato sempre in Figura 42. La struttura rappresentata con il simbolo grafico



è un esempio di ***generalizzazione***, definita nello schema tra la entità genitore Studente e le entità figlie Studente italiano e Studente straniero. Abbiamo già parlato della generalizzazione nel Capitolo 9 del primo libro della Enciclopedia. Dire che Studente è generalizzazione di Studente italiano e Studente straniero significa affermare che ogni studente italiano è uno Studente, e che ogni studente straniero è uno Studente, e che l'insieme degli studenti italiani e studenti stranieri esaurisce l'insieme degli studenti.

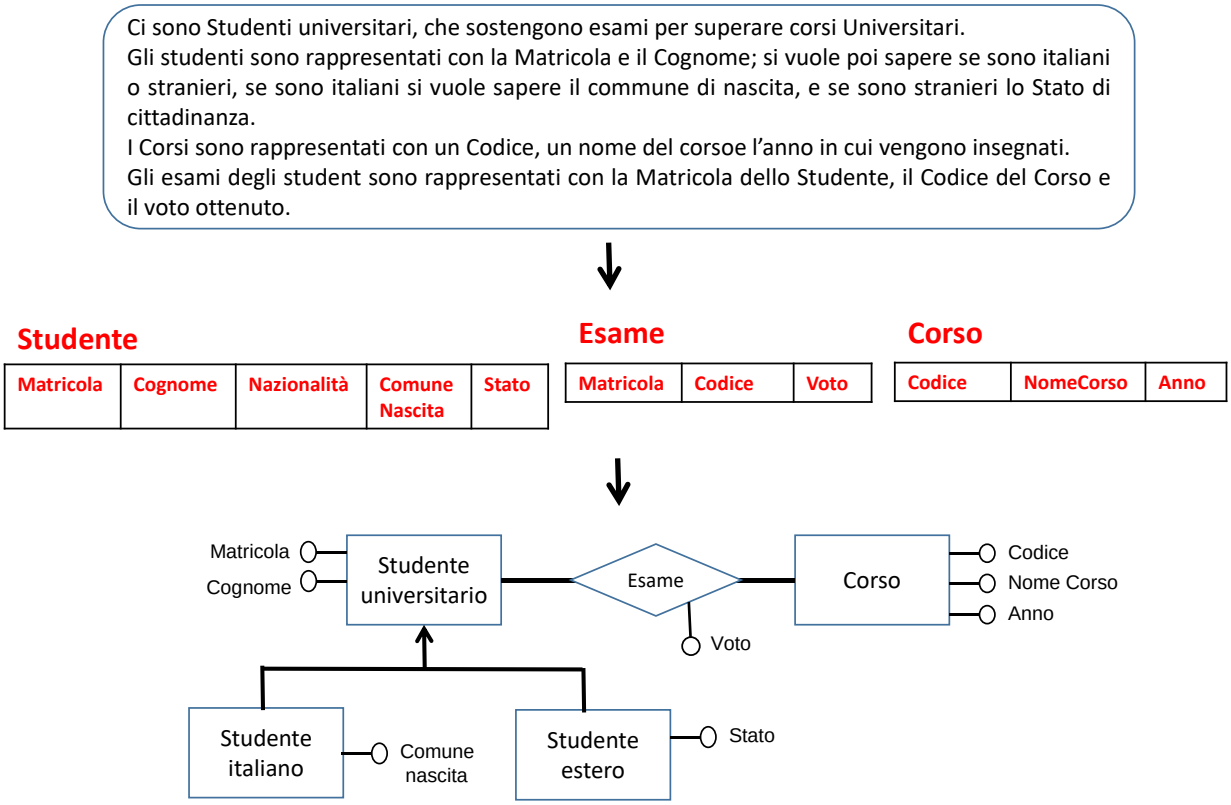


Figura 42 – Una nuova struttura: la generalizzazione

Notate anche che l'attributo Comune nascita è associato alla sola entità Studente italiano e l'attribuito Stato alla sola entità Studente Straniero. E' chiaro il tutto?

Sì, è chiaro, ma mi chiedo: se, come è ovvio ogni studente italiano e ogni studente straniero sono uno studente, allora cosa ne è dei due attributi Matricola e Cognome? Per spiegarmi meglio: sono attributi della sola entità Studente, oppure li devo ripetere anche per le entità Studente italiano e Studente straniero?

La proprietà di ereditarietà

Ancora complimenti! Hai scoperto da solo una proprietà importantissima delle generalizzazioni, la *proprietà di ereditarietà*. Andiamo con ordine.

Osserva nuovamente ia Figura 42. Ti ricordo che le Entità rappresentano classi, l'entità "Studente Universitario" rappresenta la classe di tutti gli studenti, l'entità "Studente Italiano" la classe degli student italiani e così via. Siccome tutti gli studenti italiani e gli studenti stranieri sono studenti, ogni attributo di Studente è anche un attributo delle entità figlie. Sei d'accordo?

Certo, ad esempio se ho rappresentato il fatto che tutti gli studenti universitari hanno un cognome, allora hanno un cognome sia gli studenti italiani che gli studenti stranieri, mi sembra ovvio!

Ora riconosci altre strutture di rappresentazione dello schema associate alla entità Studente che si trasferiscono anche alle entità figlie?

Stai facendo forse riferimento alla relazione Esame?

Esattamente, anche la relazione Esame è definita anche per le entità figlie; ciò è come dire: sia gli studenti italiani che quelli stranieri superano esami!

Mi piace la struttura di generalizzazione!

Certo! Ti ricordi l'esempio dell'albero nel Capitolo 9 del Primo Libro della Enciclopedia. Lo riproduco in Figura 43; quale nome useresti per indicare l'oggetto in figura tra:

1. cosa
2. albero
3. pino
4. pino romano (o domestico)?

Figura 43 – Come chiamo questo “oggetto”?

E’ probabile che tu abbia scelto il termine albero, oppure pino, mentre cosa è troppo generale, e pino romano (o domestico) richiede un minimo di conoscenza di botanica.

Quando una generalizzazione è definita tra una entità genitore e una sola entità figlia viene chiamata con la etichetta *è-un*. In Figura 44 mostriamo l’insieme completo dei simboli grafici associati ai concetti del modello Entità Relazione.

Le generalizzazioni, come molto altri concetti, non sono state scoperte dagli informatici! Di generalizzazione si parla in filosofia, in linguistica, in scienze cognitive, e in altre scienze. Gli informatici hanno “riscoperto” il concetto nel momento in cui si sono posti il problema di individuare un insieme minimo di concetti che potessero arricchire il modello relazionale, e essere utilizzati per descrivere un frammento di realtà in modo da coglierne le proprietà più importanti. Si dice anche che il modello Entità Relazione è più espressivo del modello relazionale nel descrivere la *semantica* dei dati, cioè il loro significato.

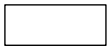
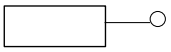
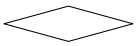
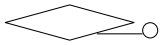
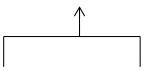
Concetto	Simbolo grafico
Entità	
Attributo di entità	
Relazione	
Attributo di relazione	
E' un	
Generalizzazione	

Figura 44 – Simboli grafici associati al modello Entità Relazione

E in effetti guardando la figura 45, in cui la stessa descrizione della realtà è modellata mediante il modello relazionale e mediante il modello Entità Relazione, credo possiamo convenire che il modello Entità Relazione è molto più ricco e chiaro del modello relazionale nel rappresentare il testo, si "fa capire molto meglio", senza bisogno di ulteriori parole o spiegazioni.

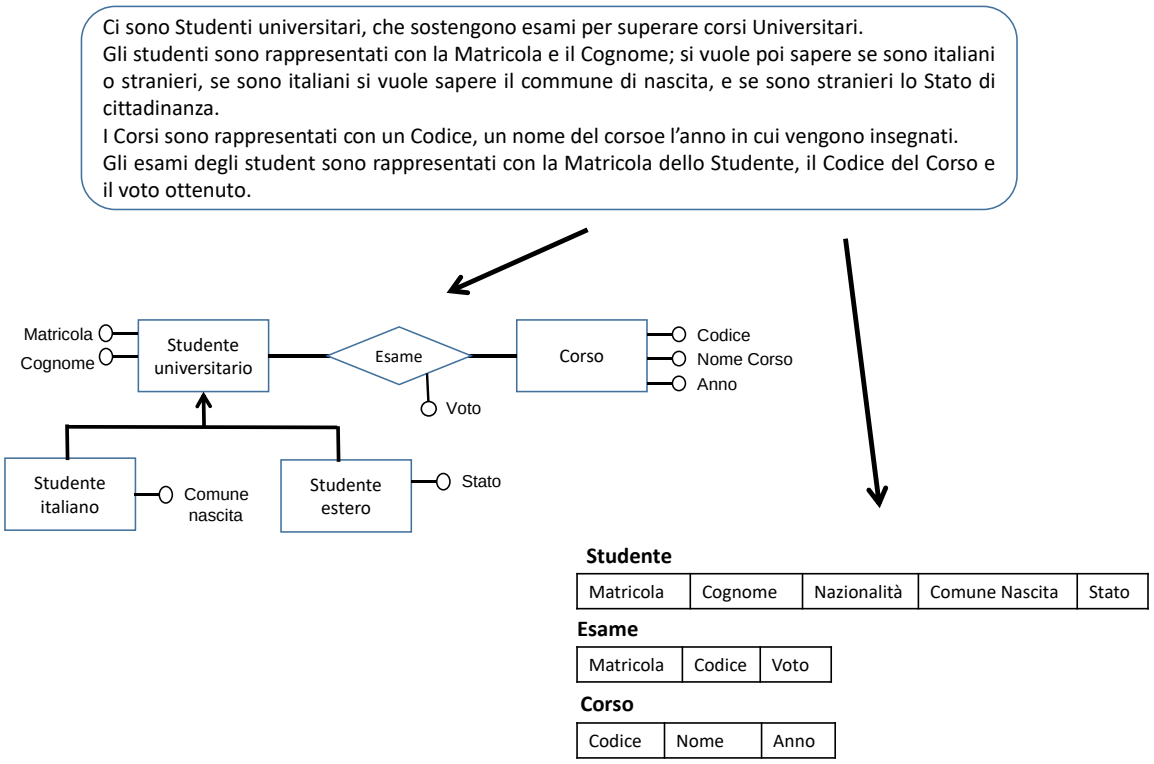


Figura 45 – Il modello Entità Relazione è più espressivo e comprensibile

Come leggere uno schema nel modello Entità Relazione

Facciamo ora un esercizio di lettura; partendo dallo schema Entità Relazione di Figura 46, provate a tradurlo, mentalmente o con un testo scritto, in parole.

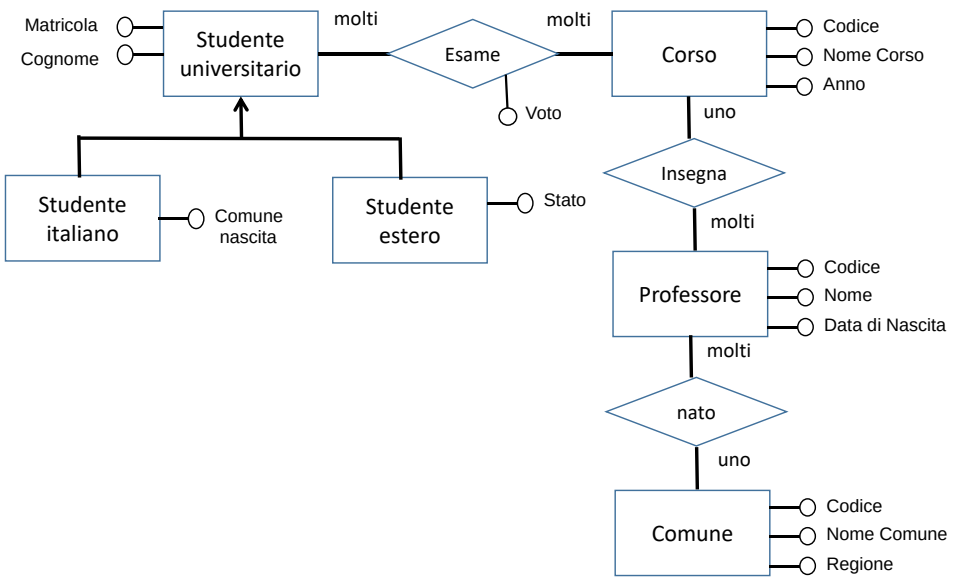


Figura 46 – Traduci lo schema in un testo italiano

In Figura 47 vediamo due testi, tra i tanti, che possono essere prodotti.

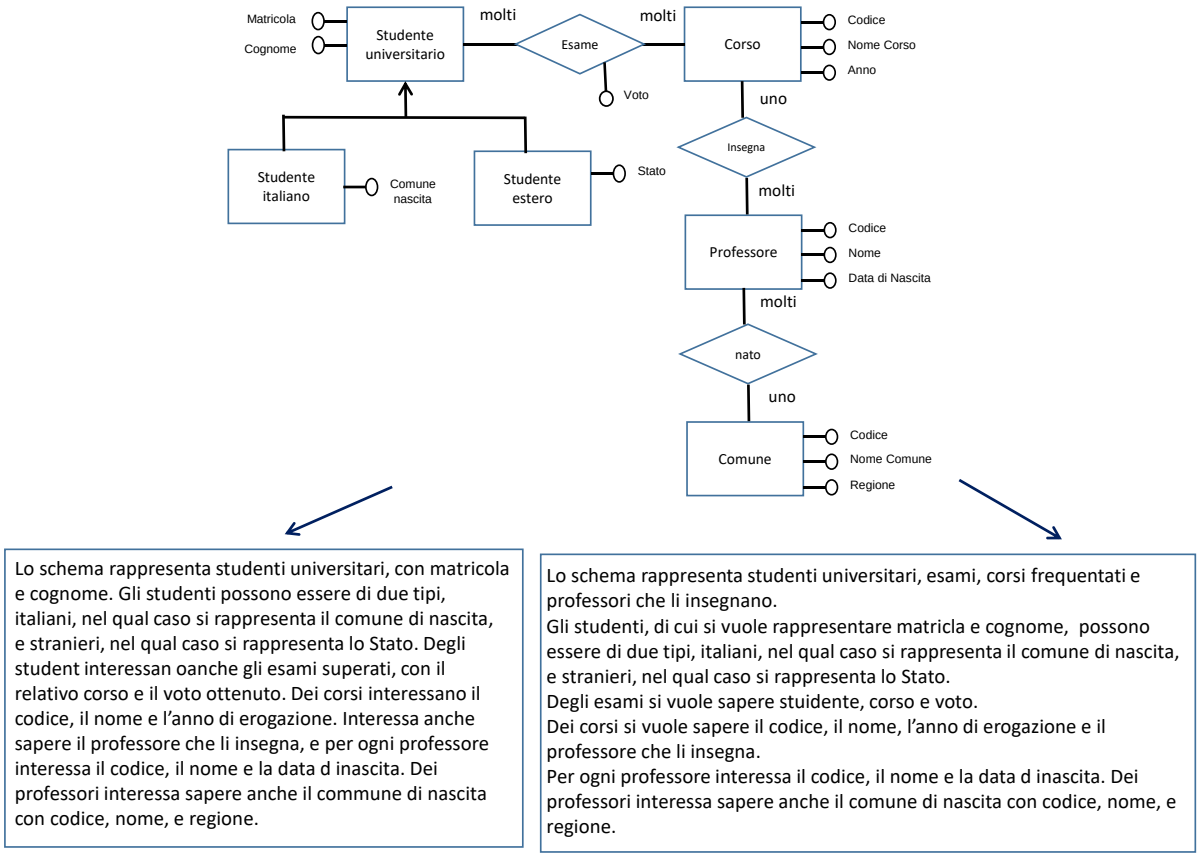


Figura 47 – Due possibili testi che descrivono lo schema

Proviamo a leggere e rileggere un'altra volta i due testi. Essi sono stati costruiti osservando lo schema Entità Relazione, e trasformando i vari frammenti di schema visitati in successive frasi in linguaggio italiano.

Facciamo ora l'esercizio inverso. Attenzione: cerchiamo di capire, partendo dai due testi, come possiamo generare lo schema Entità Relazione, simulando come gli occhi si muovono diversamente nel produrlo, vedi Figure 48, 49 e 50.

Per quanto riguarda la Figura 48, vediamo rappresentate in corsivo nei due testi le parti degli schemi circondate da linee punteggiate nei due schemi. Nel testo a sinistra, ci siamo focalizzati sulla entità principale, lo Studente, descrivendolo con tutte le sue proprietà; nel testo a destra, invece, ci si concentra inizialmente su una parte dello schema più ampia, costituita dalle entità Studente universitario, Corso, Professore, e sulle relazioni che le collegano.

Passando alla Figura 49, adesso le parti del testo in corsivo si ampliano; nel caso dello schema a sinistra, l'occhio si muove "a macchia d'olio", generando dapprima le due entità Studente Italiano e Studente Straniero, e poi la parte dello schema relativa alla entità Corso. Nel caso dello schema a destra, si procede "dal generale al particolare", introducendo gli attributi delle entità Studente, Corso e Professore, e la generalizzazione di Studente (vedi in entrambi i casi le successive linee punteggiate).

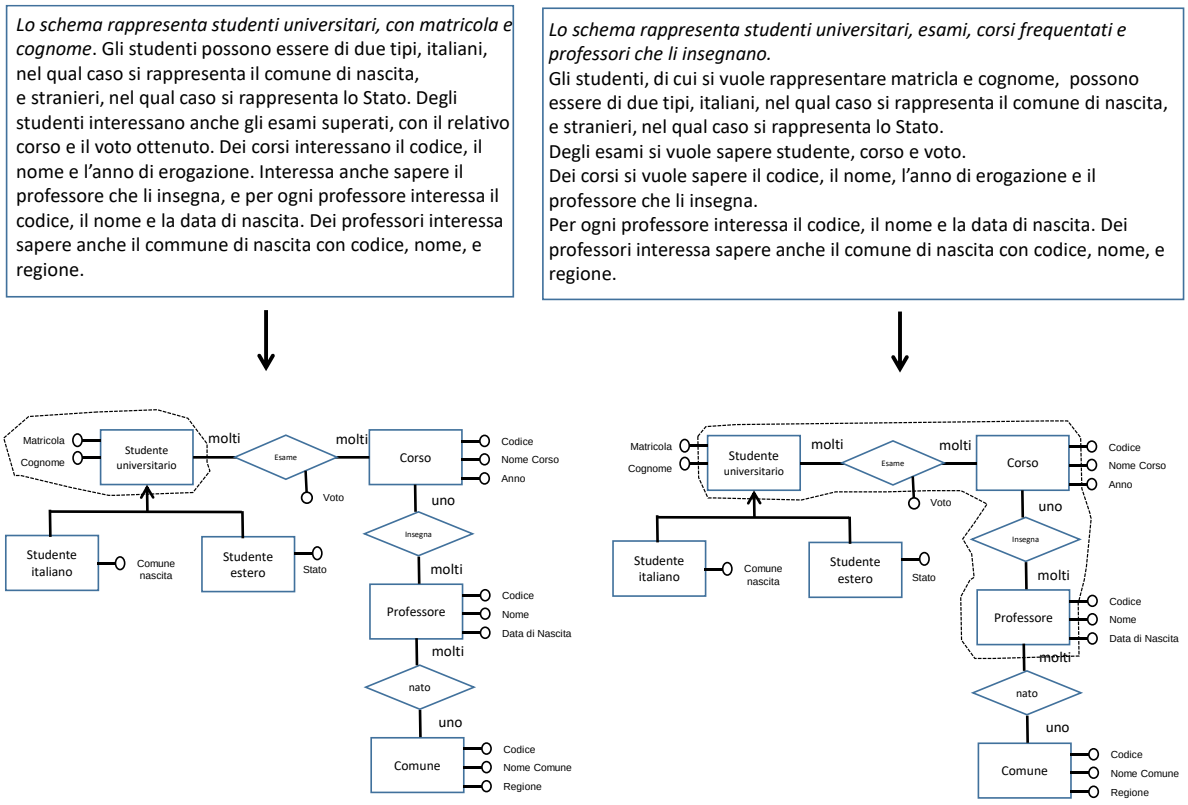
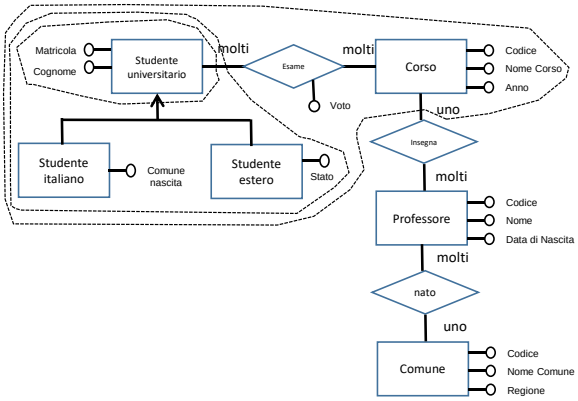


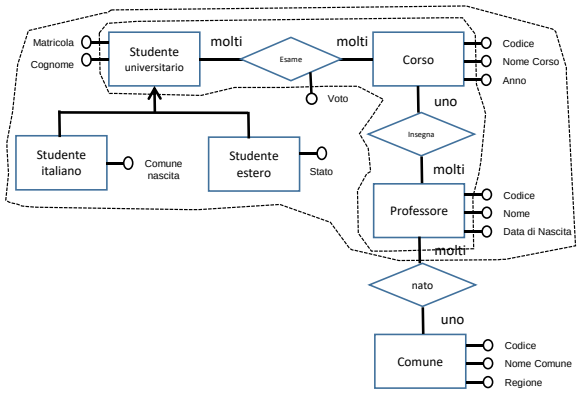
Figura 48 – Due letture diverse della stessa realtà – passo 1

Lo schema rappresenta studenti universitari, con matricola e cognome. Gli studenti possono essere di due tipi, italiani, nel qual caso si rappresenta il comune di nascita, e stranieri, nel qual caso si rappresenta lo Stato. Degli studenti interessano anche gli esami superati, con il relativo corso e il voto ottenuto. Dei corsi interessano il codice, il nome e l'anno di erogazione. Interessa anche sapere il professore che li insegna, e per ogni professore interessa il codice, il nome e la data di nascita. Dei professori interessa sapere anche il commune di nascita con codice, nome, e regione.



Esplorazione "a macchia d'olio"

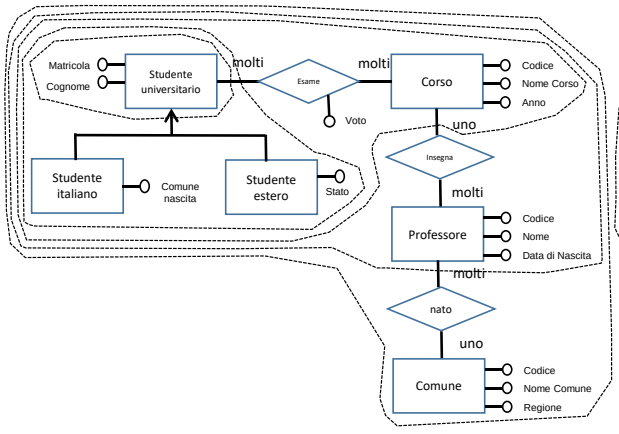
Lo schema rappresenta studenti universitari, esami, corsi frequentati e professori che li insegnano.
Gli studenti, di cui si vuole rappresentare matricla e cognome, possono essere di due tipi, italiani, nel qual caso si rappresenta il comune di nascita, e stranieri, nel qual caso si rappresenta lo Stato.
Degli esami si vuole sapere studente, corso e voto.
Dei corsi si vuole sapere il codice, il nome, l'anno di erogazione e il professore che li insegna.
Per ogni professore interessa il codice, il nome e la data di nascita. Dei professori interessa sapere anche il comune di nascita con codice, nome, e regione.



Esplorazione "dal generale al particolare"

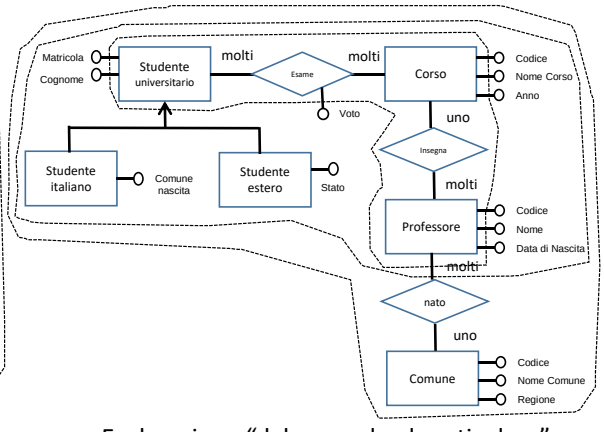
Figura 49 – Due letture diverse della stessa realtà – passo 2

Lo schema rappresenta studenti universitari, con matricola e cognome. Gli studenti possono essere di due tipi, italiani, nel qual caso si rappresenta il comune di nascita, e stranieri, nel qual caso si rappresenta lo Stato. Degli studenti interessano anche gli esami superati, con il relativo corso e il voto ottenuto. Dei corsi interessano il codice, il nome e l'anno di erogazione. Interessa anche sapere il professore che li insegna, e per ogni professore interessa il codice, il nome e la data di nascita. Dei professori interessa sapere anche il commune di nascita con codice, nome, e regione.



Esplorazione "a macchia d'olio"

Lo schema rappresenta studenti universitari, esami, corsi frequentati e professori che li insegnano.
Gli studenti, di cui si vuole rappresentare matricla e cognome, possono essere di due tipi, italiani, nel qual caso si rappresenta il comune di nascita, e stranieri, nel qual caso si rappresenta lo Stato.
Degli esami si vuole sapere studente, corso e voto.
Dei corsi si vuole sapere il codice, il nome, l'anno di erogazione e il professore che li insegna.
Per ogni professore interessa il codice, il nome e la data di nascita. Dei professori interessa sapere anche il comune di nascita con codice, nome, e regione.



Esplorazione "dal generale al particolare"

Figura 50 – Due letture diverse della stessa realtà – passo 2

Infine nella Figura 50 tutto il testo è rappresentato dalle due linee punteggiaste finali. Riassumendo, possiamo dire che la strategia di produzione dello schema dal testo è "a macchia d'olio", perché assomiglia a una macchia d'olio che si espande su una superficie, dal punto inziale verso l'esterno. Questa modalità di esplorazione è la più intuitiva che ci possa venire in mente.

La strategia di produzione dello schema a destra procede invece dal generale al particolare, e viene spesso detta *top-down*. Fornisce dapprima una visione sintetica dello schema, attraverso i soli nomi delle entità principali, Studente, Corso, Professore. A questo punto, estende l'orizzonte agli attributi delle tre entità, alle relazioni che le collegano, alle due tipologie di studenti e alle loro proprietà.

Come modificare uno schema nel modello Entità Relazione

Vediamo ora come possiamo modificare uno schema quando arricchiamo un testo con nuovi particolari. C'è qui una gerarchia di difficoltà che state affrontando: leggere uno schema è più facile che modificarlo, e modificarlo è più facile che progettarlo dal nulla.

Partiamo dallo schema riprodotto nuovamente in Figura 51, e proviamo ad arricchirlo con le parti in neretto del testo mostrate in figura. Provate a usare una matita per disegnare i nuovi simboli, e poi confrontate la soluzione con lo schema in Figura 52 nella prossima pagina.

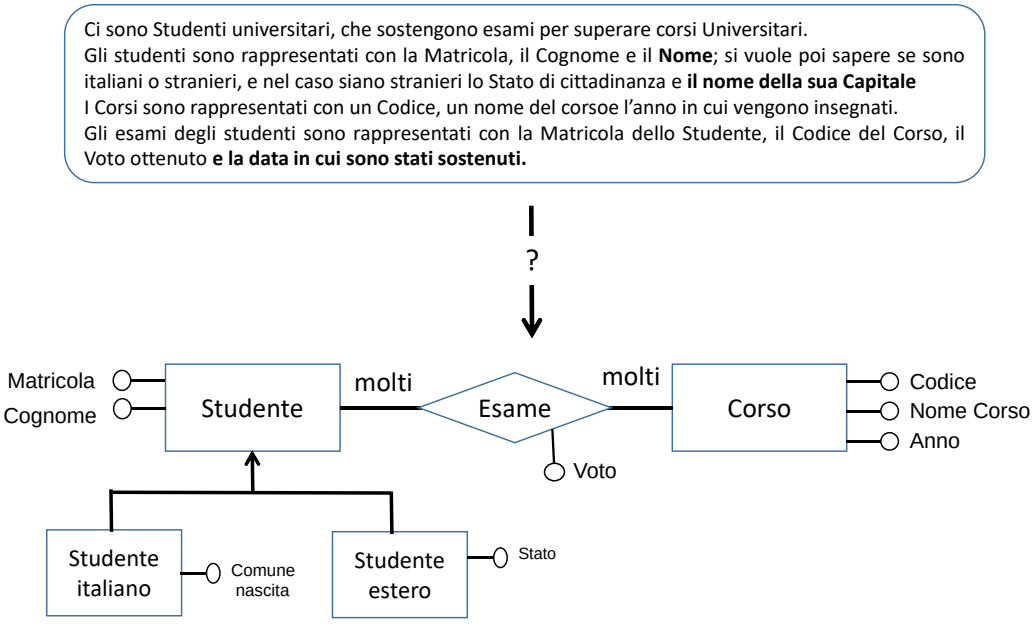


Figura 51 – Le modifiche da effettuare e lo schema da modificare

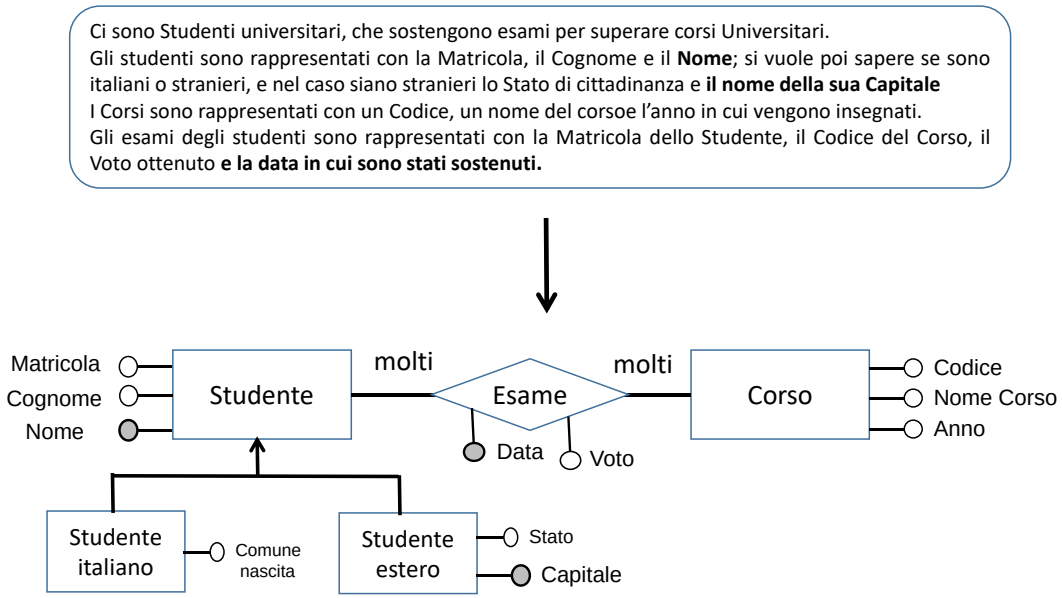


Figura 52 – Soluzione

Per ogni nuova frase che compare nel testo, dobbiamo capire a quale concetto dello schema fare riferimento. Per esempio, per quanto riguarda il Nome, è chiaro che è un nuovo attributo di Studente, mentre invece il "nome della sua Capitale" va riferito a Studente estero, e la "data in cui sono stati sostenuti" fa riferimento non alla entità Studente, non alla entità Corso, ma, piuttosto, alla relazione Esame tra Studente e Corso.

Spero di essere riuscito a convincere i lettori che il modello Entità Relazione è allo stesso tempo più espressivo e più facile da usare del modello relazionale; è più espressivo perché ha un numero di concetti più ampio per descrivere una realtà di nostro interesse, è più facile perché ha una rappresentazione grafica che lo rende comprensibile in maniera più intuitiva.

Quindi, tutte le volte che volete capire un frammento della realtà che vi circonda, provate ad usare il modello Entità Relazione, e vedete cosa viene fuori! Naturalmente, non riuscirete a rappresentare tutto, ma gli insiemi o classi di dati, le proprietà delle classi, le relazioni tra le classi, le generalizzazioni tra le classi, quelle sì che riuscirete a rappresentarle! Spesso queste strutture sono sufficienti per comprendere molte cose!

Come "rimettere insieme" (o, integrare) frammenti di mondo con il modello Entità Relazione

Fino ad ora abbiamo visto come il modello Entità Relazione si può usare per rappresentare un frammento di mondo, generando uno schema, e come, inversamente, partendo da uno schema, questo si possa "leggere" per ricostruire un frammento di mondo.

In entrambi i casi stiamo parlando di un unico frammento di mondo e un unico schema concettuale. Ma la realtà è così varia e così complessa che possiamo ben arrivare alla conclusione che tanti diversi frammenti di mondo siano da rappresentare con tanti schemi

concettuali. E allora diventa importante cercare di capire cosa accade quando io abbia a disposizione due schemi Entità Relazione, e voglia metterli insieme per crearne uno che li comprenda entrambi.

Il tema che ci accingiamo a discutere riguarda un aspetto profondo della nostra vita, quello che viviamo ogni giorno quando confrontiamo la nostra visione del mondo con quella di un parente, di un amico, di una conoscente. Quante volte ci accade di voler confrontare la nostra visione del mondo, di un pezzetto di mondo, con quella di un interlocutore? E quante volte fatichiamo a trovare un punto di contatto, una visione comune? Badate, non è indispensabile che tutti la pensiamo allo stesso modo, questo è l'obiettivo che cercano le dittature, il pensiero unico. L'importante è confrontarsi, capire ciò su cui si è d'accordo e ciò in cui si dissente. E anche qui, gli schemi Entità Relazione non sono la panacea, ma possono aiutare.

Supponiamo di avere prodotto nel passato i due schemi Entità Relazione di Figura 53. Vogliamo generare un nuovo schema che li contenga entrambi, uno schema, insomma, che rappresenti entrambi i frammenti di mondo rappresentati dai due schemi di partenza.

Calma. Fermiamoci un attimo, ho la necessità di capire. Fino a poco fa, hai sempre rappresentato mediante una o più frasi in italiano un frammento di mondo, e poi uno schema che lo rappresenta nel modello relazionale o nel modello Entità Relazione. Adesso c'è una grande novità, che non voglio far passare sotto silenzio: tu sei direttamente partito da due schemi, senza rappresentare prima i due frammenti di mondo, c'è qualcosa che non va!

Anche questa è un'ottima osservazione, che mi permette (spero) di spiegarmi meglio. Una volta che io ho rappresentato la realtà mediante uno schema, da questo momento in poi posso usare lo schema *invece che* il testo in linguaggio naturale, ad esempio, per esplorarlo, per capire quali sono i percorsi logici che posso seguire, insomma per capire meglio il frammento di mondo cui fa riferimento.

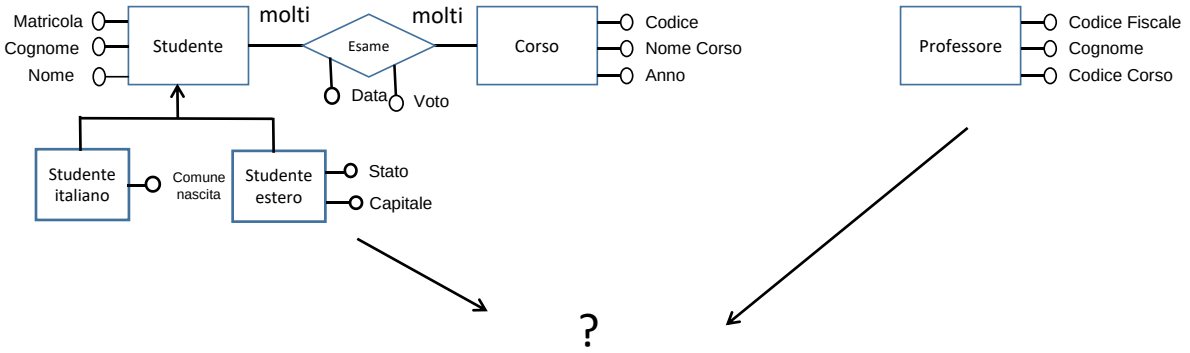


Figura 53 – Cosa significa integrare due schemi

Quindi è naturale che io possa "lavorare" sugli schemi, scordandomi temporaneamente del mondo. Quando voglio fare riferimento al mondo, posso sempre "leggere" lo schema trasformandolo, ad esempio, in un testo in italiano, oppure anche rappresentando solo mentalmente il mondo decritto dallo schema. E questo è anche ciò che posso fare quando io

abbia a disposizione due schemi (o tanti schemi) e voglia costruire un nuovo schema che li comprenda entrambi. Che è ciò che vi propongo di fare in questa sezione.

Torniamo alla Figura 53. Cercate di rispondere inizialmente alla seguente domanda: hanno i due schemi qualche concetto in comune? La risposta nella prossima pagina.

A ben vedere l'unico concetto in comune è il Codice corso, che è rappresentato come attributo di Corso nel primo schema e come attributo di Professore nel secondo schema, vedi Figura 54.

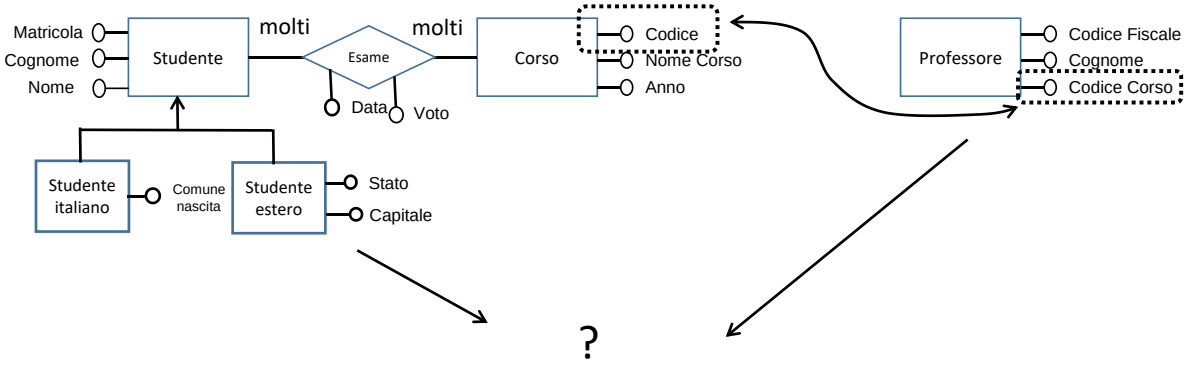


Figura 54 – I due concetti che si corrispondono nei due schemi

Quindi i due schemi hanno un concetto in comune, ma associato a due entità diverse. Nel primo schema Codice Corso è una proprietà di corso, nel secondo è una proprietà di Professore, è come dire "ogni professore insegna un (solo) corso, che è rappresentato per mezzo del suo codice". Sono due modi diversi di vedere i corsi, nel primo caso come entità dotata di proprietà, nel secondo caso come una proprietà di Professore, appunto il codice del corso che insegna. Possiamo riconciliare questi due punti di vista diversi? Quale concetto dobbiamo modificare, e in quale schema, per essere in grado di rappresentare il codice corso allo stesso modo? Provate a risolvere questo problema, la risposta nella prossima pagina.

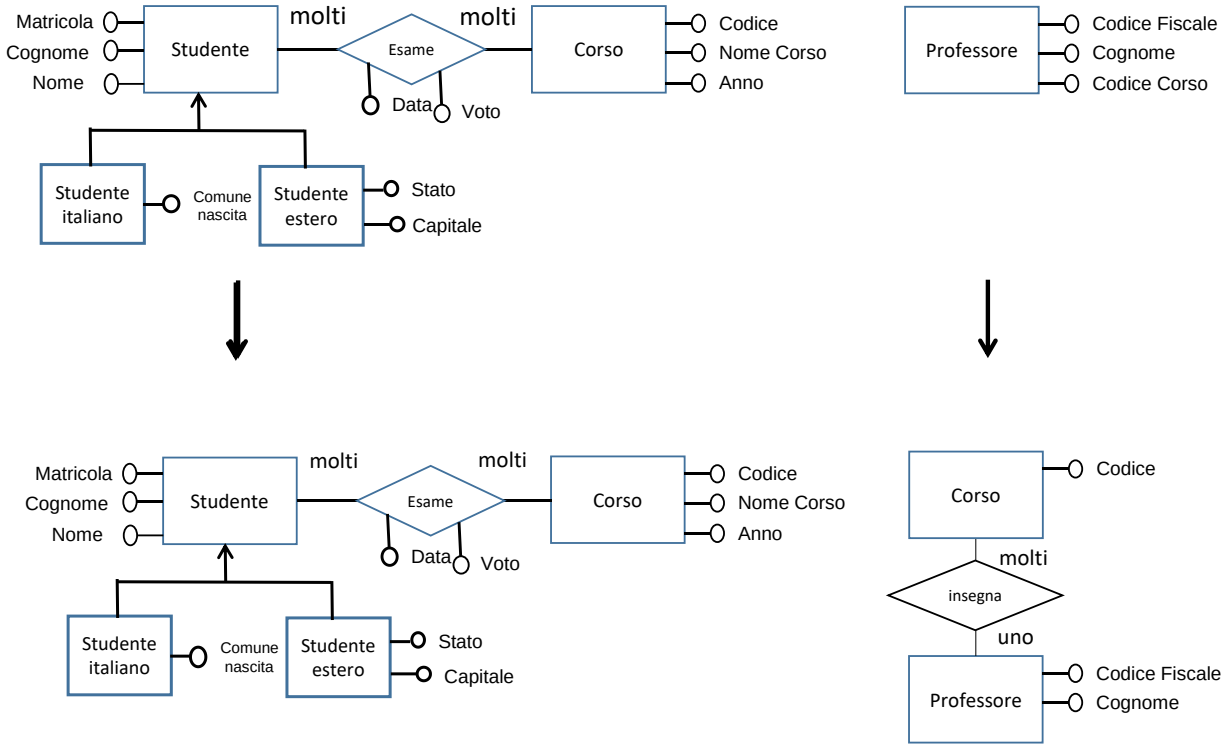


Figura 55 – Come modificare i due schemi (in realtà solo il secondo) per rappresentare Codice Corso nello stesso modo in entrambi.

Per rappresentare Codice corso nello stesso modo nei due schemi, l'unica strada percorribile è quella di introdurre una entità Corso nel secondo schema a cui associamo l'attributo Codice Corso, vedi Figura 55. Se non siete convinti provate a leggere mentalmente il secondo schema, la frase che generate è esattamente quella di prima: "ogni professore insegna un (solo) corso, che è rappresentato per mezzo del suo codice".

Mi puoi spiegare da quale ragionamento derivano le due cardinalità "uno, molti" della relazione insegna?

Certo, ogni professore insegna un corso, perché nello schema da cui siamo partiti codice corso era attributo di Professore, quindi assumeva un solo valore. Dall'altra parte, effettivamente prima di rappresentare la cardinalità di Corso dovremmo informarci presso la università: dato un corso, quanti professori diversi possono insegnarlo? Qui abbiamo assunto: più di uno, da cui la cardinalità molti.

E' chiaro che se a questo vogliamo rappresentare i due schemi con un unico schema, siccome Codice Corso è rappresentato nello stesso modo nei due schemi, possiamo semplicemente sovrapporli. Vedi Figura 56. Questa attività viene anche chiamata *integrazione* dei due schemi. Vedi anche in Figura 57 lo schema integrato e i due schemi di partenza rappresentati insieme.

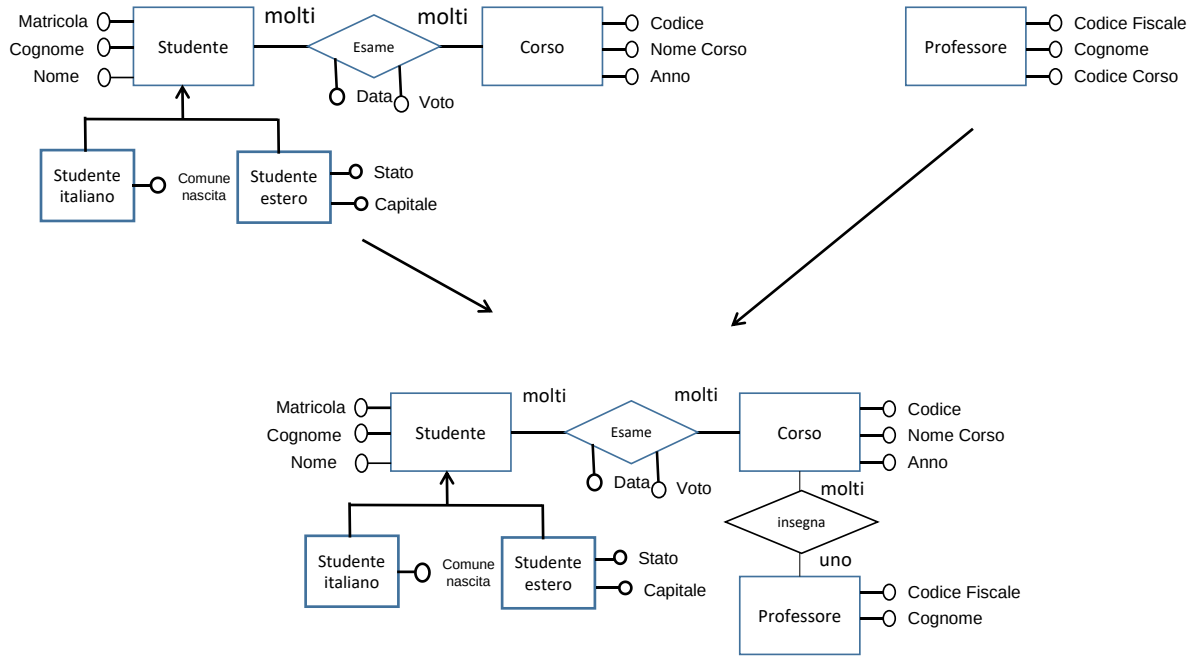


Figura 56 – Lo schema risultato della integrazione dei due schemi di partenza

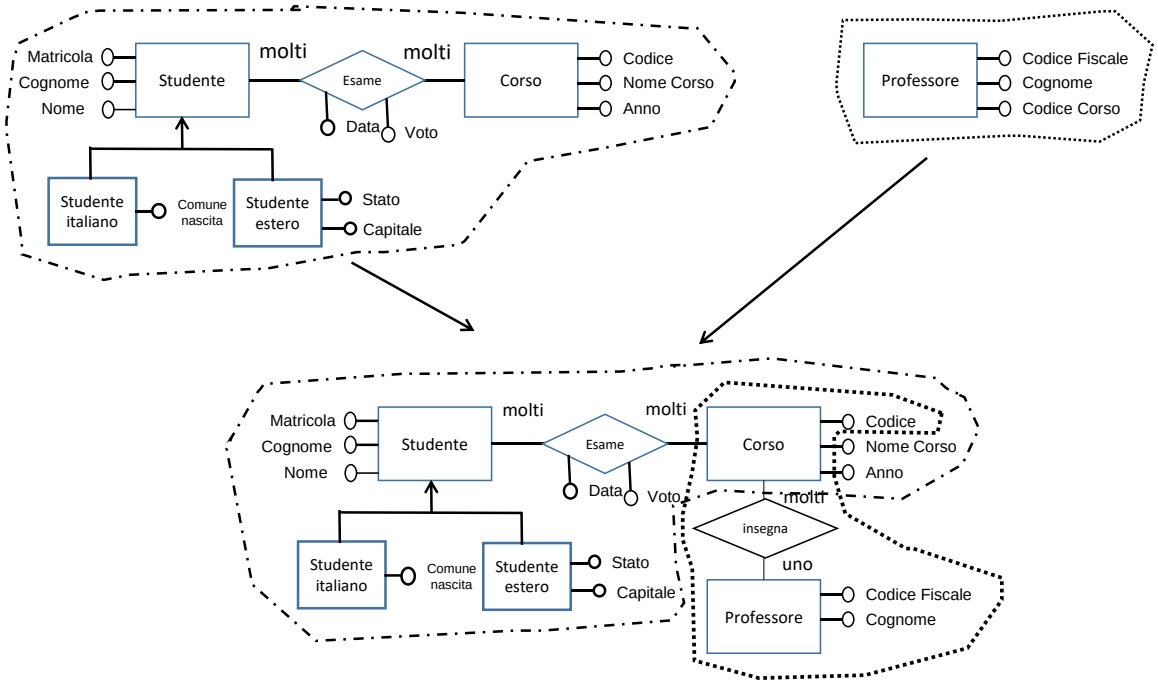


Figura 57 – I due schemi di partenza rappresentati nello schema integrato

Nota che la parola integrazione/integrare è spesso usata anche nella vita di ogni giorno, ad esempio quando si parla di migranti che vanno integrati in una comunità. Ha un significato nobile, importante, perché la integrazione porta sempre ricchezza, ad esempio ora noi sul nuovo schema integrato possiamo fare molte più interrogazioni di quanto accadeva sui due schemi separati.

Intuisco quello che dici, ma mi fai un esempio?

Certo, nel nuovo schema io posso sapere ad esempio quali studenti hanno superato l'esame del corso che insegna, e tante altre informazioni che ricostruisco navigando tra il primo schema e il secondo, e viceversa.

Ora prova tu a integrare i due schemi di figura 58. E poi guarda la soluzione nella pagina seguente.

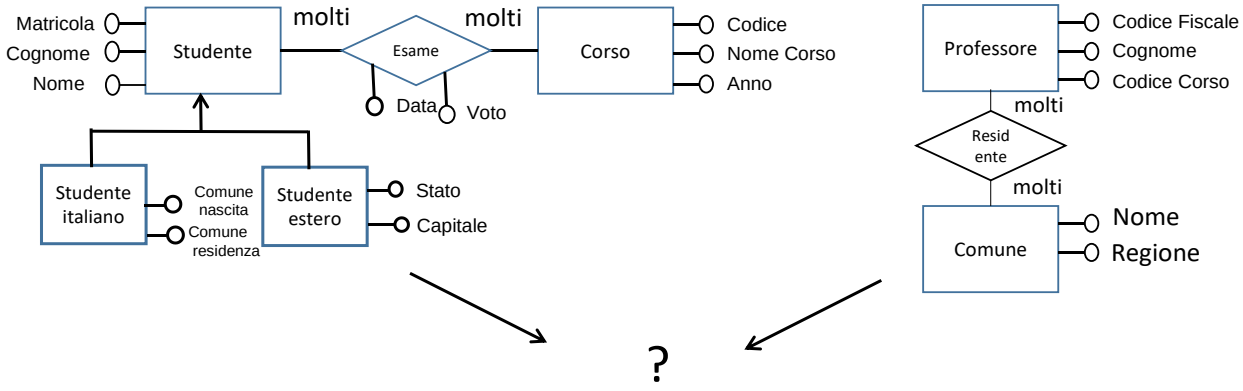


Figura 58 – Provate a integrare questi schemi…..

Ecco la soluzione, vedi Figura 59.

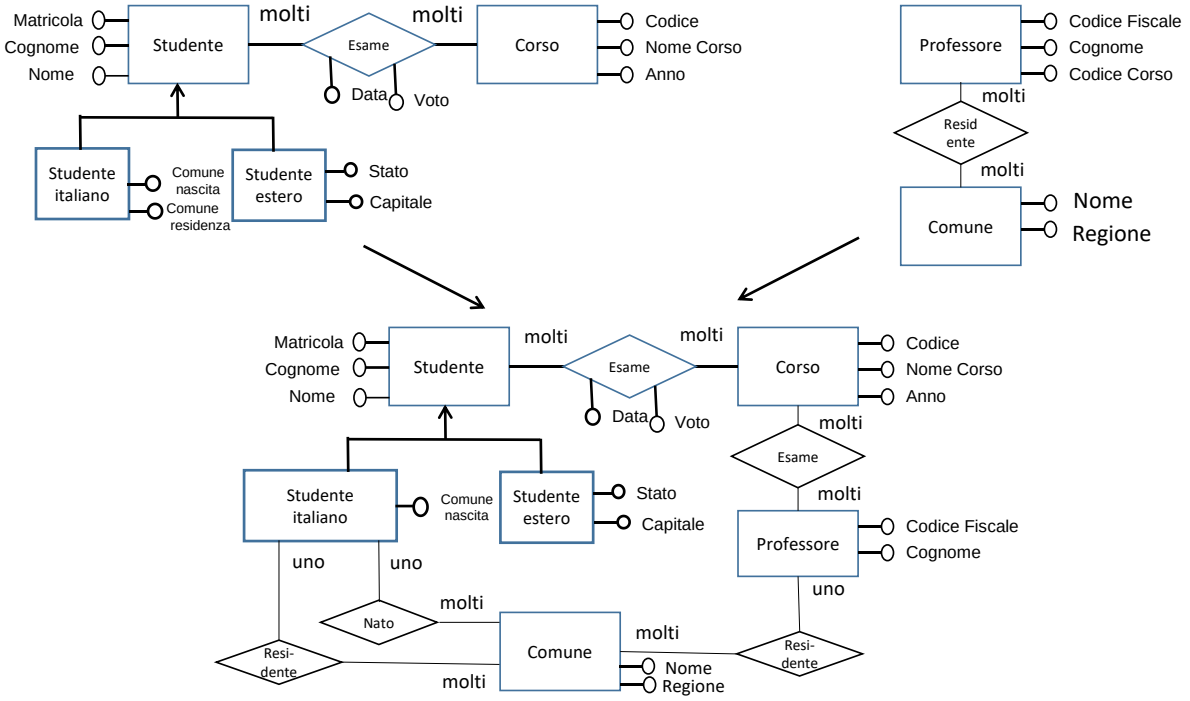


Figura 59 – Soluzione

In questo esercizio c'erano da superare alcune difficoltà. Prima di tutto, Comune era presente nello schema a destra come entità e nello schema a sinistra tramite due attributi, Comune di nascita e Comune di Residenza. La situazione è simile alla precedente, ma in questo caso Comune compare con due significati differenti nello schema a sinistra, per cui quando integriamo i due schemi, questi due "ruoli" di Comune vanno rappresentati come due diverse relazioni nello schema integrato.

Diciamo che ho capito, però, non si poteva continuare a mantenere Comune come entità, e continuare a rappresentare Comune di nascita e Comune di residenza come due attributi di Studente italiano?

Come dire, tutto si può fare, ma nel tuo schema integrato Comune è rappresentato tre volte! Mentre il suo significato è sempre lo stesso! Quindi, per raggiungere l'obiettivo di rappresentare ogni aspetto del frammento di mondo una volta sola, conviene rappresentarlo una sola volta, chiaramente come entità.

Ho capito!

Capitolo 4

La conoscenza come rete di concetti - I grafi semantici

Il modello relazionale e il modello Entità Relazione sono stati per molto tempo il modello di riferimento per le basi di dati, che sono la tecnologia utilizzata nei sistemi informativi delle aziende e delle pubbliche amministrazioni di tutto il mondo per produrre beni e offrire servizi. Nei sistemi informativi tradizionali, i dati sono utilizzati all'interno delle organizzazioni, e solo in misura limitata vengono scambiati con gli utenti finali o con altre organizzazioni.

Ad esempio, una compagnia di trasporti che venda viaggi in treno, dovrà, come ad esempio accade nel caso dell'alta velocità con Trenitalia e Italo, esporre ai clienti gli orari dei viaggi, le stazioni di partenza e di arrivo, il costo dei biglietti. Altri dati, come la anagrafe dei dipendenti, i treni in manutenzione, il bilancio della azienda, le anagrafi dei clienti che hanno una carta fedeltà, sono dati interni, spesso gelosamente nascosti alla concorrenza come i dati sui clienti.

In questo mondo fatto di sistemi informativi sostanzialmente chiusi, e di dati spesso accessibili solo a un numero limitati di soggetti, irrompe negli anni 90 il Web. Il Web è una immensa prateria in cui ciascuno di noi può condividere ciò che vuole; le reti sociali ci permettono di condividere pensieri, foto, suoni, video. Il Web è l'essenza stessa della condivisione; nel Web i dati possono essere illimitatamente condivisi, non ci sono gerarchie, tutti gli utenti sono potenzialmente parte di una comunità tra pari!

Ora, quando il Web viene utilizzato per condividere dati digitali, non possiamo immaginare che ciò possa essere fatto, banalmente, pubblicando delle tabelle di dati strutturati. Se una persona, come in Figura 60 pubblica sul Web una tabella in linguaggio italiano sui premi Nobel italiani, e in Australia, a Sydney, qualcuno pubblica una tabella in inglese sui premi Nobel in Letteratura, come si fa a collegare sul Web le due tabelle per i dati comuni, per esempio, in figura 60 i dati relativi a Luigi Pirandello e Grazia Deledda? Come si fa a capire che stiamo parlando della stessa persona, anche se i nomi degli attributi sono in lingue diverse, il cognome è prima del nome nella tabella a sinistra e dopo il nome in quella a destra, le date sono date di nascita in entrambi i casi, ma hanno un formato diverso?

Ma soprattutto, come si fa a collegare le due tabelle, visto che nel modello relazionale i collegamenti tra dati avvengono per mezzo di valori?

Premi Nobel italiani

Cognome	Nome	Ambito	Data Nascita
Deledda	Grazia	Letteratura	28091971
Natta	Giulio	Chimica	26021903
Pirandello	Luigi	Letteratura	28061867
Rubbia	Carlo	Fisica	31031934
…	……	…….	…..

← ?→

Nobel Prizes in Literature

Given Name	Last Name	Date of Birth	Place of Birth
……	…	…	…
Luigi	Pirandello	28/06/1867	Agrigento
Grazia	Deledda	28/09/1971	Nuoro
……..	….	….	….

Figura 60 - Due tabelle, una creata da un romano, l'altra creata da un abitante di Sydney

Il problema è ancora più critico quando, come in Figura 61, per rappresentare i dati sono utilizzati in un caso una tabella e nel secondo caso un testo in linguaggio naturale. Qui la rappresentazione dei dati nei due casi è diversa, in un caso dati dotasti di una struttura, nel secondo caso un testo in lingua italiana.

In questo caso si potrebbe pensare di isolare le parole nel testo che corrispondono ai termini che compaiono nella tabella, ad esempio "Luigi", "Pirandello", ma accade per molti formati di dati che siano *proprietari*, siano cioè noti solo al proprietario del formato, che il proprietario non vuole rivelare per motivi commerciali; in questi casi come facciamo?

Inoltre, se il proprietario del dato vuole essere il solo ad avere il diritto di condividere il dato con altri, come accade nel caso di copyright, per esempio per gli articoli scientifici o i libri pubblicati da molte case editrici, come sarà mai possibile condividere i dati?

Premi Nobel italiani

Cognome	Nome	Ambito	Data Nascita
Deledda	Grazia	Letteratura	28091971
Natta	Giulio	Chimica	26021903
Pirandello	Luigi	Letteratura	28061867
Rubbia	Carlo	Fisica	31031934
…	……	…….	…..

← ?→

A list of Nobel Prizes in Literature is
…..
Luigi Pirandello, born on june 28th 1867 in Agrigento,
Grazia Deledda born on september 28 1971 in Nuoro,
….

Figura 61 – Ancora più difficile….

Le cinque stelle di Berners Lee

Come accennato nel Libro primo della Enciclopedia, il primo ricercatore che si è posto il problema di stabilire regole e definire modelli per la condivisione dei dati sul Web è stato Tim Berners Lee, che nell'anno 2010 ha proposto cinque gradi di maturità per i dati pubblicati sul Web, corrispondenti a cinque livelli (chiamati anche stelle, *), essi sono:

1. Una * - Rendere disponibili i dati sul Web, in qualunque formato siano rappresentati, con una licenza open che li rende utilizzabili da tutti. Questa prima stella fa riferimento alla proprietà dei dati, che ai fini della condivisione non possono essere proprietari, ma aperti.

2. Due ** - Rendere disponibili i dati sul Web come dati strutturati, ad esempio in Excel invece che in formato scannerizzato. Questa seconda * fa riferimento alla possibilità di interrogare e elaborare i dati messi in condivisione, azione molto complessa o impossibile in un documento scannerizzato.

3. Tre *** - Rendere disponibili i dati sul Web in formato strutturato non proprietario (ad esempio Excel è un formato proprietario, mentre il formato CSV è aperto). In questo caso, l'enfasi è sulla possibilità di garantire l'accesso ai dati senza incertezza presente e futura riguardo ai diritti legali o le specifiche tecniche.

Quattro **** - Usare identificatori universali di risorsa per denotare gli oggetti descritti dai dati, così che la comunità degli utenti possa fare riferimento ai dati attraverso tali identificatori universali, collegandoli tra loro. Qui ci si riferisce al fatto che il dato abbia nel Web un "indirizzo" condiviso che permetta a tutti di accedere al dato.

La quarta stella di Berners Lee risolve il problema della identificazione unica di un dato nel Web; se ci mettiamo d'accordo in tutto il mondo su come associare un nome o indirizzo unico

nel Web, e accettiamo di condividere il formato di questo identificatore unico, allora potremo costruire un mondo virtuale di dati ciascuno accessibile da un nome o indirizzo caratteristico di quel dato.

5. Cinque ***** - Collegare i dati ad altri dati nel Web per condividerli ed integrarli nella comunità mondiale degli utenti.

La quinta stella di Berners Lee affronta e risolve il problema del collegamento trai dati che abbiamo discusso commentando le figure 58 e 59. Se voglio collegare dati sul Web devo adottare per la loro rappresentazione un modello in cui i collegamenti avvengono citando il loro nome o indirizzo unico, quello introdotto dalla quarta stella. Questo significa che nel modello ho bisogno di archi, di rami, che collegano i singoli dati, a cui possiamo associare nodi del grafo che così si viene formando; dunque non più modelli basati su valori, ma modelli a grafo, formati da nodi e da archi che collegano i nodi.

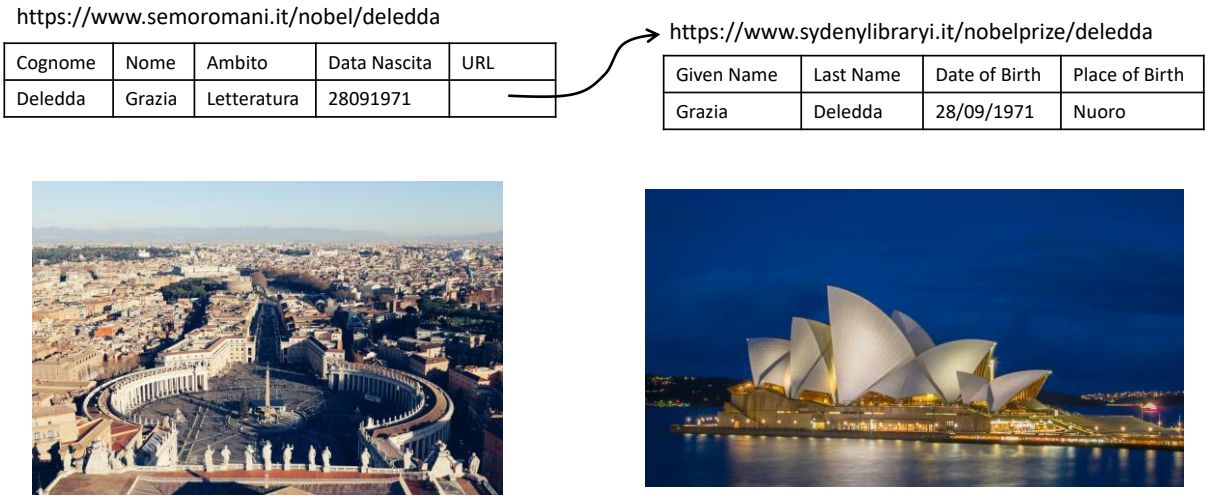


Figura 62 – Come collegare i dati nel Web.

Universal Resource Location e Uniform Resource Name

In Figura 62 mostro un esempio semplificato di ciò che nel Web è chiamato URL, Universal Resource Location, Indirizzo Universale di Risorsa, un indirizzo che, ad esempio, viene usato per indicare pagine web di siti, di giornali, di organizzazioni: una sintassi analoga viene usata per i dati, ed infatti per indicare con un termine unico ciò che può essere indirizzato da URL si usa il termine *risorsa*.

Accanto alle URL, accenno anche un altro meccanismo di identificazione nel Web, gli URN, Universal Resource Name, un nome o codice che identifichi in tutto il Web una risorsa, senza ancora esprimere la pagina Web dove si trova. Un esempio di URN è l'ISBN, International Standard *Book Number*, o "numero di riferimento internazionale del libro". Ogni libro pubblicato nel mondo ha associato un URN univico, valido ovunque.

I modelli a grafo

I grafi, che sono tipicamente utilizzati come formalismi matematici per rappresentare tantissimi tipi di reti, dalle reti di strade alle reti elettriche e reti idriche ecc., e strutture ad albero come alberi genealogici, tassonomie, ecc. vengono utilizzati per rappresentare la rete dei concetti condivisi da tutti coloro che nel mondo accettano di condividere dati e conoscenza!

Anche qui il paragone con quanto sta avvenendo nel mondo con il fenomeno delle migrazioni è calzante: alcuni innalzano muri e steccati, altri si battono perché vengano creati ponti tra popolazioni, ponti che per quanto riguarda i dati sono associabili ai collegamenti della quinta stella di Berners Lee.

I modelli a grafo hanno introdotto altre novità nel mondo dei modelli per dati.

La prima consiste nel fatto che si è fatto un grande sforzo per ampliare il più possibile la ricchezza espressiva del modello rispetto a quelli tradizionali. Abbiamo visto come il modello Entità Relazione sia decisamente più ricco del modello relazionale, attraverso la introduzione delle relazioni viste come aggregazioni di entità, e attraverso le generalizzazioni, che esprimono un legame concettuale tra entità che noi ritroviamo molto spesso nel linguaggio e nelle scienze.

Riguardo alle relazioni, i modelli cosiddetti a grafo semantico (la parola semantica, ricordo, fa riferimento al significato dei concetti del modello) arricchiscono il concetto, permettendo di definire vari tipi di relazioni cui possiamo associare proprietà e regole diverse. Un esempio lo abbiamo visto nel modello Entità Relazione con il concetto di cardinalità, che dà luogo a diversi tipi di relazioni, uno a uno, uno a molti, molti a molti. Per questa ragione i grafi semantici sono anche chiamati grafi di conoscenza o knowledge graphs, ad indicare che sono rappresentazioni progressivamente più ricche della realtà rispetto ai modelli tradizionali usati nelle basi di dati.

Nella Figura 63 vediamo un grafo semantico in cui sono rappresentate essenzialmente *classi*, che corrispondono alle Entità del modello Entità Relazione: la classe delle piante, quella degli animali, quella dei mammiferi ecc. Le classi sono rappresentate mediante delle ellissi.

Nel grafo in Figura 63 ritroviamo la generalizzazione espressa dalla relazione “è un”; una seconda relazione, la relazione “ha” che lega Faggio a Foglie e Giraffa a Zoccoli, esprime piuttosto una aggregazione, a cui possiamo associare diverse proprietà, come ad esempio, nel caso di Faggio, il fatto che possa averle o non averle, a seconda della stagione, e nel caso di Giraffa il numero degli zoccoli.

Non approfondisco questo aspetto, che puoi studiare meglio nei testi citati negli approfondimenti, ma mi preme dire che quando queste nuove proprietà sono espresse in un linguaggio di tipo logico, quindi con una semantica rigorosa, allora i grafi semantici (o grafi di conoscenza) prendono il nome di *ontologie*.

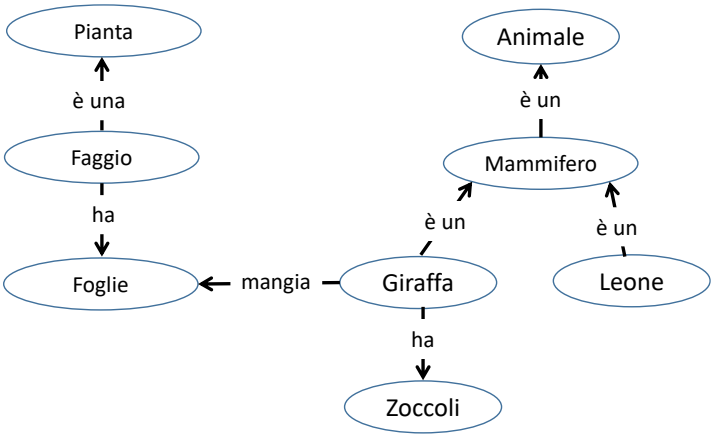


Figura 63 – Un grafo semantico

Un altro aspetto che caratterizza il modo in cui nel Web sono utilizzati i grafi semantici riguarda lo schema e i valori. Schema e valori nel modello relazionale sono rigorosamente e inscindibilmente legati tra loro; non posso avere solo schemi di tabelle senza valori (anche chiamate istanze), e non posso avere valori/istanze senza schema.

Ora, poiché il grafo semantico è creato e popolato nel tempo aggiungendo dinamicamente sia classi che istanze, in modo non necessariamente strutturato come avviene nelle basi di dati, ovvero, come vedremo tra poco, il grafo semantico nasce dal collegamento di grafi semantici preesistenti, non si vede perché non possano esserci classi senza istanze e istanze senza classi.

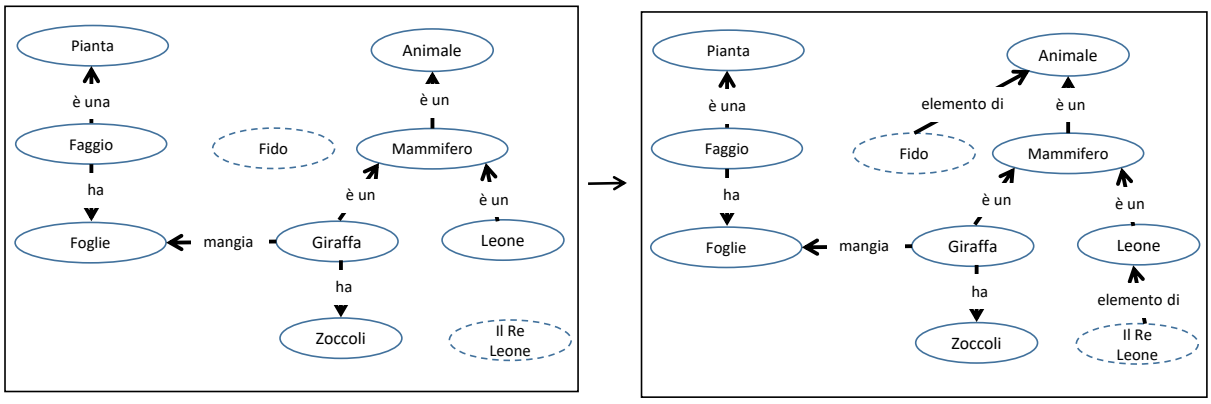


Figura 64 – Come si possono aggiungere dinamicamente istanze a classi

Supponiamo ora, vedi Figura 64, che al grafo semantico di Figura 63 vengono aggiunte due istanze, Fido e Il Re Leone. Il Re Leone, anche se immaginario, è chiaramente una istanza della classe Leone, mentre per quanto riguarda Fido supponiamo di sapere che sia una istanza di cane, il cane di un nostro amico.

Attenzione! Nel libro Primo della Enciclopedia gli esempi di grafi semantici che rappresentavano personaggi dei Demoni rappresentavano solo istanze, graficamente descritte da ellissi con linee chiuse. In questo capitolo rappresentiamo grafi semantici con

classi e istanze; le classi sono graficamente descritte da ellissi con linee chiuse, mentre le istanze sono graficamente descritte da ellissi con linee tratteggiate.

Questa conoscenza, esterna al grafo semantico, propria della nostra esperienza e percezione del mondo, ci permette di arrivare alla conclusione che possiamo arricchire il grafo semantico aggiungendo due collegamenti di tipo "elemento di" tra Il re Leone e la classe Leone, e tra Fido e Animale.

Un attimo! Avevi detto che Fido era una cane, e non mi puoi dire che hai fatto uso di tutto quello che sapevi, collegando Fido con Animale, che è molto più generale di Cane!

Vedo che sei sempre più attento! Hai ragione. In Figura 64 ho fatto una scelta conservativa; in pratica ho pensato: quale è il concetto (o i concetti) già presente nello schema a cui posso collegare Fido? Ho guardato lo schema, e ho associato Fido a Animale….

Continuo a non capire! Faccio questo ragionamento, seguimi tu questa volta: so che Fido è un cane, e lo so perché tu lo hai detto fin dall'inizio; dunque, posso vederlo come elemento di Mammifero.

Non solo, ma posso essere un po' più coraggioso e riversare nel grafo semantico tutta la conoscenza che ho su Fido; posso, cioè introdurre una nuova classe Cane.

Questa classe Cane la posso poi collegare a Mammifero con una relazione "è un" e infine collegare Fido al concetto Cane con una relazione "elemento di". O no?

Che bello! Sta accadendo ciò che mi è accaduto qualche volta nei miei corsi universitari, la inversione dei ruoli tra me e gli studenti, un momento abbastanza magico in cui lo studente ne sa più di me, e mi corregge e mi propone una soluzione più completa e di migliore qualità della mia. Complimenti, sono totalmente d'accordo con te!

Riportiamo in Figura 65 la soluzione che hai proposto.

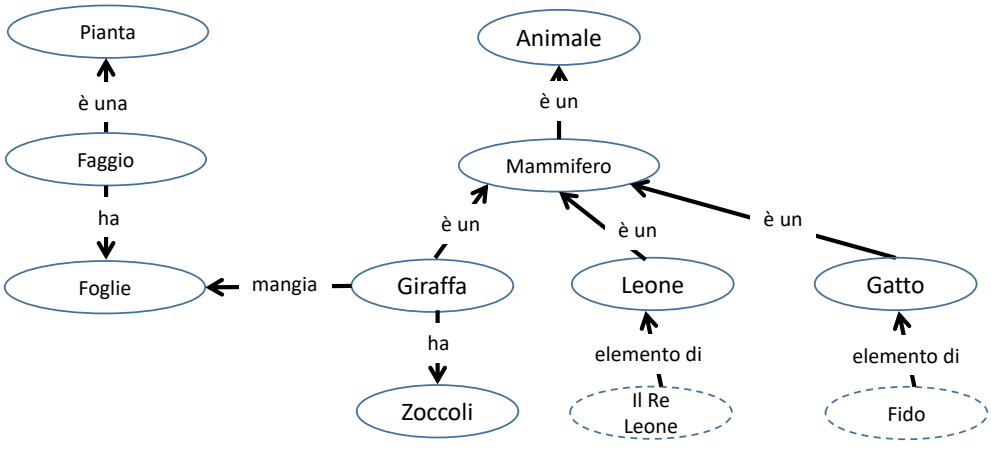


Figura 65 – Un grafo semantico "migliore" del precedente

Ho capito. Senti ma quella attività consistente nell'integrare due schemi che mi hai fatto vedere nel caso del modello Entità Relazione si può fare anche con i grafi semantici?

Certamente!

Bene, lo pensavo anche io, ma ho capito bene che in questo caso non si integra solo lo schema, si integrano anche gli elementi, le istanze?

Certamente, si integrano anche le istanze, perché nei grafi semantici non c'è più distinzione. Ad esempio, in Figura 66 noi abbiamo due grafi semantici, uno in lingua italiana e uno in lingua inglese. Indagando un po' possiamo arrivare alla conclusione che Leone e Lyon siano due classi che hanno lo stesso significato, e lo stesso per le loro istanze "Re Leone" e "King Lyon". Ebbene, a questo punto possiamo integrare i due grafi semantici, producendo il grafo che compare nella parte bassa della Figura 66.

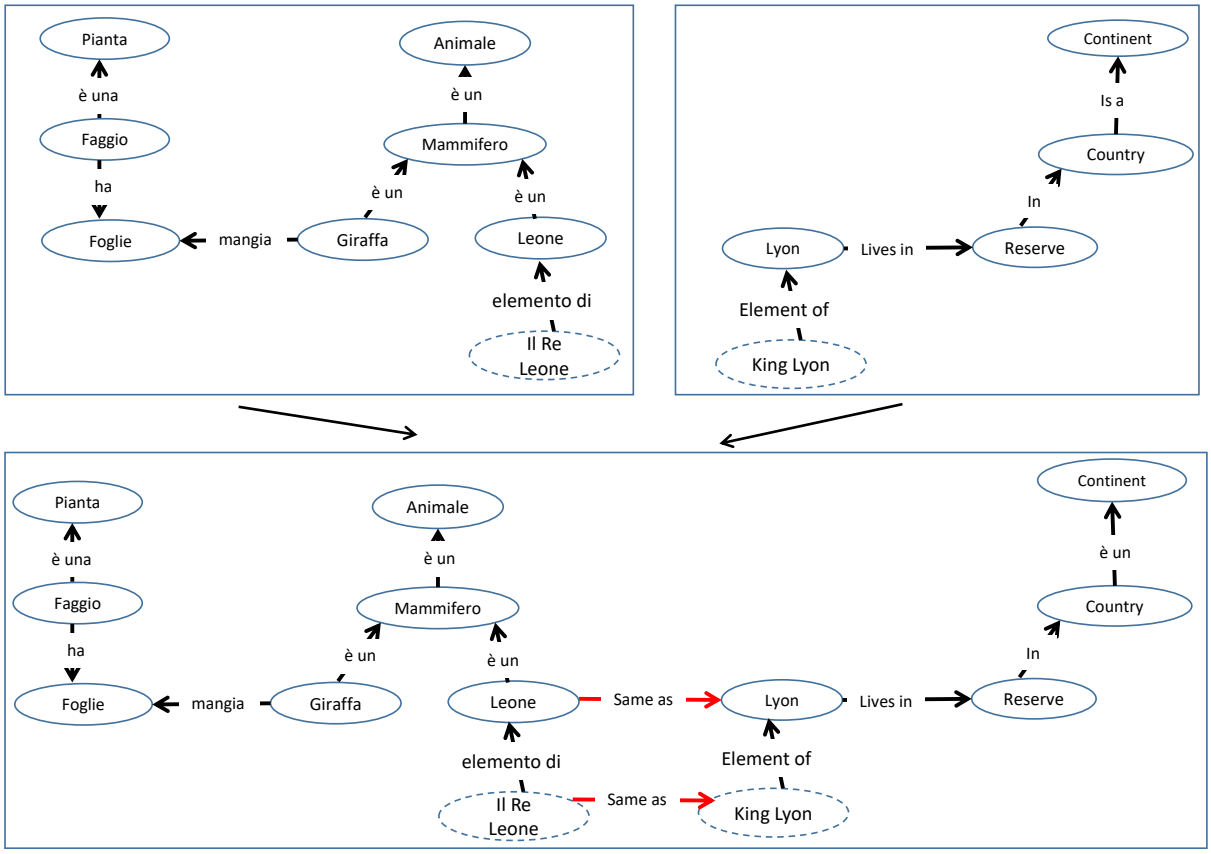


Figura 66 – Come collegare grafi semantici

Il linked open data Cloud

Berners Lee propone le sue cinque stelle nel 2010, ma i modelli a grafo semantico hanno qualche anno in più. Più o meno dal 2007 molti ricercatori hanno cominciato a rappresentare i *loro data set* mediante un modello a grafo. Quando uso il termine *data set* intendo un termine molto generale per esprimere un insieme di dati come una rubrica, un vocabolario,

una anagrafe, un elenco di cantanti, una collezione di CD, descrittivi, come posso dire, di qualunque aspetto della realtà rappresentabile mediante dati.

Coloro che nel mondo hanno contribuito a rappresentare i dataset con il modello a grafo, e li hanno pubblicati sul Web, non si sono limitati a questo, si sono posti il problema di collegare questi dataset, per arrivare a una rappresentazione veramente condivisa di una conoscenza collettiva.

Questo ha portato a costruire una immensa base di conoscenza che nel 2007 (vedi Figura 67) era composta da 11 dataset, e nel 2020 da 1260 dataset, con decine di miliardi di terne (anche dette triple) di nodi-archi-nodi.

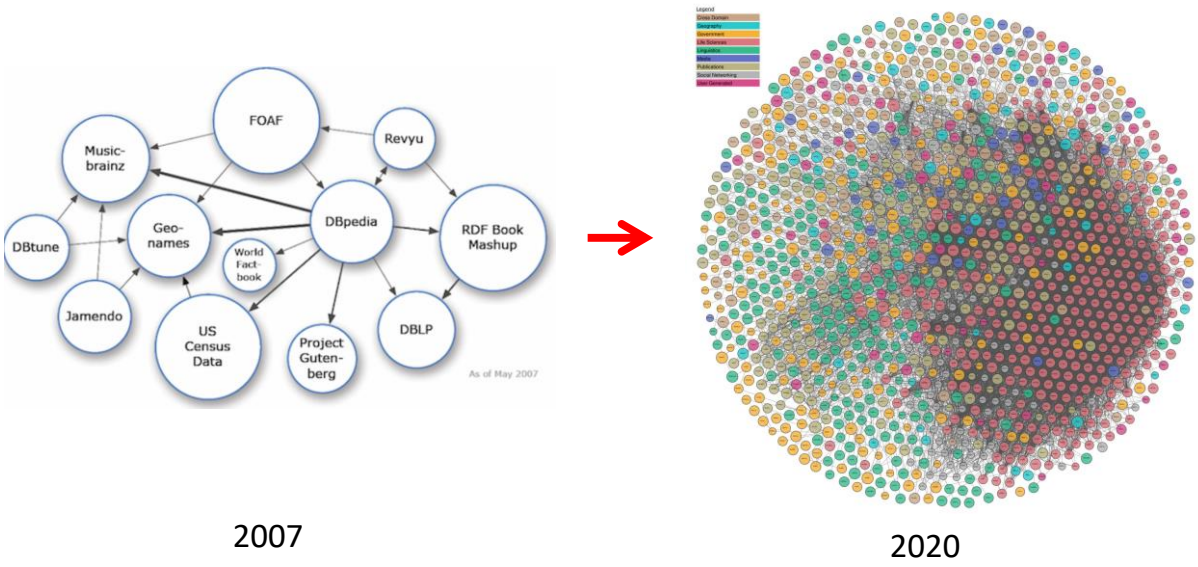


Figura 67 - Evoluzione del linked open data cloud

Come si vede nel grafo del 2007, faceva già parte dell'Open Linked data Cloud il dataset DBPedia (vedi https://wiki.dbpedia.org/), un progetto nato nel 2007 con lo scopo di estrarre informazioni strutturate da Wikipedia (https://it.wikipedia.org/wiki/ che probabilmente conoscete) e pubblicarle sul Web come Linked Open Data.

Le risorse del Web semantico: DBPedia e Wikipedia

Per capire la importanza che hanno assunto DBPedia e QWiipedia nello sviluppo del *Web semantico*, cioè quella parte del Web formata da risorse descritte con modelli di tipo a grafo semantico, dobbiamo un momento tornare agli esempi fatti poco fa di modifica incrementale di un grafo semantico (Figura 64 e 65) e di integrazione di due grafi semantici (Figura 66). Quando abbiamo aggiunto nodi o abbiamo collegato nodi, ciò è stato il risultato di un nostro ragionamento, di una nostra decisione.

Tra i ricercatori che si sono appassionati alla costruzione di dataset rappresentati mediante grafi semantici collegati tra loro nel Linked open data Cloud è emersa una idea, un progetto:

perché non costruiamo algoritmi che, utilizzando il Linked open data Cloud, permettano di automatizzare attività che in genere sono svolte da umani?

A cosa ti riferisci con "attività che sono svolte da umani?"

Per esempio, l'aggiunta di elementi o concetti a un grafo semantico (Figura 64 r 65), ovvero la scoperta di collegamenti tra due grafi semantici (Figura 67), ovvero ancora la trasformazione di tabelle o testi in grafi semantici per favorire la integrazione di dati.

Invece che fare affermazioni generali, non potresti farmi vedere degli esempi?

Va bene, d'accordo. Anzitutto, guarda la Figura 68. Sono rappresentate un piccolo frammento di DBPedia e un pezzetto della voce in Inglese sui premi Nobel.

Il frammento di DBPedia ci fa vedere alcuni dati che descrivono il concetto NobelPrize. Se guardi con attenzione, si dice anzitutto che Nobel Prize ha una superclasse Award, nella nostra terminologia *è un* Award. Questa relazione di generalizzazione tra Nobel Prize e Award è un piccolissimo frammento dellì'insieme di generalizzazioni definite in Wikipedia, e che puoi consultare sul sito www.DBpedia.com.

Sempre guardando il frammento di DBpedia, scopri che una ulteriore serie di dati presente riguarda un insieme di labels o etichette, che ci forniscono i nomi del concetto Nobel Prize in varie lingue tra cui l'italiano e l'inglese.

Guardando invece la parte della figura dedicata a Wikipedia, troviamo un frammento della voce Nobel Prize in inglese.

DBPedia

NobelPrize (Show in class hierarchy)

Label (it):	Premio Nobel
Label (el):	Βραβείο Νόμπελ
Label (ga):	Duais Nobel
Label (fr):	Prix Nobel
Label (de):	Nobelpreis
Label (es):	Premio Nobel
Label (ja):	ノーベル賞
Label (en):	Nobel Prize
Label (nl):	Nobelprijs

Super classes: Award

Properties on *NobelPrize*:

Name	Label	Domain	Range	Comment

Wikipedia

The **Nobel Prize** (/ˈnoʊbɛl/, *NOH-bel*; Swedish: *Nobelpriset* [nʊˈbêlːˌpriːsɛt]; Norwegian: *Nobelprisen*) is a set of annual international awards bestowed in several categories by Swedish and Norwegian institutions in recognition of academic, cultural, or scientific advances. The will of the Swedish chemist, engineer and industrialist Alfred Nobel established the five Nobel prizes in 1895. The prizes in Chemistry, Literature, Peace, Physics, and Physiology or Medicine were first awarded in 1901.[1][3][4] The prizes are widely regarded as the most prestigious awards available in their respective fields.[5][6][7]

In 1968, Sveriges Riksbank, Sweden's central bank, established the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. The award is based on a donation received by the Nobel Foundation in 1968 from Sveriges Riksbank on the occasion of the bank's 300th anniversary. The first Prize in Economic Sciences was awarded to Ragnar Frisch and Jan Tinbergen in 1969. The Prize in Economic Sciences is awarded by the Royal Swedish Academy of Sciences, Stockholm, Sweden, according to the same principles as for the Nobel Prizes that have been awarded since 1901.[8] However, as it is not one of the prizes that Alfred Nobel established in his will in 1895, it is not a Nobel Prize.[9]

Figura 68 – Due frammenti di DBPedia e Wikipedia

Ora ti faccio vedere come posso utilizzare una estensione del frammento di DBPedia per trovare le relazioni semantiche tra la tabella strutturata in italiano e il testo in inglese che compaiono in Figura 68. Mi segui?

Si, ma mi piacerebbe che arrivassi rapidamente alle conclusioni, e non girassi in tondo!

Eh, non è facile rendere semplici questioni che semplici non sono, devi avere un pò di pazienza…..

Guarda ora la Figura 69, in cui ci concentriamo su DBPedia, più avanti prenderemo in considerazione anche Wikipedia. Osserviamo i due frammenti di dati, costituiti dalla riga nella tabella a sinistra e dal testo a destra circondati da cornici.

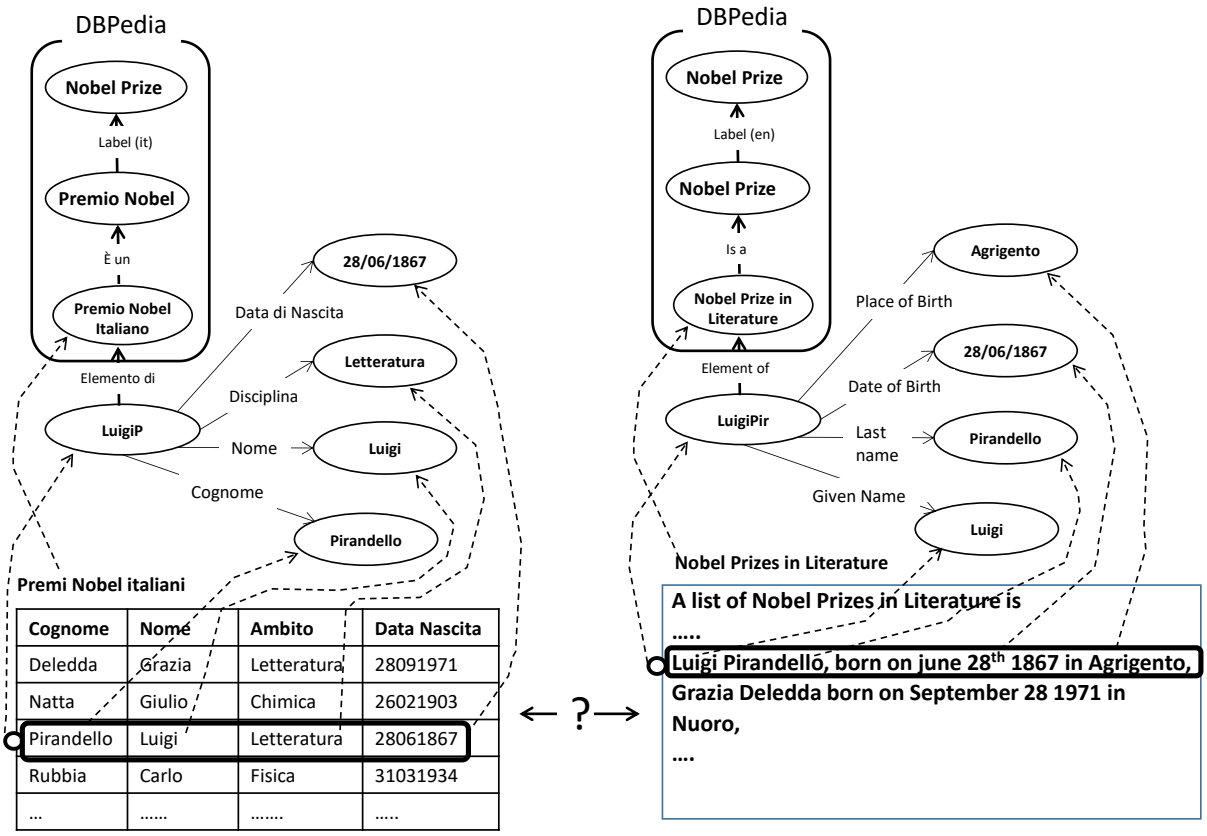


Figura 69 – Riga della tabella e testo e loro trasformazioni in grafo semantico

In Figura 69 il nostro obiettivo è trasformare la riga della tabella e il testo a destra in due grafi semantici, così che poi un programma automatico potrà confrontarli per capire se stiamo parlando dello stesso Luigi Pirandello oppure di due Luigi Pirandello diversi.

Quanto alla riga della tabella, un programma automatico deve riconoscere anzitutto che "Pirandello" è un elemento di Cognome; questo intuitivamente non è difficile, perché "Pirandello" nella tabella è un valore che compare nella colonna Cognome. Analogamente per "Luigi", per "Letteratura" e per "28061867".

Nota che la stringa "28061867", oltre che venire riconosciuta come data, viene anche trasformata in un formato standard, in cui giorno, mese e anno vengono rappresentati da numeri separati dal simbolo "/".

E perché questo?

Beh, per avvicinare la rappresentazione delle date ad una rappresentazione standard che possa essere condivisa da tutti e possa facilitare le attività di integrazione.

Viene anche creato un elemento “LuigiP” che rappresenta la riga nella sua interezza. La riga viene poi messa in relazione con classi di DBPedia, tra esse la seconda e la terza (premio Nobel, Nobel Prize) sono menzionate nella Figura 69, e la prima “premio Nobel italiano” è costruita mediante una funzione di elaborazione in linguaggio naturale che opera sul nome della tabella relazionale “Premi Nobel italiani”, normalizzandolo nel singolare.

Per la frase in linguaggio inglese il discorso è un po' più complicato. In questo caso il programma dovrà:
1. riconoscere nella frase le singole parole (es. “Luigi”) o gruppi di parole che sintatticamente sono legate tra di loro (es. June 28th 1967),
2. associare a ciascuna di esse un nodo del grafo semantico, e
3. associare a tali nodi, sulla base di risorse disponibili o in DBPedia o in altre parti del Linked Open Data Cloud, delle relazioni che le legano all’elemento LuigiPi, associando a tali relazioni nomi in inglese come Last Name, ecc.
4. Infine il titolo del documento va associato a una classe “Nobel prize in Literature”.

Mi segui o ti sei perso?

Mah, più o meno ti seguo…..devo rileggermi il tutto per essere sicuro di aver capito…
In ogni caso, è un po' faticoso seguire tutti i singoli passaggi, sembra che ti diverti a complicare le cose con tante piccole questioni.

Sei un po' ingeneroso! Io non pretendo di spiegare il tutto dandoti una competenza completa sulla materia, tutto quello che voglio fare e mostrarti un percorso che ti convinca che il programma automatico si può produrre, così che anche tu possa avere confidenza di essere in grado di fare tutto ciò….

OK, grazie.

Bene, ora facciamo un altro passaggio. Finora abbiamo prodotto due gradi semantici, ora dobbiamo confrontarli. Questo è ciò che faccio nella Figura 70. Anche questa attività può essere affidata ad un programma software, che, esplorando i due grafi e le parti comuni in DBPedia, può tracciare un insieme di legami semantici.

Possiamo immaginare di iniziare a cercare legami a partire dai concetti più generali, quelli in DBPedia; come abbiamo visto nel frammento di Figura 68, in DBPedia sono rappresentate per il concetto “Nobel Prize” le diverse denominazioni in varie lingue, tra cui inglese e italiano. Quindi un programma software che accede a DBPedia, con interrogazioni simili a quelle che ho mostrato nella prima parte del capitolo, e su cui non entro ora nel merito, può concludere che “Premio Nobel” e “Nobel Prize” sono lo stesso concetto, associando ai due concetti la relazione “same as”.

Analogamente per le altre coppie di concetti in Figura 70.

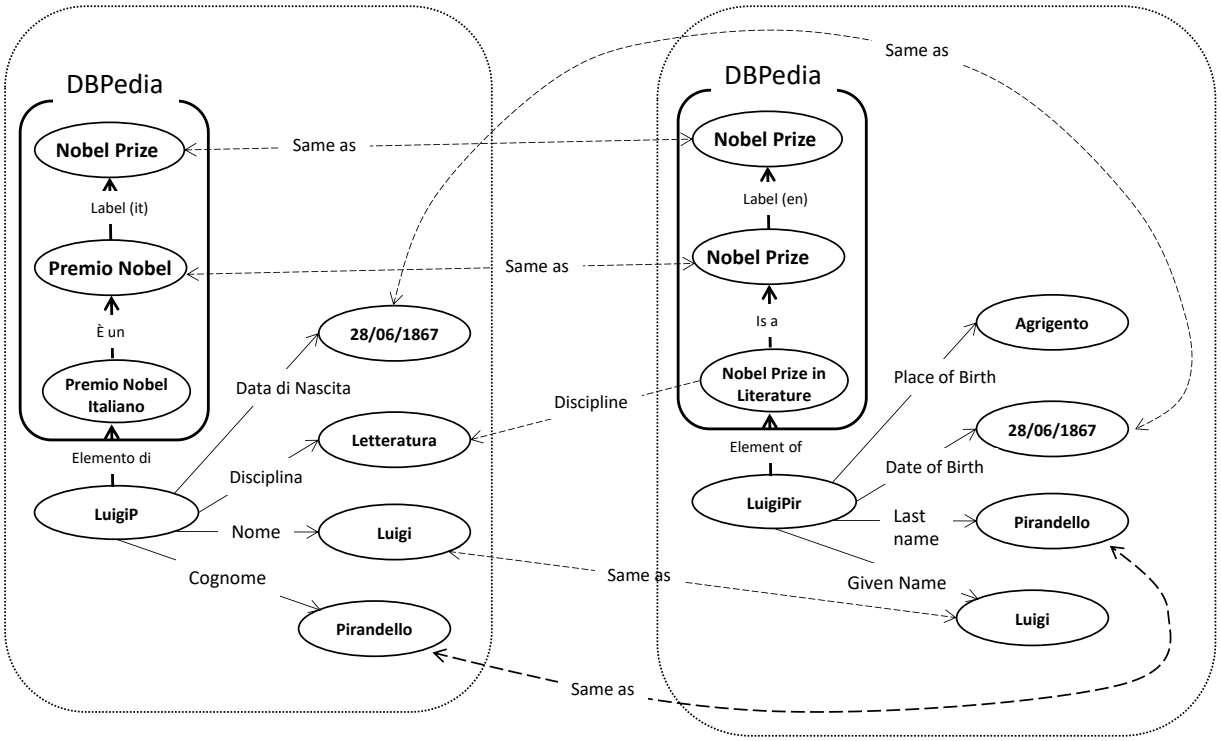


Figura 70 – Corrispondenze tra i due grafi semantici

Ora non ti voglio stancare con la descrizione dettagliata di come possiamo scoprire tutte le corrispondenze di Figura 70, mi pacerebbe solo che tu fossi convinto che esplorando i due grafi effettivamente queste corrispondenze sono sensate. Approfondisco solo la relazione di tipo "Discipline" tra "Nobel Prize in Literature" e Literature": questa relazione compare in DBPedia e il compito del programma è quello di individuarla e inserirla come relazione tra "Nobel Prize in Literature" e Literature" nell'ambito dei due grafi associati alla riga della tabella e al testo.

Che fatica!

Certo, è una fatica, ma non la devi fare tu, la fa il programma. La grande novità rispetto al passato è che ora i programmi software trovano nel Web tutta questa conoscenza, che non deve essere fornita loro da un essere umano.

Se è tutto chiaro, vorrei fare un secondo esempio. La conoscenza costituita da Wikipedia può essere sfruttata per attività di disambiguazione del linguaggio naturale.

Cosa intendi per disambiguazione?

Eh, guarda ad esempio la Figura 71, e leggi il testo a sinistra, che possiamo immaginare essere un frammento di una informativa anonima che giunge a un investigatore. Il testo fa riferimento a Palermo, un nome che può essere associato a due località, Palermo la città siciliana, ovvero Palermo visto come quartiere di Buenos Aires.

Come fa un programma software a eliminare la ambiguità tra i due significati? Capita spesso che noi usiamo termini cui possiamo associare differenti significati, in questo caso differenti località….

Ecco, ci può venire in aiuto Wikipedia, che associate a Palermo ha due voci, che corrispondono alla città siciliana e al quartiere di Buenos Aires.

Nella figura 71 vediamo un semplice algoritmo di disambiguazione in cui isoliamo tutte le parole rilevanti nel testo e le cerchiamo nelle due voci di Wikipedia. Dopodichè, siccome abbiamo trovato tutte e tre le parole nella voce che corrisponde a Palermo in Sicilia e solo due nella voce Palermo come quartiere di Buoenos Aires, possiamo arrivare alla conclusione che è più probabile che il testo faccia riferimento a Palermo in Sicilia.

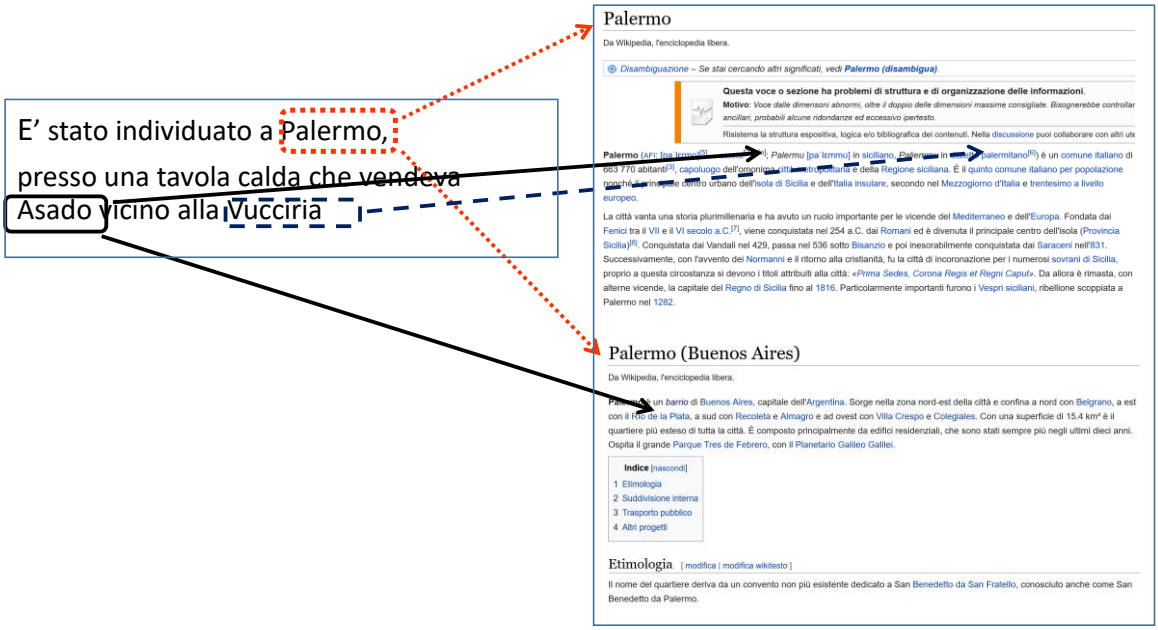


Figura 71 – Palermo in Sicilia o Palermo quartiere di Buenos Aires?

Interessante! Però, attenzione! Hai detto "è più probabile", non hai detto "è certo", come si fa a raggiungere la certezza?

Eh, la certezza non è di questo mondo! Voglio dire che, come accade nelle attività umane, è ben difficile che noi arriviamo alla assoluta certezza sulla validità di una conclusione mediante un programma automatico. Comunque, se vuoi una discussione più approfondite su questi problemi, ti rimando al Capitolo/Libro 5 sulla qualità dei dati.

Che lunga cavalcata che abbiamo fatto sui modelli dei dati! Siamo partiti vedendo i modelli come strumenti per rappresentare il mondo, e abbiamo finito vedendo i modelli come strumenti per ragionare sul mondo…

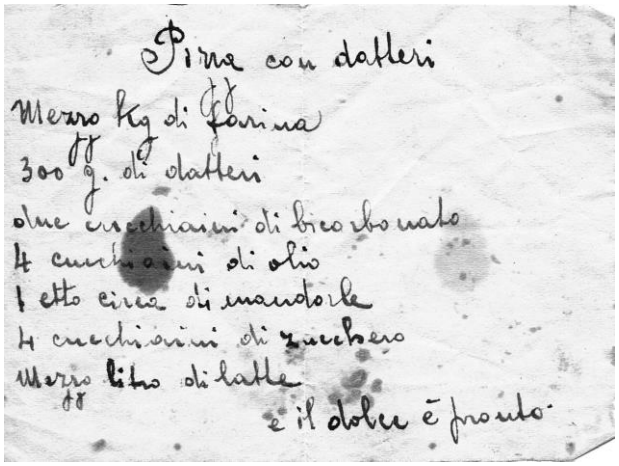
Capitolo 5

I limiti dei modelli dei dati

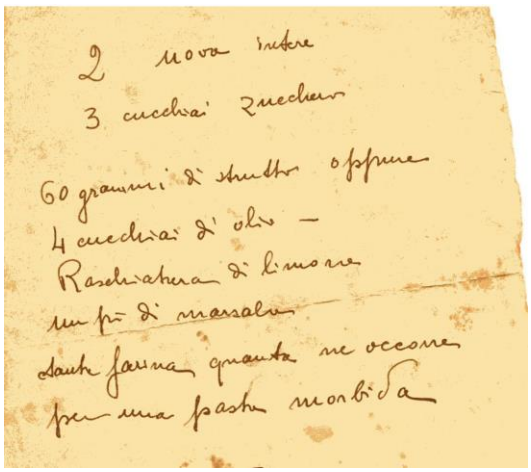
I modelli dei dati che abbiamo visto nel capitolo, anche nelle versioni più espressive, presentano significativi limiti nel rappresentare un frammento di mondo. Per spiegarmi meglio, tornerò in questa sezione al modello relazionale, modello che viene usato diffusamente nei sistemi informativi delle pubbliche amministrazioni e delle aziende, tanto è che la tecnologia delle basi di dati è quella più utilizzata nei sistemi informativi nel mondo.

Le vecchie ricertte di cucina

Quando ero piccolo, mi piaceva rovistare nei cassetti della casa dove abitavamo con i miei genitori e i miei fratelli. Quando con i miei fratelli chiudemmo la casa dei miei genitori, perché ormai scomparsi, rovistando in cantina in una delle mie ricerche trovai un fascicolo con tante ricette che nella prima parte del secolo scorso furono scritte su fogli di carta, come si faceva un tempo, dalla mia mamma, dalle mie zie, e forse anche dai miei nonni. Mostro due di queste ricette nella Figura 72.



Ricetta della Pizza con datteri



Ricetta della ??

Figura 72 – Due antiche ricette di cucina: si riescono a descrivere per mezzo di tabelle?

Se proviamo a modellare queste due ricette con il modello relazionale, ci troviamo in difficoltà con alcune espressioni, come vediamo in Figura 73.

Pizza con datteri

Nome ingrediente	Unità di misura	Quantità
Farina	kilogrammo	1/2
Datteri	grammo	300
Bicarbonato	cucchiaino	2
Olio	cucchiaino	4
Mandorle	etto	1 circa
Zucchero	cucchiaino	4
Latte	litro	1/2

??

Nome ingrediente	Unità di misura	Quantità
Uova intere	numero	2
Zucchero	cucchiaio	3
Strutto	grammi	60
Olio	cucchiaio	4
Limone	raschiatura	??
marsala	cucchiaino	Un pò
farina	litro	Quanta ne occorre per una pasta morbida

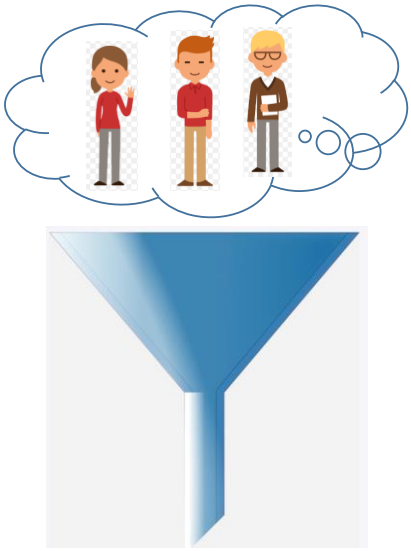
Figura 73 – Difficile tentativo di rappresentare le ricette con il modello relazionale

Cosa significa infatti “un etto circa”, cosa significa “quanta ne occorre per una pasta morbida”? Quanto significa “un po'”? E che nome diamo alla ricetta a destra? Siamo costretti a “forzare” il frammento di mondo descritto nella ricetta, ovvero ad approssimarne il significato, ovvero ancora a rinunciare a rappresentarlo.

Certo, possiamo sempre rappresentare un etto circa con un valore “1 circa”, possiamo riportare la frase “quanta ne occorre per una pasta morbida”, ma abbiamo solo spostato il problema; abbiamo bisogno di modelli più espressivi di quelli che usiamo nelle basi di dati e nel Web semantico, modelli più espressivi su cui la ricerca nel mondo sta lavorando alacremente.

Insomma, ogni volta che usiamo un modello per descrivere un frammento di mondo, noi effettuiamo una approssimazione, è come se usassimo un imbuto semantico (vedi la Figura 74) che necessariamente isola alcune caratteristiche degli oggetti rappresentati, e ne nasconde moltissime altre.

Se ad esempio le persone di cui rappresentiamo alcune caratteristiche sono coinvolte in un concorso, per valutarle non ci possiamo basare sui soli attributi di Figura 74, e scusate la banalità dell’esempio, ma ne dovremmo raccogliere e rappresentare molte altre. E a seconda di quelle che rappresentiamo, e di quelle che non rappresentiamo, potremmo applicare metodi di valutazione che volontariamente o involontariamente introducono discriminazioni tra i candidati. Il problema che ho appena toccato è diventato ancora più rilevante nella nostra epoca in cui si sono diffusi moltissimo le tecniche predittive e decisionali basate sul machine learning e sulla intelligenza artificiale, di cui discuteremo in un capitolo successivo.



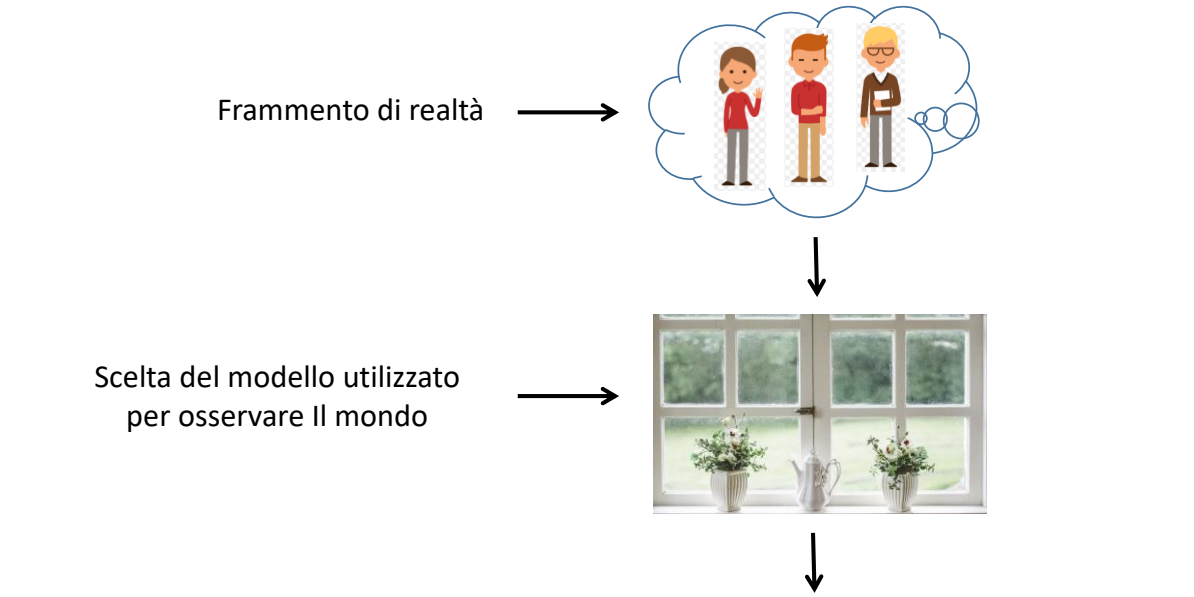
Persone Residenti a Milano

Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 74 – L'imbuto semantico che usiamo sempre
nel rappresentare mediante dati un frammento di mondo

Concetti introdotti in questo Libro

In questo libro abbiamo discusso dei modelli dei dati digitali. In questa sezione riassumo i concetti visti nel capitolo. Guardate nella Figura 75, in cui mostriamo come un frammento di realtà può essere descritto per mezzo di un modello di dati, i principali concetti introdotti.



Cognome	Nome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
Bini	Carlo	Pescara	7/6/1949	Via Rossini 18	20127
Verdi	Giulio	Roma	9/7/1999	Via Rossini 18	20127
Rossi	Laura	Milano	20/12/1996	Via Verdi 31	20123

Figura 75 - I principali concetti introdotti nel capitolo

Ogni termine che useremo nella sintesi che segue e evidenziato in ***neretto corsivo*** ha una definizione nella lista successiva.

Per osservare e rappresentare un frammento di mondo, la Scienza dei dati propone di usare ***modelli dei dati***; attraverso una attività di ***modellazione*** noi rappresentiamo il frammento di mondo per mezzo di ***strutture di rappresentazione*** che cambiano da modello a modello. Nel ***modello relazionale*** queste strutture sono le ***tabelle*** e gli ***attributi***; le tabelle sono ***classi*** di ***istanze*** di oggetti del mondo, ogni istanza corrisponde a una riga della tabella. Gli attributi, che rappresentano proprietà degli oggetti del mondo, sono classi di **valori**, e ogni valore è il dato che compare in una cella di una riga. Un insieme di tabelle è chiamato ***base di dati***, composta di a. un insieme di ***schemi*** di tabelle e b. un insieme di ***istanze*** di tabelle.

Ogni oggetto del mondo rappresentato in una tabella deve avere una ***identità*** unica nella tabella, nel modello relazionale tale identità è costituita dalla ***chiave***. Una base di dati può essere interrogata mediante un linguaggio di ***interrogazione***, nel modello relazionale il

linguaggio SQL, che esprime le proprietà dei dati che si vuole estrarre. I linguaggi di interrogazione possono essere ***operazionali*** o ***dichiarativi***.

Una tabella può essere decomposta in più tabelle, e in questo caso una proprietà importante della ***decomposizione*** è di essere ***senza perdita di informazione***.
Oltre al modello relazionale, vi sono altri modelli dei dati cui corrispondono diverse strutture di rappresentazione, come il ***modello gerarchico*** e il ***modello Entità Relazione***, che è caratterizzato da una ***rappresentazione grafica*** che ne aumenta la comprensibilità. Nel modello Entità Relazione sono rilevanti le **cardinalità** delle relazioni.

Tra le classi e le istanze di classi rappresentate nei vari modelli con le strutture di rappresentazione sussistono diversi meccanismi di composizione (di classi e istanze in altre classi e istanze), tra cui la **aggregazione** e la **generalizzazione**. Le classi coinvolte in una generalizzazione godono della proprietà di ereditarietà.

Un'ulteriore insieme di modelli sono i ***modelli a grafo semantico***, che sono soprattutto usati nel descrivere i dati condivisi sul Web. Perché i dati siano condivisi sul Web devono rispettare le cinque stelle (o livelli) di Tom Berners Lee; l'esempio più significativo di dati condivisi e creati in modo collaborativo sul Web è il ***Linked open data cloud***. Sul Web sono state create attraverso una modalità collaborativa altre risorse di dati, come **DBPedia** e **Wikipedia**.

Definizioni

1. ***Modello dei dati*** - E' una rappresentazione in cui un frammento di mondo è descritto per mezzo di un insieme di strutture, o ***strutture di rappresentazione***, che, ad esempio, nel modello relazionale, sono le tabelle e gli attributi.

2. ***Modellazione*** – Una attività in cui noi osserviamo un frammento del mondo reale e lo rappresentiamo mediante un insieme di dati, organizzati, ad esempio, in una tabella, in un testo, in un grafo.

3. ***Struttura di rappresentazione*** – Una particolare modalità con cui possiamo organizzare un insieme di dati, come, ad esempio, una tabella, una riga di una tabella, una colonna di una tabella. Le strutture di rappresentazione sono importanti perché permettono di operare sui dati in modo efficiente quando vogliamo ritrovarne uno o più, ovvero quando vogliamo eseguire delle interrogazioni o altri tipi di elaborazioni.

4. ***Modello relazionale dei dati*** – E' una rappresentazione in cui un frammento di mondo è descritto per mezzo di un insieme di strutture costituite da ***tabelle***.

5. ***Istanza*** e ***classe di istanze*** – Nella rappresentazione di un frammento del mondo reale fornita da un modello, noi abbiamo bisogno di rappresentare singoli oggetti (ad esempio singole persone che vivono a Milano), e classi di oggetti (ad esempio insiemi di persone che vivono a Milano). Nel modello relazionale, gli oggetti sono rappresentati da istanze (di oggetti), che corrispondono alle righe delle tabelle, le classi di oggetti sono rappresentate dall'insieme o classe delle righe delle tabelle. In generale, un modello di dati fornisce rappresentazioni di istanze e classi di istanze.

6. ***Tabella*** – E' la più semplice delle possibili organizzazioni di un insieme di dati. In una tabella i dati (descrittivi degli oggetti del mondo) sono rappresentati in un insieme di righe e di colonne, che formano un insieme di celle, in ognuna delle quali è rappresentato un dato. Ogni riga corrisponde a una istanza, ogni colonna corrisponde a un attributo. Ogni tabella ha un nome.

7. ***Attributo*** di tabella – L'insieme dei dati rappresentato in una colonna di una tabella. Corrisponde a una proprietà degli oggetti della tabella, come, ad esempio nel caso di cittadini residenti a Milano, il Nome, il Cognome, la Data di Nascita. Ogni attributo ha un nome e un insieme di possibili valori chiamato dominio (ad esempio per l'attributo Età, il dominio può assumere i valori da 0 a 120 anni.

8. ***Valore*** – E' il dato rappresentato in una cella di una tabella.

9. ***Base di dati nel modello relazionale*** – E' un insieme di tabelle che rappresentano un frammento di mondo. Quindi mentre il modello relazionale è l'insieme delle strutture di rappresentazione costituite dalla tabella e dagli attributi, la base di dati è l'insieme delle tabelle.

10. ***Schema nel modello relazionale*** – Dato un insieme di tabelle, lo schema ad esse associato è l'insieme dei nomi delle tabelle e per ciascuna dei nomi degli attributi. Lo schema, in altre parole non descrive istanze e valori, ma strutture di rappresentazione corrispondenti a tabelle e attributi.

11. ***Identità*** – E' la proprietà per cui un oggetto del mondo reale è rappresentato in una base di dati in modo tale che sia univocamente distinguibile da tutti gli altri oggetti rappresentati.

12. ***Chiave di una tabella*** – E' un attributo o un insieme di attributi i cui valori in una riga di una tabella individuano univocamente tutti gli altri valori nella riga (ad esempio il codice fiscale di un insieme di persone rispetta quelsta proprietà).

13. ***Interrogazione nel modello relazionale*** – E' una operazione su una o più tabelle di una base di dati, che permette di estrarre le sole righe delle diverse tabelle che rispettano determinate condizioni.

14. ***Linguaggio SQL*** – E' il linguaggio universalmente adottato nei sistemi informatici che adottano il modello relazionale per esprimere interrogazioni.

15. ***Linguaggio di interrogazione operazionale*** -– Linguaggio di interrogazione in cui le righe delle tabelle da selezionare sono individuate mediante un insieme di operazioni elementari organizzate in passi concatenati tra di loro.

16. ***Linguaggio di interrogazione dichiarativo*** – Linguaggio di interrogazione in cui le righe delle tabelle da selezionare sono individuate esprimendo semplicemente le condizioni di selezione senza entrare nel merito delle singole operazioni elementari e passi da eseguire.

17. **Decomposizione senza perdita di informazione** – E' la proprietà per cui data una tabella del modello relazionale, è possibile decomporre la tabella in un insieme di tabelle con un minor numero di attributi, e successivamente ricostruire esattamente la tabella originaria .

18. ***Modello di dati gerarchico*** – E' un modello in cui contrariamente a quanto accade ad esempio nel modello relazionale, i dati non sono rappresentati in tabelle in cui ogni cella rappresenta un valore elementare, ma, piuttosto, secondo una rappresentazione gerarchica o a livelli, in cui si può passare dal livello superiore ai livelli inferiori, ad esempio il livello superiore può rappresentare studenti, quello inferiore gli esami svolti, quello ancora inferiore i professori che hanno svolto i vari esami, ecc.

19. ***Modello Entità Relazione*** – E' un modello il cui insieme di strutture di rappresentazione è più ricco di quelle del modello relazionale, ed è costituto dalle entità, le relazioni, gli attributi di entità e relazioni, le gerarchie è-un tra due entità, le generalizzazioni tra una entità genitore e un insieme di entità figlie.

20. ***Rappresentazione grafica del modello Entità Relazione*** – Insieme di simboli grafici associati alle strutture di rappresentazione del modello Entità Relazione che permettono di rappresentare uno schema Entità Relazione così da renderlo leggibile e comprensibile in modo immediato.

21. ***Cardinalità*** – Sono proprietà delle relazioni nel modello Entità Relazione. Le relazioni nel modello collegano istanze di entità; le cardinalità, uno a uno, uno a molti, molti a molti ci dicono quante istanze di una entità sono collegate in generale a una istanza dell'altra.

22. ***Aggregazione*** – Partiamo da un esempio; una data è composta, si dice anche è una aggregazione, di un nome, di un mese e di un anno. Facendo riferimento alle classi, una classe a cui corrispondono un insieme di sottoclassi che sono parti della classe (il mese è una parte di una data), è detta classe aggregazione delle sue parti.

23. ***Generalizzazione*** – Anche qui partiamo da un esempio. Le classi degli studenti italiani e degli studenti stranieri di una università sono sottoclassi della classe degli studenti. Possiamo dire, nel modello Entità Relazione, che la entità Studente è generalizzazione delle entità Studente italiano e Studente straniero, in quanto la classe degli studenti italiani e la classe degli studenti stranieri sono entrambe sottoclassi della classe degli studenti. La classe Studente è chiamata classe genitore, le classi Studente italiano e Studente straniero sono chiamate classi figlie. Qualora la relazione di sottoclasse sia definita tra due classi, ad esempio Studente e Studente Straniero, si dice allora che tra la classe figlia e la classe genitore è definita una **gerarchia is-a**, dall'inglese ***è un*** o ***è una***.
24. ***Proprietà di ereditarietà*** – Esempio: ogni proprietà di una entità Studente è anche proprietà della entità Studente straniero ad essa collegata da una gerarchia è-un. La proprietà esprime la caratteristica intuitiva per cui se una proprietà è definita per una classe essa è anche definita per tutte le sue sottoclassi.

25. ***Modello a grafo semantico*** - E' un modello in cui al contrari di quanto avviene nel modello relazionale, i dati sono rappresentati per mezzo di nodi e archi tra nodi, le cosiddette *triple*; i nodi possono essere sia classi che istanze, e sia le classi che le istanze possono essere collegate tra loro da proprietà che in alcuni modelli a grafo semantico possono essere più espressive e ricche di proprietà di quanto accada, per le classi, nel modello Entità Relazione. I modelli a grafo semantico, rispetto al modello relazionale utilizzati nelle basi di dati, sono sempre più spesso utilizzati per una varietà di scopi diversi, alcuni dei quali sono stati esemplificati nella sezione 12. Per tale ragione, i modelli a grafo semantico, pur rappresentando classi ed istanze, non rispettano la rigida regola tipica delle basi di dati per cui prima è creato lo schema di una tabella (nome di relazione e insieme dei nomi degli attributi) e poi viene nel tempo creata e aggiornata la istanza, ma, piuttosto, possono essereci classi senza instanze e istanzxe senza classi, garantendo in questo modo maggiore flessibilità all'uso del grafo semantico.

26. ***Le cinque stelle di Tom Berners Lee*** – Sono cinque livelli successivi, proposti da Tom Berners Lee, che un insieme di dat ideve rispettare per poter essere pubblicati nel Web e collegato.

27. ***Linked open data cloud*** – E’ un vastissimo e sempre cscente negli anni insieme di dati rappresentati per mezzo di grafi semantici che utenti di tutto il mondo hanno pubblicato e continuano a pubblicare collaborativamente sul Web, rispettando nella loro pubblicazione tutte le cinque stelle di Berners Lee, in modo tale che i dati pubblicati da utenti diversi possano essere collegati tra loro

28. ***DBPedia*** – Collezione di dati costituita da dati strutturati estratti da Wikipedia e pubblicati sul Web come Linked Open Data in formato RDF, che è un particolare modello a grafo semantico.

29. ***Wikipedia*** – E’ un'enciclopedia online, libera e collaborativa. Grazie al contributo di volontari da tutto il mondo, Wikipedia è disponibile in oltre 300 lingue. Chiunque può contribuire alle voci esistenti o crearne di nuove, affrontando sia gli argomenti tipici delle enciclopedie tradizionali sia quelli presenti in almanacchi, dizionari geografici e pubblicazioni specialistiche

Approfondimenti

Se siete interessati ad approfondire la conoscenza del modello Relazionale, del modello Entità Relazione, l'SQL e i grafi semantici potete accedere alle seguenti risorse.

Sul modello relazionale e modello Entità Relazione potete

1. Scaricare dal sito http://hdl.handle.net/10281/97114 le trascrizioni del corso su Data Base
   Modeling and Design di Carlo Batini.

   Sia per il modello relazionale che per il modello Entità Relazione trovate una trattazione degli argomenti che abbiamo affrontato qui più strutturata e meno discorsiva, con definizioni, esempi ed esercizi che potete svolgere, e che sono risolti all'interno della dispensa.

2. Accedere alle lezioni video dello stesso corso sul sito
   https://open.elearning.unimib.it/course/index.php?categoryid=15

   In questo caso trovate gli stessi argomenti delle dispense erogati tramite lezioni video, quindi attraverso una comunicazione un po' più "calda".

3. Accedere alle lezioni in Power Point commentate a voce sul sito
   https://open.elearning.unimib.it/course/index.php?categoryid=15

   Queste lezioni comprendono oltre che i temi legati ai modelli dei dati anche il linguaggio di interrogazione SQL. Anche qui vengono proposti esercizi con soluzione.

Sui grafi di conoscenza potete leggere il Capitolo 7: Dati e Semantica scritto da Matteo Palmonari del libro "La scienza dei dati", che potete scaricare gratuitamente dal link

https://boa.unimib.it/handle/10281/295980

Ringraziamenti

Ringrazio la professoressa Graziella Casà e le sue studentesse e studenti che hanno collaborato con entusiasmo e dedizione a una revisione della prima versione del testo, fornendomi indicazioni importanti per comprendere quali fossero le parti meno chiare.

Ringrazio

Francesco Angelone
Laura Batini
Gabriele Facciolongo
Claudio Salone
Gaetano Santucci
Teresa Serafini
Fabio Stella

per le loro osservazioni su una prima versione del testo.