Basi di dati

Capitolo 9 del testo

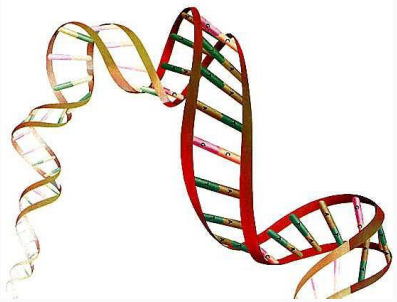
Alberto Policriti

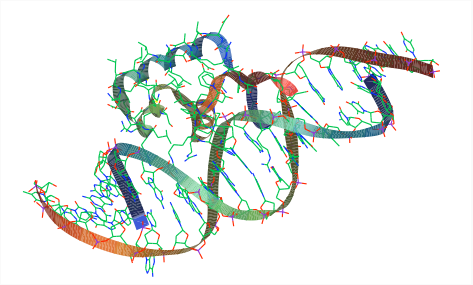
IGA
ISTITUTO DI GENOMICA APPLICATA

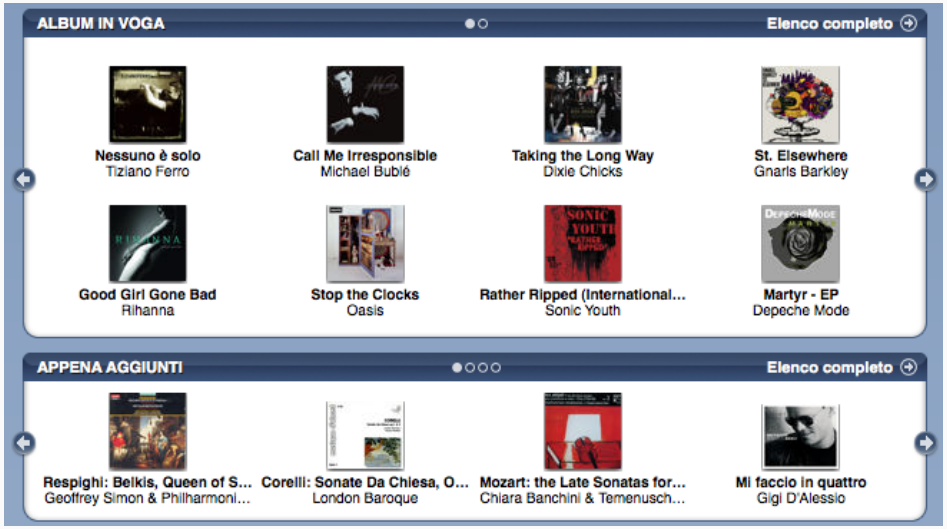
UNIVERSITAS STUDIORUM UTINENSIS

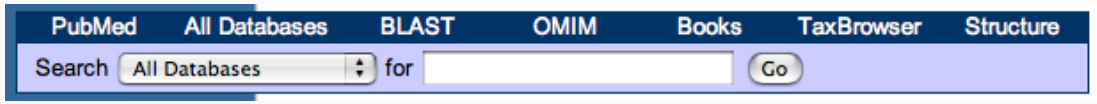
5 Dicembre, 2019

Informazioni









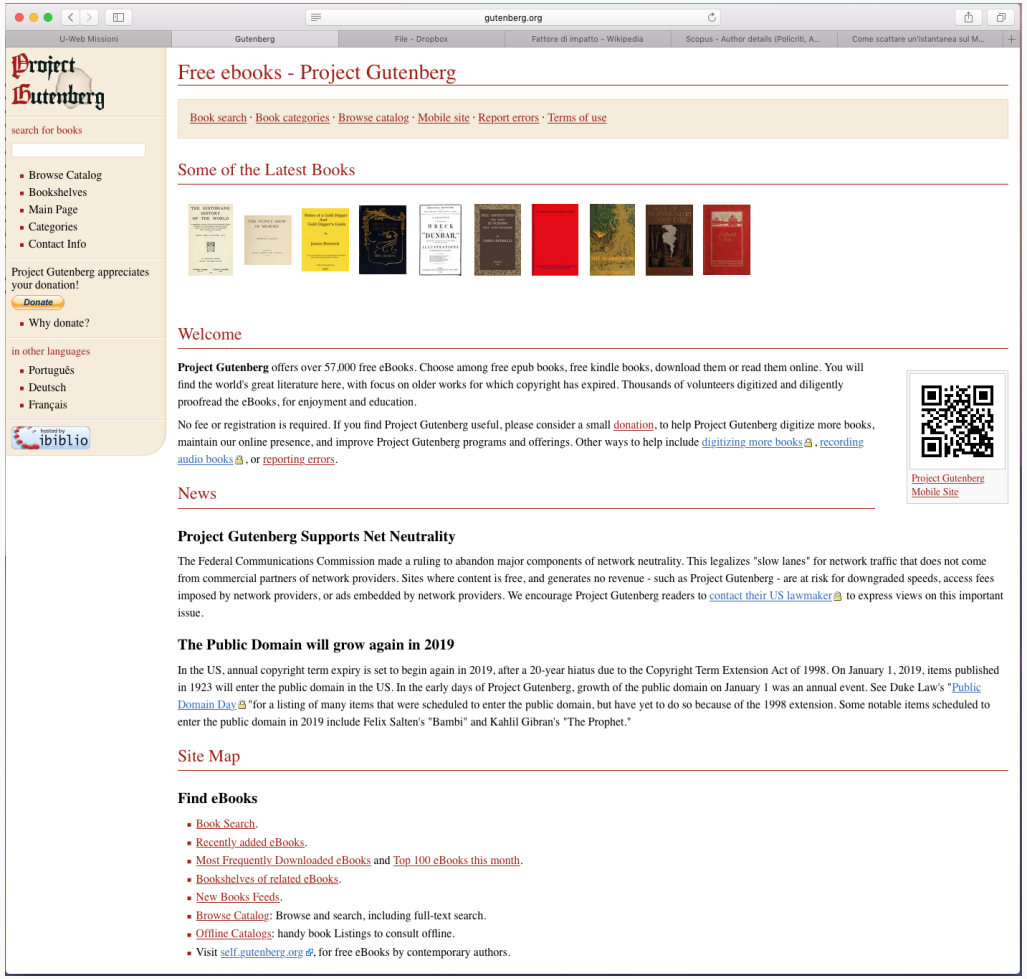
Attività di un'organizzazione

- Raccolta delle informazioni
- Rappresentazione delle informazioni (dati)
- Organizzazione dei dati
- Conservazione dei dati
- Reperimento/interrogazione dei dati

. . . per il perseguimento degli scopi dell'organizzazione

Sistema informativo

Sistema informativo vs sistema informatico



http://www.gutenberg.org/

Base di dati

Collezione strutturata di dati...

- di dimensioni arbitrarie
- persistente
- condivisa

DBMS

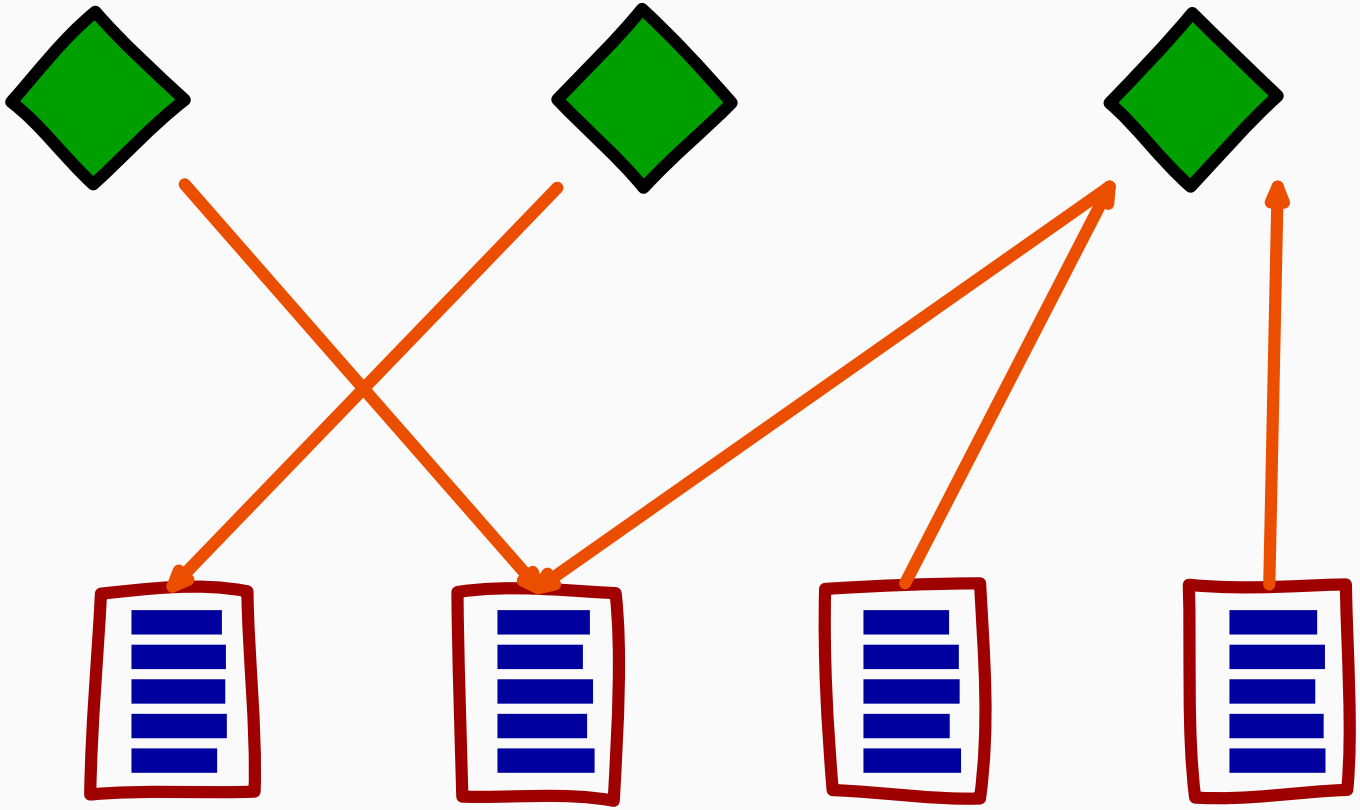
DataBase Management System (Sistema di gestione di basi di dati)

- strumento software (collezione di programmi)
- per la creazione e manipolazione di basi di dati
- di qualunque dimensione e per qualunque scopo

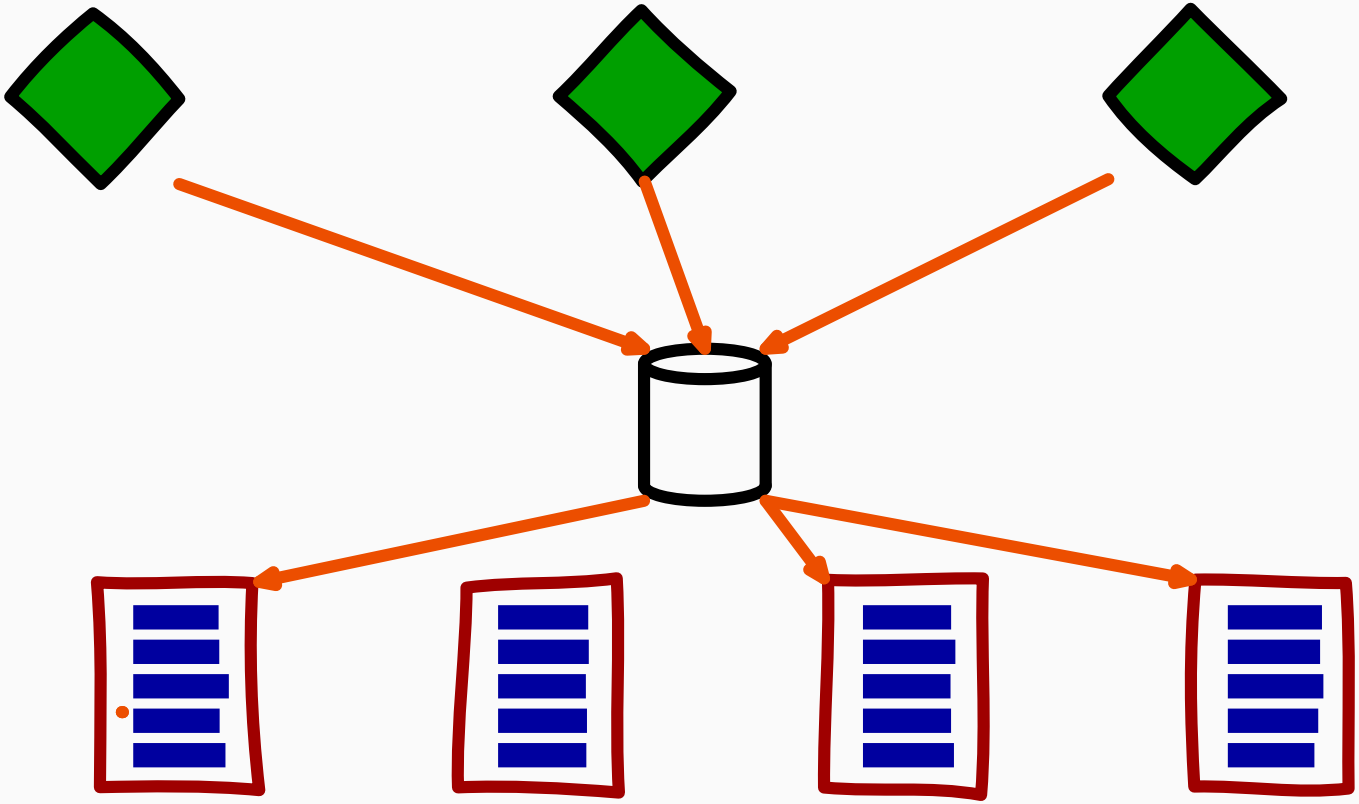
Esempi:

- Oracle
- PostgreSQL
- MySQL
- ...

File + programmi



Tecnologia delle BdD



Modello dei dati

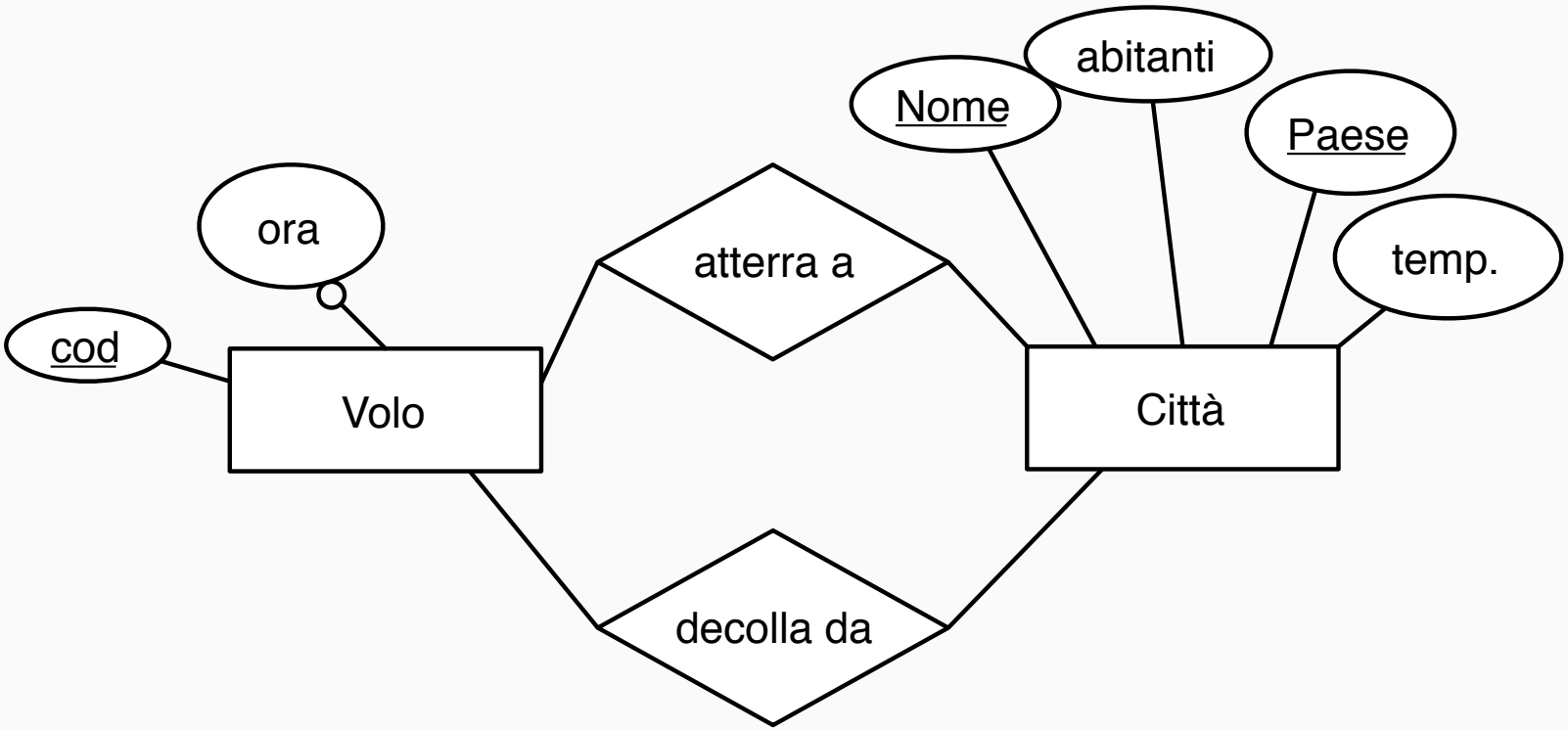
Collezione di concetti e regole per la descrizione dei dati, delle relazioni tra i dati e dei vincoli di consistenza sui dati.

- Proprietà statiche:
  - oggetti di informazione, entità
  - proprietà degli oggetti (attributi)
  - relazioni tra gli oggetti
  - vincoli su oggetti e relazioni
- Proprietà dinamiche:
  - operazioni su oggetti e relazioni
  - relazioni tra operazioni (transazioni)
  - vincoli sull'evoluzione degli oggetti e delle relazioni

Modelli dei dati

- Modelli dei dati primitivi (ma vedi i ,“flat file”. . .)
  - basati sulla gestione diretta di file
  - Operazioni sui dati $\equiv$ operazioni sui file
- Modelli dei dati classici
  - Fondamento dei DBMS attuali
  - Obiettivo: indipendenza dei dati
  - Modello relazionale
- Modelli dei dati semantici
  - Strumenti per la progettazione
  - Modello entità-relazione (ER)

Modello entità-relazione



Per una completa descrizione (con esempi) si veda:

https://users.dimi.uniud.it/~massimo.franceschet/teatro-sql/index.html

Il modello relazionale

- Storia:
  - Proposto da E. Codd nel 1970
  - Acquista popolarità negli anni Ottanta
  - Attualmente, il modello più diffuso
- Caratteristiche:
  - Semplicità: l'utente percepisce la base di dati come un insieme di tabelle
  - Le relazioni tra le tabelle sono implicitamente rappresentate dai valori
  - Linguaggi di manipolazione dichiarativi
    - specificano che risultato ottenere, non le modalità per ottenere il risultato)
  - Prospettiva algebrica: insieme di strutture dati e operatori
  - Prospettiva logica: SQL

Schema di relazione

$$R(A_1, \ldots, A_n)$$

- Nome di relazione: R
- Attributi: $A_1, \ldots, A_n$
- A ciascun A_i è associato un dominio.

Lo schema di una base di dati è un insieme di schemi di relazione

Esempio

- Genoma(organismo, dimensione)
- Sequenza(accession, seq_grezza, specie)
- Blast(database, data, risultato, seq_id)

Rappresentazione tabulare:

Genoma	
organismo	dimensione

Sequenza		
accession	seq_grezza	specie

Blast			
db	data	risultato	seq_id

Istanza di base di dati

Genoma

organismo	dimensione
Homo Sapiens	3000
Arabidopsis Thaliana	100

Sequenza

accession	seq_grezza	specie
1234	ATGCT...	Homo Sapiens
567	GTCCGT...	Arabidopsis
890	TGGGGA...	Homo Sapiens

Blast

db	data	risultato	seq_id

Relazioni e tabelle

Concetto relazionale:

- relazione
- attributo
- grado
- tupla
- cardinalità

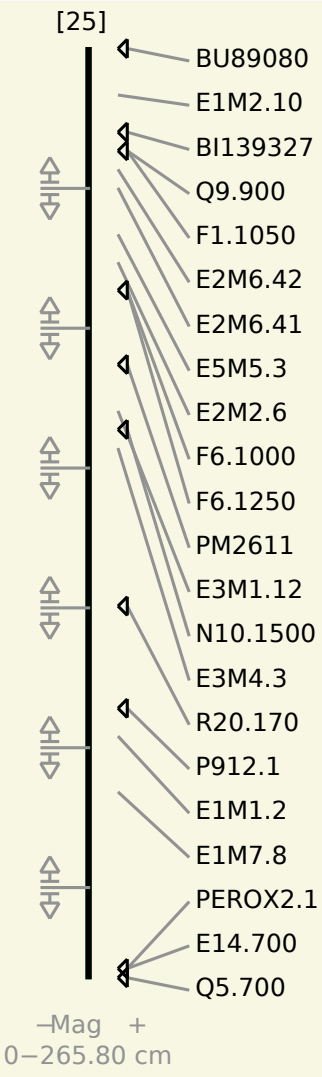
Equivalente informale:

- tabella
- colonna
- numero di colonne
- riga
- numero di righe

Differenze tra tabelle e relazioni:

- in una tabella le colonne e le righe sono ordinate
  - nelle relazioni non sussistono tali ordinamenti
- in una tabella possono esistere righe ripetute
  - una relazione è un insieme, pertanto non ci possono essere tuple ripetute

Mappe biologiche: schema



- Mappa(id, specie, nome, tipo, unità, start, stop)
- Marker(id, mappa, tipo, nome, start, stop)

Mappe: un'istanza

Mappa

id	specie	nome	tipo	unità	start	stop
1	F1	LG I	gen	cM	0	70.6
2	F1	LG II	gen	cM	0	122.4
3	D2	Chr IV	seq	bp	0	1437893

Marker

id	mappa	tipo	nome	start	stop
0	1	SSR	PM1234	36.8	NULL
1	1	SSR	PM3245	18.4	NULL
2	3	SSR	PM334	748876	748967
3	1	AFLP	E5M1.3	8.3	NULL

Vincoli d'integrità

Proprietà, associate a uno schema di base di dati, che devono essere soddisfatte da tutte le istanze che rappresentano informazioni corrette della base di dati.

- Si specificano insieme allo schema della base di dati
- il DBMS verifica la consistenza dei dati rispetto ai vincoli
- Vincoli intrarelazionali
  - coinvolgono una o più tuple di una stessa relazione
- Vincoli interrelazionali
  - coinvolgono tuple di relazioni diverse

Esempio

Mappa(id, specie, nome, tipo, u, start, stop)
Marker(id, mappa, tipo, nome, start, stop)

- I valori di start e stop devono essere non negativi
- Non ci possono essere id ripetuti
- Il valore di mappa in Marker deve corrispondere a qualche mappa
- I nomi dei marcatori devono essere sempre noti
- ...

Chiavi

- Tupla $\equiv$ entità del mondo reale
- Identificabile mediante un sottoinsieme di valori
  - univoci
  - sempre noti
- Chiavi: sottoinsiemi minimali con tali caratteristiche
- Chiave primaria: scelta dal progettista tra le chiavi candidate

Esempio

Esame(matricola, corso, data, voto, lode)

Quali sono le chiavi? L'unica è $\{\text{matricola}, \text{corso}\}$

Visita(data, ora, stanza, medico, paziente, referto)

Quali sono le chiavi?

1. $\{\text{data}, \text{ora}, \text{stanza}\}$
   - A una data ora di un certo giorno ci può essere solo una visita per stanza
2. $\{\text{data}, \text{ora}, \text{medico}\}$
   - A una data ora di un certo giorno un medico non può fare più di una visita
3. $\{\text{data}, \text{ora}, \text{paziente}\}$
   - A una data ora di un certo giorno un paziente non può essere sottoposto a più visite

Integrità referenziale

I valori di uno o più attributi di una relazione devono occorrere in attributi corrispondenti di un'altra relazione (e.g., specie $\rightarrow$ organismo)

Chiave esterna

Genoma

organismo	dimensione
Homo Sapiens	3000
Arabidopsis T.	100

Sequenza

accession	seq_grezza	specie
1234	ATGCT...	Homo Sapiens
1235	GTCCGT...	Arabidopsis T.

Esercizi

1. Definire uno schema relazionale completo per rappresentare l'informazione relativa ai libri letti dai soci di una biblioteca.
2. Definire uno schema relazionale completo per rappresentare l'informazione relativa ai voli internazionali di una compagnia aerea. Ciascun volo è identificato da un codice e da una data, e decolla e atterra in determinati aeroporti (che devono essere sempre noti), che si trovano in determinate città, di cui interessa sapere il numero di abitanti. Si tenga conto che:
   - una città può avere diversi aeroporti;
   - città di stati diversi possono avere lo stesso nome.

Beh, perché non usare un foglio elettronico?

- File + programmi?!
- Assenza di linguaggi d'interrogazione $\equiv$ limitate possibilità di estrarre le informazioni desiderate
- Inconsistenze, assenza di condivisione e concorrenza, etc. . .

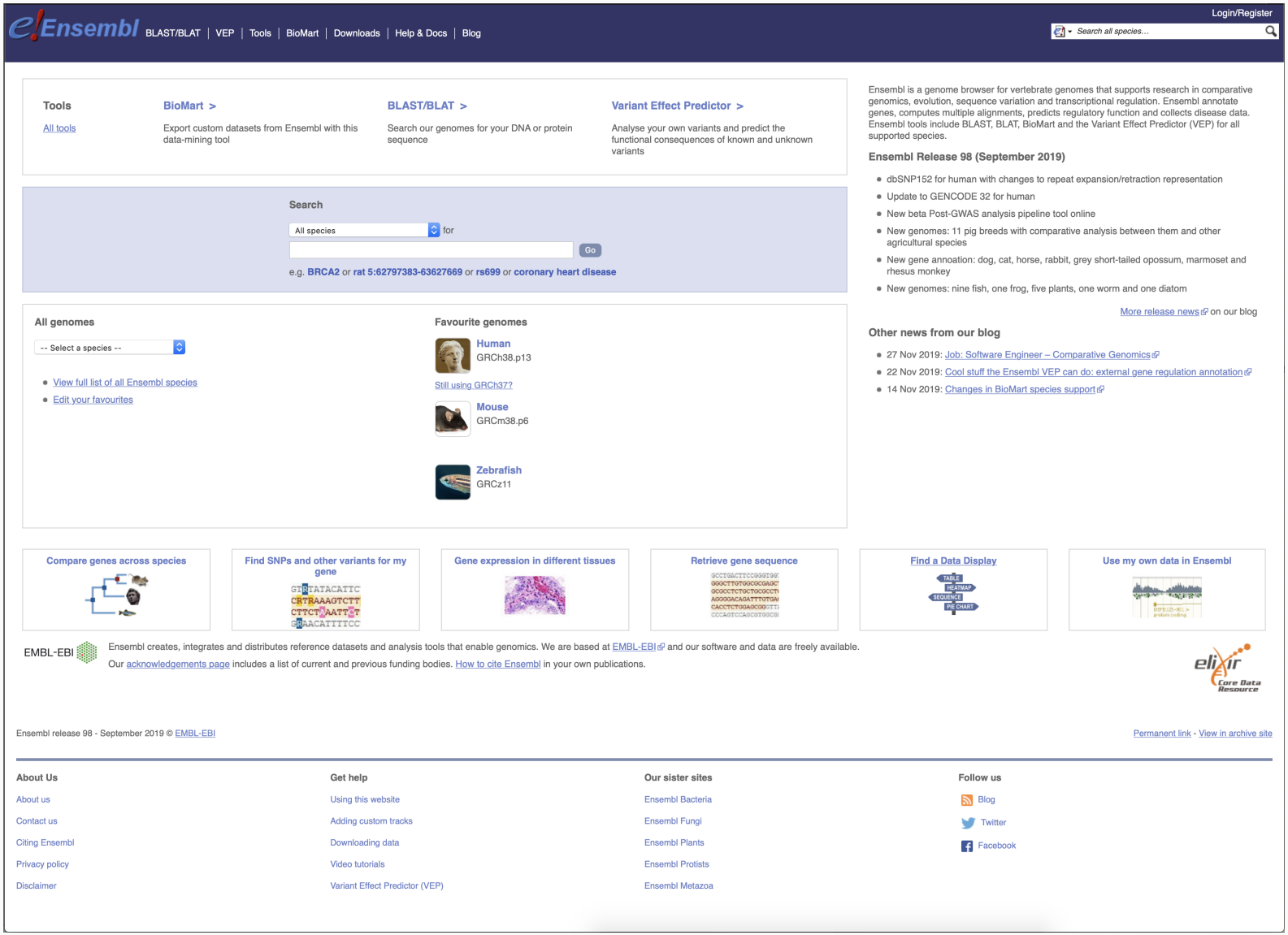
Ensembl

Joint project to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

- Many databases, even for the same organism
- All data sets in the Ensembl system are stored in relational databases (MySQL)
- Data can be obtained by FTP (`ftp://ftp.ensembl.org/pub/`)
- Data can be queried directly (the database schema is public)

  ```
  mysql -u anonymous -h ensembldb.ensembl.org
  ```

Ensembl Home Page



Protein Data Bank

PDB (`http://www.rcsb.org/pdb/home/home.do`) provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

- Single database (MySQL): 461 tables
- The database schema is public, but the db can be queried only through the web interface
- Complex queries through ,“Advanced search”
- Experimental and theoretical models are kept distinct
- Integrates also data from Swiss-Prot, Genbank, PubMed.

Protein Data Bank Home Page

PDB Database: Reference

Nucleic Acids Res. 2005 January 1; 33(Database Issue): D233–D237.
Published online 2004 December 17. doi: 10.1093/nar/gki057.

The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema

Nita Deshpande,[1] Kenneth J. Addess,[1] Wolfgang F. Bluhm,[1] Jeffrey C. Merino-Ott,[1] Wayne Townsend-Merino,[1] Qing Zhang,[1] Charlie Knezevich,[1] Lie Xie,[1] Li Chen,[3] Zukang Feng,[3] Rachel Kramer Green,[3] Judith L. Flippen-Anderson,[3] John Westbrook,[3] Helen M. Berman,[3] and Philip E. Bourne[1,2*]

Nucleic Acids Res

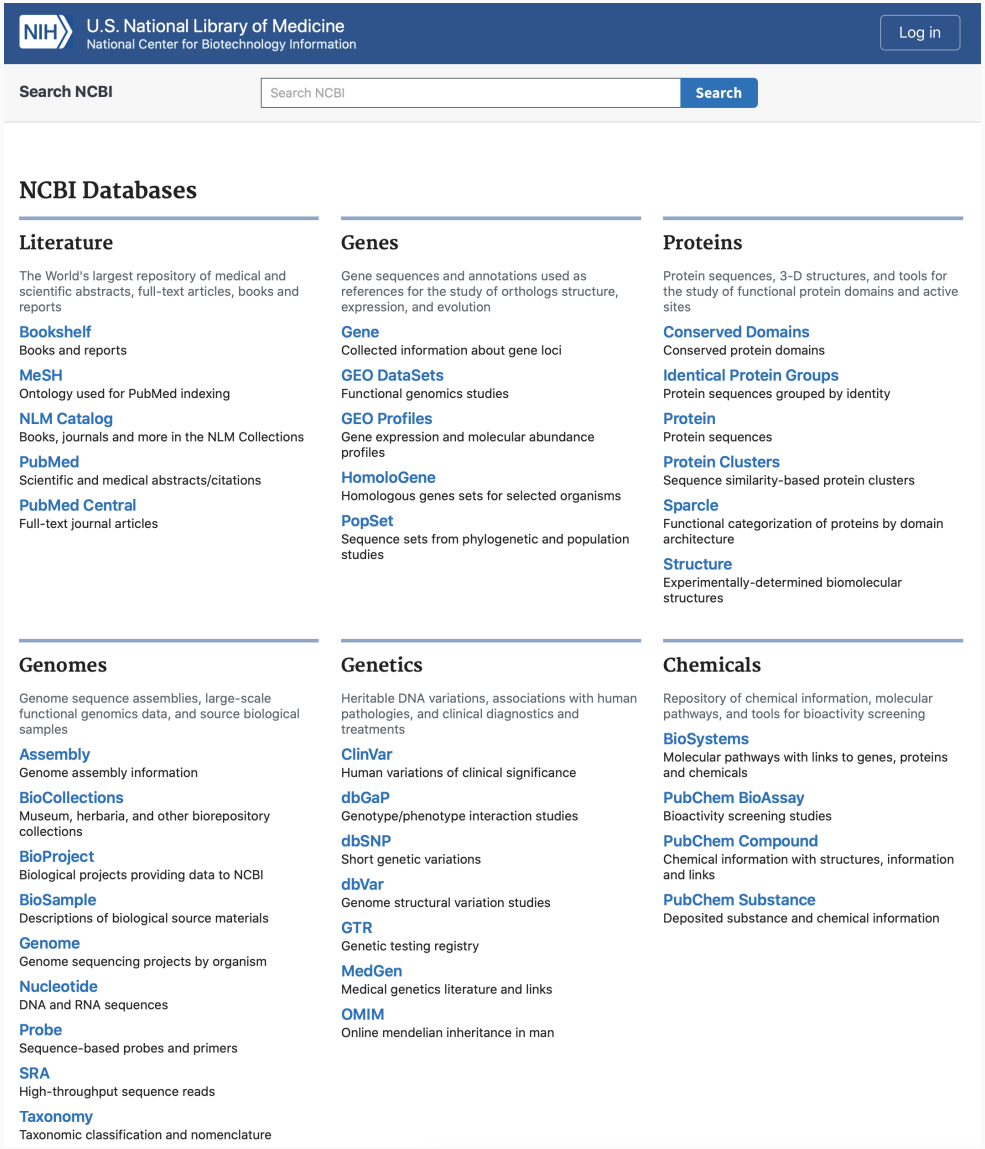
Entrez

Entrez (`http://www.ncbi.nlm.nih.gov/Entrez/`) integrates

- the scientific literature
- DNA and protein sequence databases
- 3D protein structure and protein domain data
- population study datasets
- expression data
- assemblies of complete genomes
- taxonomic information

It is a retrieval system designed for searching its linked databases.

Entrez Global Query



PubMed

PubMed is a service of the U.S. National Library of Medicine that includes over 17 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s.

PubMed Central is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature.

- One of the literature databases of Entrez
- Number of records in PubMed: 17.505.726 (8/11/07, 12:40)
- Search can be done by author, title, journal
- Makes extensive use of XML technology

A PubMed record

Functional Inactivation of EBV-Specific T-Lymphocytes in Nasopharyngeal Carcinoma: Implications for Tumor Immunotherapy.

Li J, Zeng XH, Mo HY, Rolén U, Gao YF, Zhang XS, Chen QY, Zhang L, Zeng MS, Li MZ, Huang WL, Wang XN, Zeng YX, Masucci MG.

State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou, China.

Nasopharyngeal carcinoma (NPC) is an Epstein-Barr virus (EBV) associated malignancy with high prevalence in Southern Chinese. In order to assess whether defects of EBV-specific immunity may contribute to the tumor, the phenotype and function of circulating T-cells and tumor infiltrating lymphocytes (TILs) were investigated in untreated NPC patients. Circulating naïve CD3(+)CD45RA(+) and CD4(+)CD25(-) cells were decreased, while activated CD4(+)CD25(+) T-cells and CD3(-)CD16(+) NK-cells were increased in patients compared to healthy donors. The frequency of T-cells recognizing seven HLA-A2 restricted epitopes in LMP1 and LMP2 was lower in the patients and remained low after stimulation with autologous EBV-carrying cells. TILs expanded in low doses of IL-2 exhibited an increase of CD3(+)CD4(+), CD3(+)CD45RO(+) and CD4(+)CD25(+) cells and 2 to 5 fold higher frequency of LMP1 and LMP2 tetramer positive cells compared to peripheral blood. EBV-specific cytotoxicity could be reactivated from the blood of most patients, whereas the TILs lacked cytotoxic activity and failed to produce IFNgamma upon specific stimulation. Thus, EBV-specific rejection responses appear to be functionally inactivated at the tumor site in NPC.

A PubMed XML record

```
<PubmedArticle>
   <MedlineCitation Owner="NLM" Status="In-Data-Review">
      <PMID>17987110</PMID>
      <DateCreated>
         <Year>2007</Year>
         <Month>11</Month>
         <Day>07</Day>
      </DateCreated>
      <Article PubModel="Electronic">
         <Journal>
            <ISSN IssnType="Electronic">1932-6203</ISSN>
            <JournalIssue CitedMedium="Internet">
               <Volume>2</Volume>
               <Issue>11</Issue>
               <PubDate>
                  <Year>2007</Year>
               </PubDate>
            </JournalIssue>
            <Title>PLoS ONE</Title>
            <ISOAbbreviation>PLoS ONE</ISOAbbreviation>
         </Journal>
         <ArticleTitle>Functional Inactivation of EBV-Specific T-Lymphocytes in Nasopharyngeal Carcinoma: Implications for
         Tumor Immunotherapy.</ArticleTitle>
         <Pagination>
            <MedlinePgn>e1122</MedlinePgn>
         </Pagination>
         <Abstract>
            <AbstractText>Nasopharyngeal carcinoma (NPC) is an Epstein-Barr virus (EBV) associated malignancy with high
            prevalence in Southern Chinese. In order to assess whether defects of EBV-specific immunity may contribute to the
            tumor, the phenotype and function of circulating T-cells and tumor infiltrating lymphocytes (TILs) were
            investigated in untreated NPC patients. Circulating naïve CD3(+)CD45RA(+) and CD4(+)CD25(-) cells were decreased,
            while activated CD4(+)CD25(+) T-cells and CD3(-)CD16(+) NK-cells were increased in patients compared to healthy
            donors. The frequency of T-cells recognizing seven HLA-A2 restricted epitopes in LMP1 and LMP2 was lower in the
            patients and remained low after stimulation with autologous EBV-carrying cells. TILs expanded in low doses of IL-2
            exhibited an increase of CD3(+)CD4(+), CD3(+)CD45RO(+) and CD4(+)CD25(+) cells and 2 to 5 fold higher frequency of
            LMP1 and LMP2 tetramer positive cells compared to peripheral blood. EBV-specific cytotoxicity could be reactivated
            from the blood of most patients, whereas the TILs lacked cytotoxic activity and failed to produce IFNgamma upon
            specific stimulation. Thus, EBV-specific rejection responses appear to be functionally inactivated at the tumor
            site in NPC.</AbstractText>
         </Abstract>
         <Affiliation>State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou,
         China.</Affiliation>
         <AuthorList CompleteYN="Y">
            <Author ValidYN="Y">
               <LastName>Li</LastName>
               <ForeName>Jiang</ForeName>
               <Initials>J</Initials>
```