Biomedical Databases 2018-19

- **Slides available at**
  https://iol.unibo.it/course/view.php?id=22990

- **Textbook: "Introduction to Bioinformatics" by Arthur M. Lesk**

- **Final assignment will consist in problems similar to those assigned during class hours.**

Biomedical Databases 2018-19 – main goals

- Knowing the basic infrastructure of bioinformatics, i.e. the sites of the major publicy available archival projects and their content.
- Becoming skilled at information retrieval.
- Knowing how to analyse and interpret the retrieved data.

Biomedical Databases 2018-19

- **Questionnaire on previous experience**.

Nucleic Acids Research

The 2015 *Nucleic Acids Research* Database Issue and Molecular Biology Database Collection

Michael Y. Galperin[1,*], Daniel J. Rigden[2] and Xosé M. Fernández-Suárez[3]

+ Author Affiliations

*To whom correspondence should be addressed. Tel: +1 301 435 5910; Fax: +1 301 435 7793; Email: nardatabase@gmail.com

Received November 10, 2014.
Accepted November 11, 2014.

The NAR online Molecular Biology Database Collection now includes > 1790 databases sorted into 15 categories.

The first NAR Data base issue, published on July 1, 1993, consisted of 24 articles (24 databases).

Nucleic Acids Research

2015 NAR Database Summary Paper Category List

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes

List of biological databases

From Wikipedia, the free encyclopedia

Biological databases are stores of biological information.[1] The journal *Nucleic Acids Research* regularly publishes special issues on biological databases and has a list of such databases. The 2018 issue has a list of about 180 such databases and updates to previously described databases.[2]

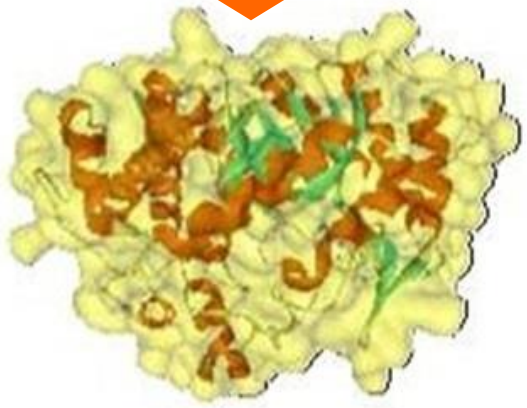
Contents [hide]

https://en.wikipedia.org/wiki/List_of_biological_databases#Meta_databases

The basic Information Flow

GenBank: 932‘095‘164 sequences

>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MTSFPNSFLFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPTPQNFDE
SKQDVTEVEINKNELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRSGMMNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEHADDGWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRIVVKRTEKGTVSLGGYGHGCEKNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYELYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLIKVDYNT
KRLYWRPSALVYKEIATNGAITDEIEELNSVPVVKPLRR



UniProt/Tremble: 133‘507‘323 sequences

UniProt/SwissProt: 558‘590 sequences

PDB: 146‘093 structures

Update: October 2018

Biomedical Databases - Overview

- **Databases at NCBI (Bethesda, Maryland)**

- GenBank	Nucleotide sequences
- PubMed	Literature

- **Databases at EMBL-EBI (Hinxton, UK)**

- UniProt	Protein sequences
- Ensembl	Genomes

- **Database at Rutgers (Univ. New Jersey)**

- PDB	Protein structures

Biomedical Databases – Classifications

- **Content** (DNA, protein, etc)
- **Type of data** (primary // secondary)
- **Annotation** (absent // present, manual // automatic)

Biomedical Databases – Classifications

Type of data

Primary data (experimental results)

- Genomes
- Protein Sequences
- Protein Structures Interactions
- Expression
- ...

Care about the experimental methods

Secondary data (derived information)

- Protein folds
- Protein families
- Genome comparisons
- ...

Care about the primary source and derivation

Biomedical Databases – Classifications

Annotation / Curation

Raw data

e.g. Protein Sequence only

Annotated data

e.g. Protein Sequence ***plus***
function,
localization,
expression info,
links,
…

Manual

e.g. Swiss Prot

Automatic

e.g. TrEMBL

Some general problems in biological databases

- **Quality check**

  - If errors do enter databases (either in data or annotations) they tend to propagate into other databases, local or secondary, and are difficult to estirpate.

- **Lack of standardization**

  - lack of standard schema (every database has its own)
  - lack of standard nomenclature (every database has its own names/accessions)

Database interoperability / interconnection

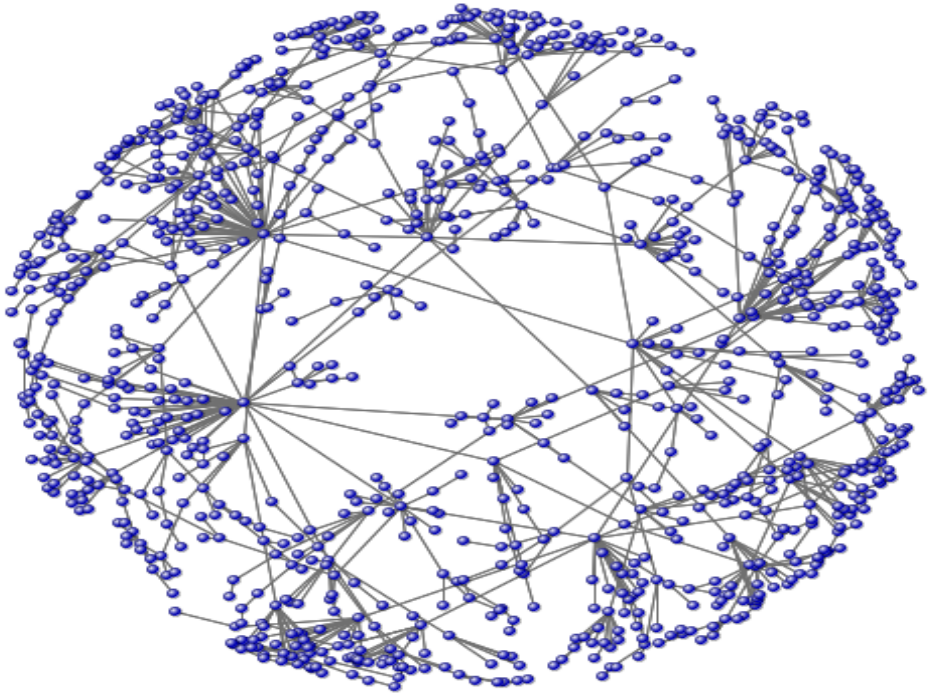
The quest for integration among different databases continue to be an area of active research

- **Fusion** of several databases into one
  (*e.g.* UniProt)
- **Links** making connection to other databases
  ("external links" or "cross-references" or "cross-linking")

Links are crucial for a database

Internal links allow navigation within the database itself.

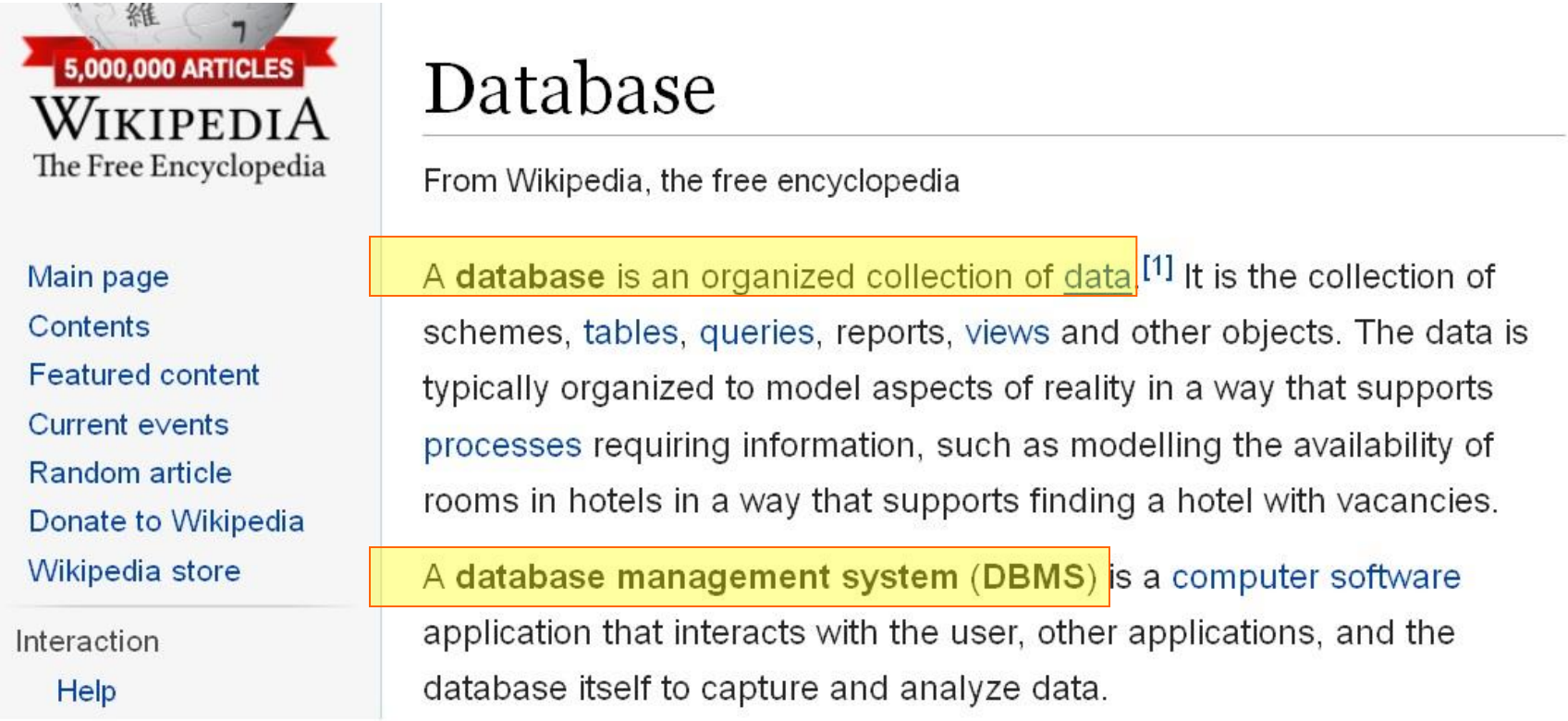
External links make connections to other databases.



Database network

***E.g.*, in Uniprot digit "crambin" and look at the external links; reach the related Pfam entry. Do the same in PDB and in NCBI Protein.**

Definition of “database”



5,000,000 ARTICLES
WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help

Database

From Wikipedia, the free encyclopedia

A **database** is an organized collection of data.[1] It is the collection of schemes, tables, queries, reports, views and other objects. The data is typically organized to model aspects of reality in a way that supports processes requiring information, such as modelling the availability of rooms in hotels in a way that supports finding a hotel with vacancies.

A **database management system** (**DBMS**) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data.

A database is a large structured set of persistent data, usually in computer-readable form.

“Organization” and “structure” allow to eventually extract meaning from the stored data.

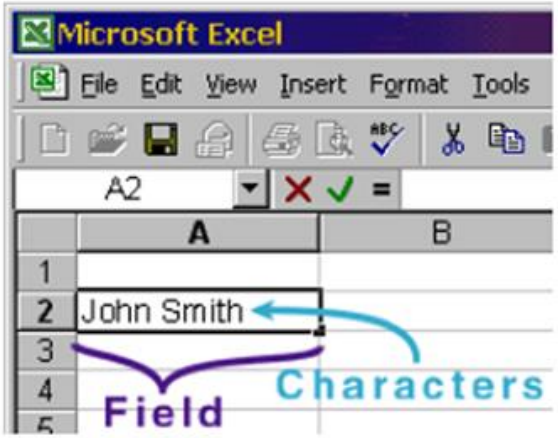
CHAOS

ORDER

Order and Chaos, M.C. Escher, 1950

Some database terms

character and field

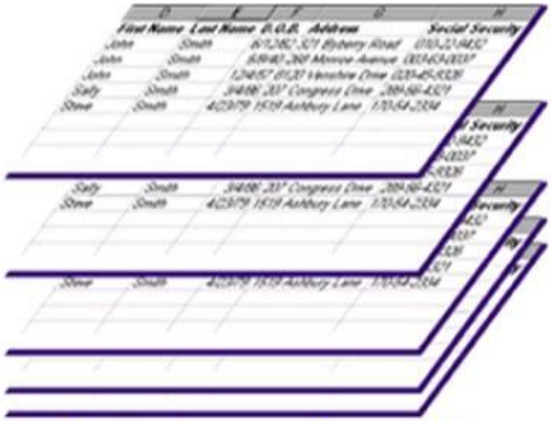


record

D	E	F	G	H
First Name	Last Name	D.O.B.	Address	Social Security
John	Smith	6/12/82	321 Byberry Road	010-22-9432
John	Smith	5/9/40	268 Monroe Avenue	003-63-0037
John	Smith	12/4/57	8120 Venshire Drive	020-45-9326
Sally	Smith	3/4/86	207 Congress Drive	289-56-4321
Steve	Smith	4/23/79	1519 Ashbury Lane	170-54-2334

This record is made up of multiple fields, First Name, Last Name, Date of Birth, Address and Social Security Number.

database file (or "table")



Related records combine to create a database file

Key (unique identifier)

D	E	F	G	H
First Name	Last Name	D.O.B.	Address	Social Security
John	Smith	6/12/82	321 Byberry Road	010-22-9432
John	Smith	5/9/40	268 Monroe Avenue	003-63-0037
John	Smith	12/4/57	8120 Venshire Drive	020-45-9326
Sally	Smith	3/4/86	207 Congress Drive	289-56-4321
Steve	Smith	4/23/79	1519 Ashbury Lane	170-54-2334

Social Security Number is the key to this spreadsheet

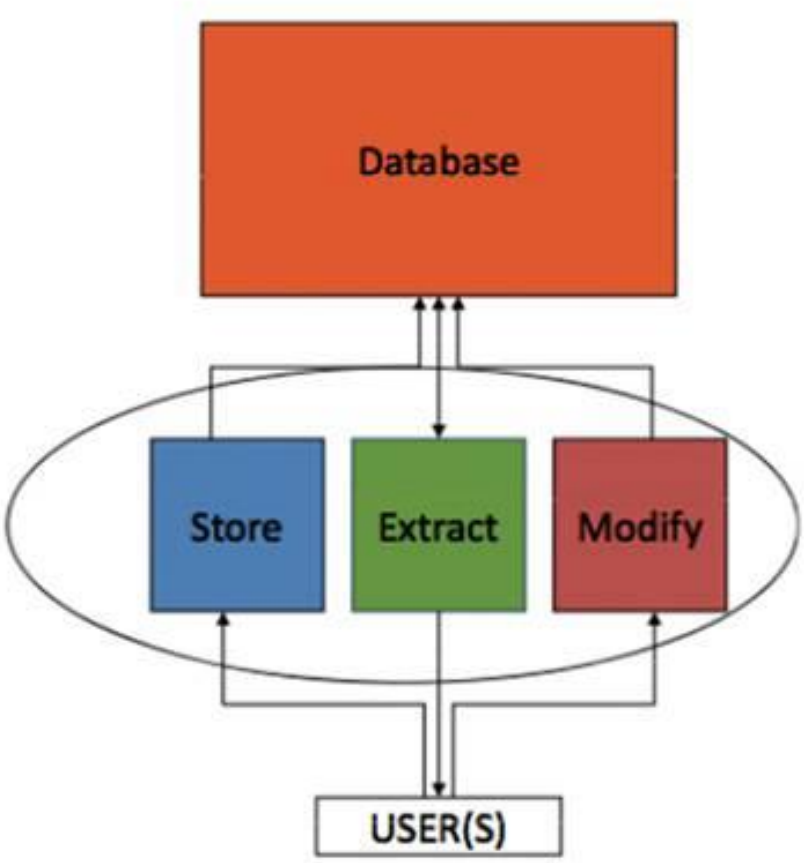
query

a request for data from a database

DBMS (Database Management System)

A DBMS is a software package that enables users:

- to access the data;
- to manipulate (create, edit, link, update) files as needed;
- to preserve the integrity of the data;
- to deal with security issues (who should have access).



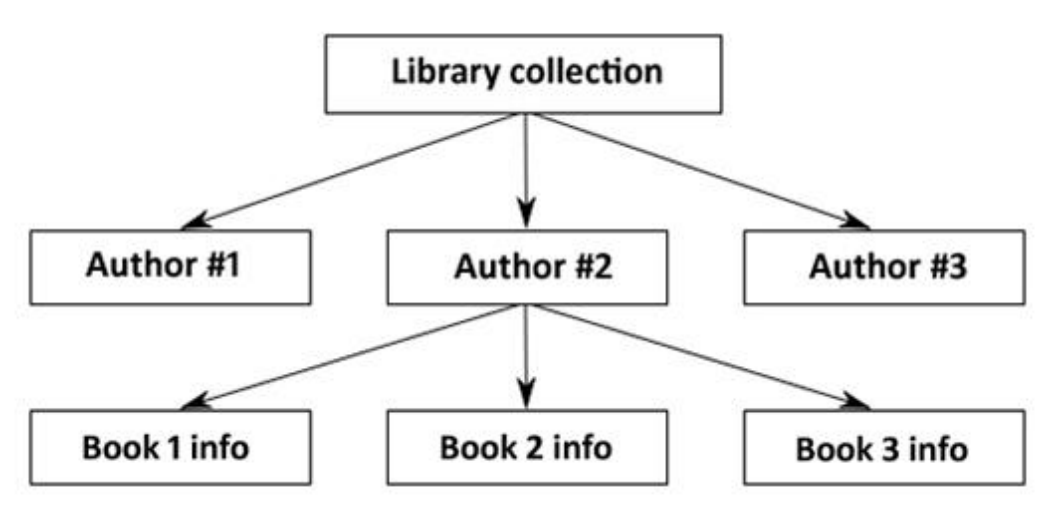
database + DBMS $\Rightarrow$ database system

DBMS models

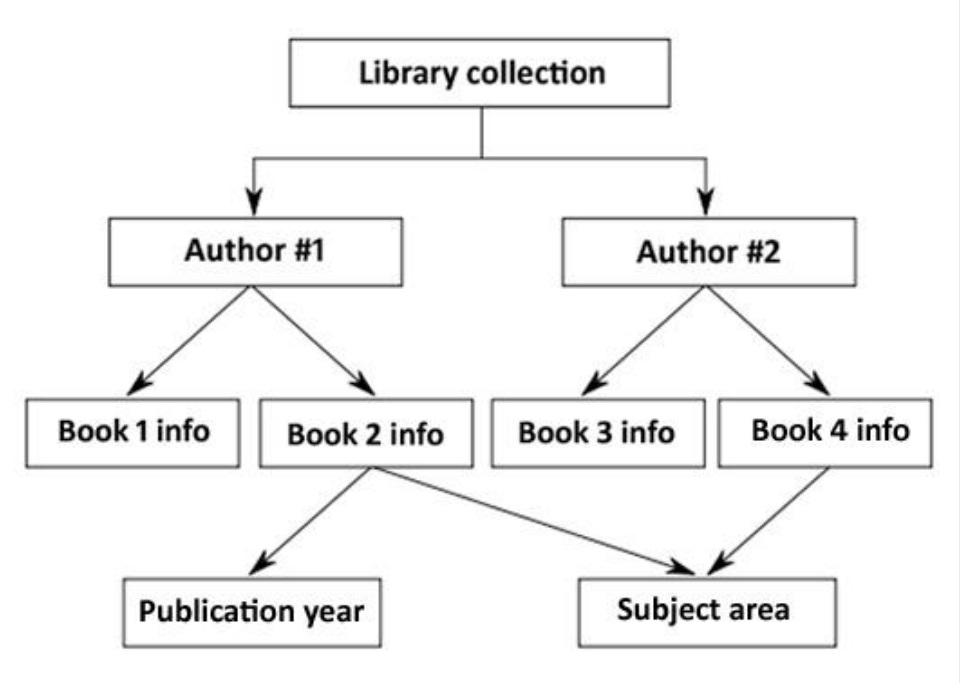
A DBMS model (database model) describes the logic by which a database is structured and organized.

- Hierarchical model
- Network model
- Relational model
- Object-oriented model

Well-known DBMSs include e.g. MySQL, PostgreSQL, SQLite, Microsoft SQL Server, Oracle (Object-Relational), IBM DB2 (Object-Relational).

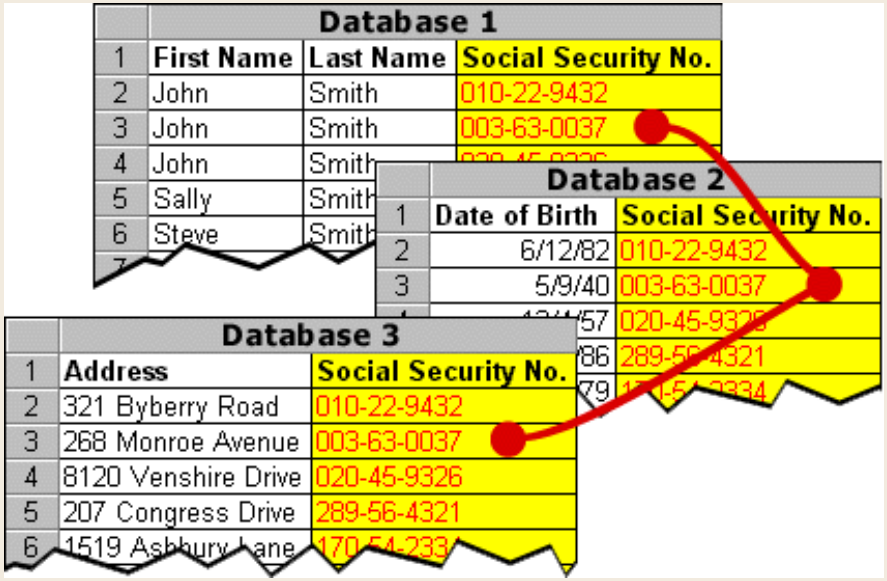


Hierarchical model



Network model

Relational model



Database 1

	First Name	Last Name	Social Security No.
2	John	Smith	010-22-9432
3	John	Smith	003-63-0037
4	John	Smith	
5	Sally	Smith	
6	Steve	Smith	

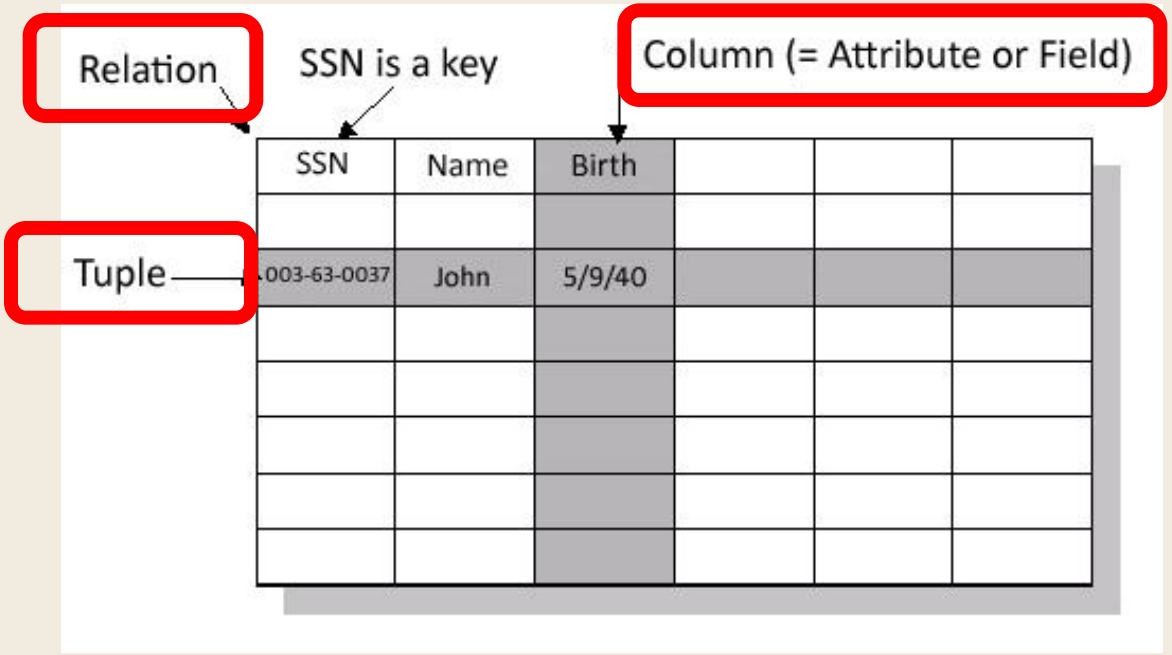
Database 2

	Date of Birth	Social Security No.
2	6/12/82	010-22-9432
3	5/9/40	003-63-0037

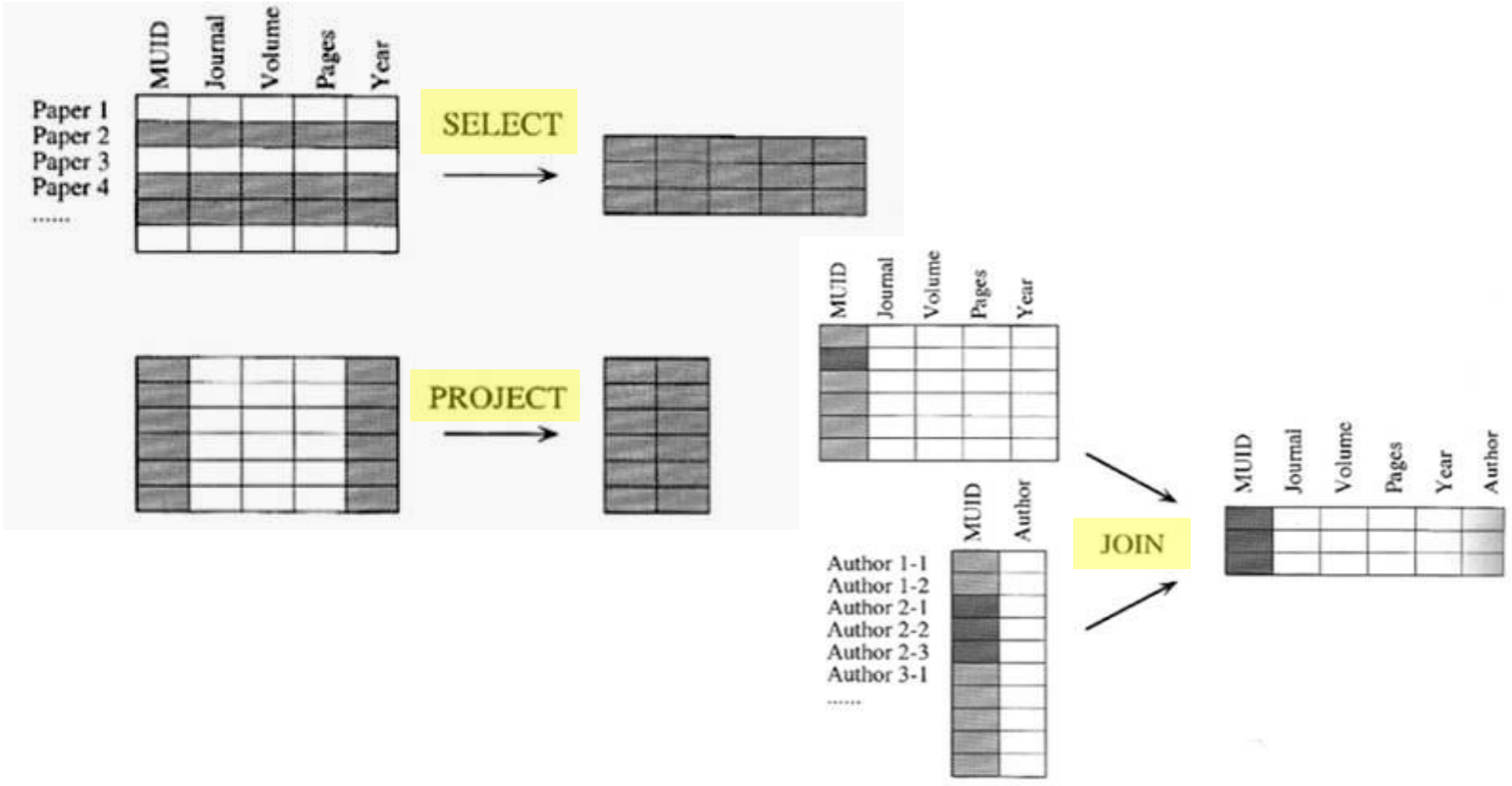
Database 3

	Address	Social Security No.
2	321 Byberry Road	010-22-9432
3	268 Monroe Avenue	003-63-0037
4	8120 Venshire Drive	020-45-9326
5	207 Congress Drive	289-56-4321
6	1519 Ashbury Lane	

Each table has a key field that uniquely identifies each row and can be used to connect one table of data to another.



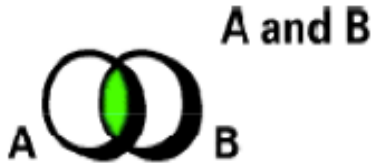
Relational operations



A few rules allow to extract information from the database according to the content of a set of attributes. The user can then choose to view the content of many additional attributes.

Boolean Operators

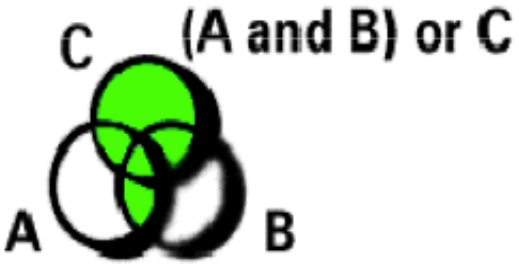
- AND (&)



- OR (|)

- NOT(!)

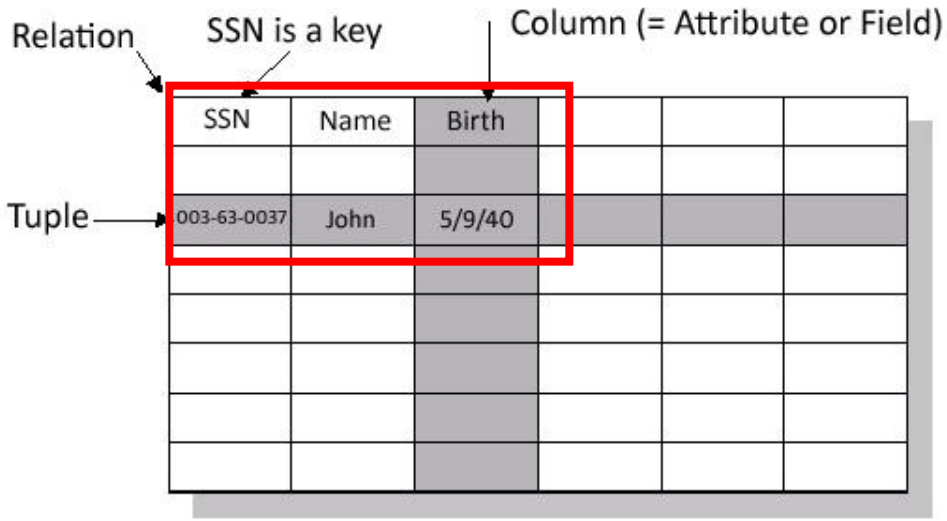
Complex queries



Schema and instance

In a database, the schema describes the logical structure of the data (the "intensional" part of the database).

The instance is the set of the actual data (the "extensional" part of the database).



The schema allows the interpretation of the data of the instance.

Basic biological entry schema

- ID/accession, name and description
- Positional Features
  - Subfeatures
- Annotations (any information not related to a portion of the sequence)
  - Keywords
  - Ontology
    - Function
    - Localization
    - ...
  - Comments (use carefully)
  - ...
- Database cross-references (DBXrefs)
- References
- Sequence (not mandatory)

A schema can be written in different file formats

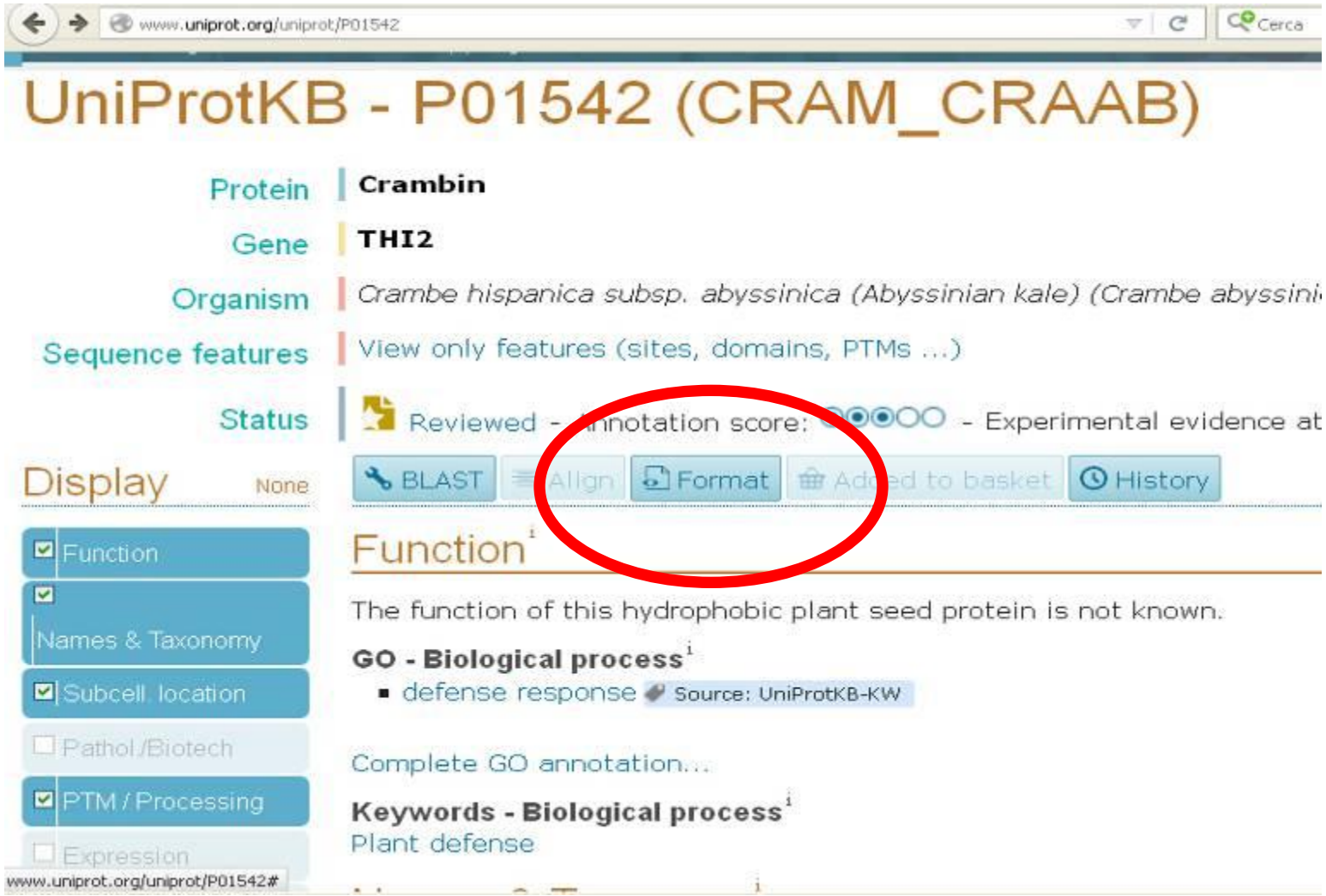
- **Flat file** – simple and available to all.

- **XML** – eXtensible Markup Language.
  (Markups instruct the software displaying the text)

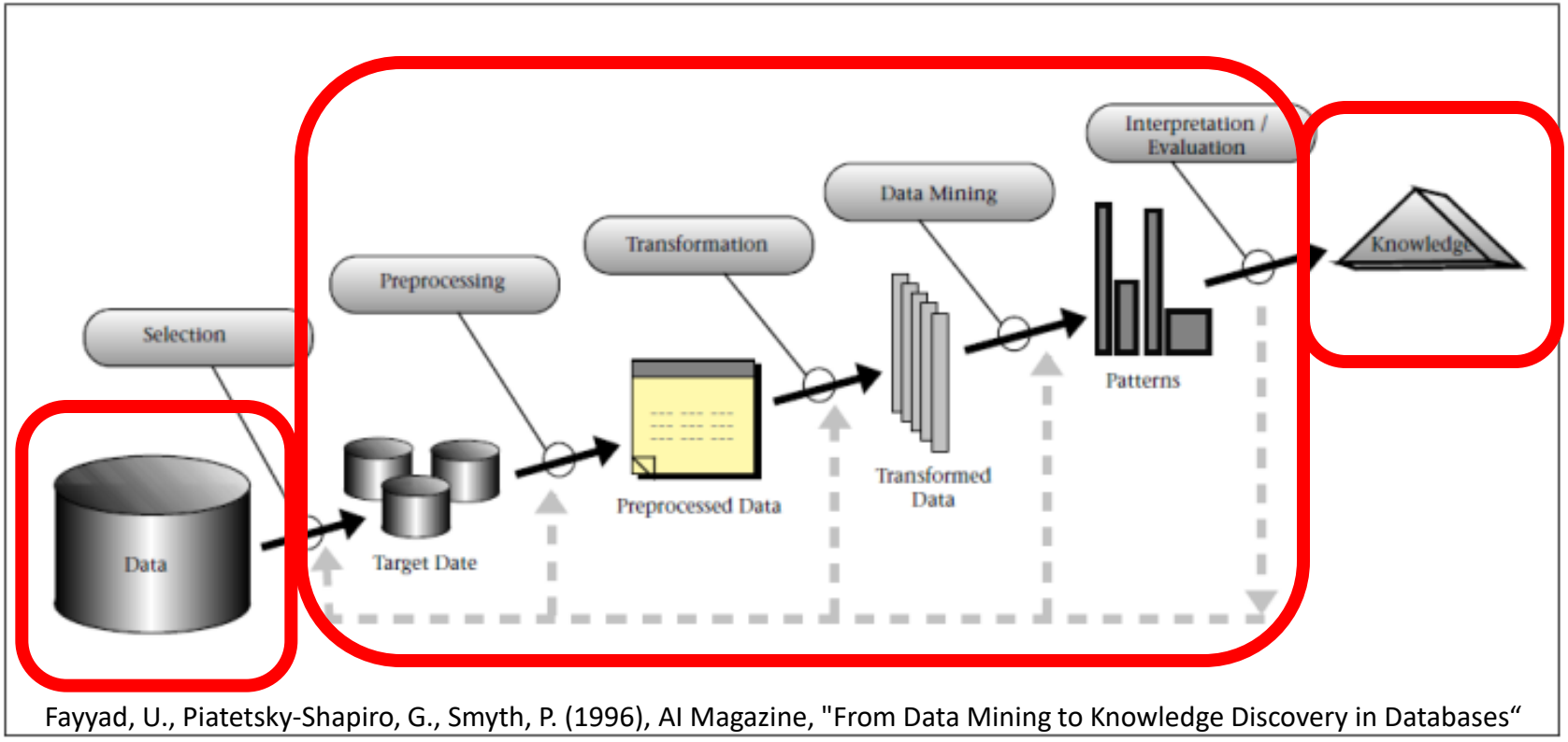
- **ASN.1** – Abstract Syntax Notation One (an International Standards Organization (ISO) format).
  NCBI uses ASN.1 for data storage and retrieval.

Uniprot schema

http://www.uniprot.org/uniprot/P01542



Knowledge Discovery in Databases (KDD)



Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996), AI Magazine, "From Data Mining to Knowledge Discovery in Databases"

Figure 1. An Overview of the Steps That Compose the KDD Process.

Data are being collected and accumulated at a dramatic pace.

→ **A new generation of computational theories and tools is needed.**

→ **These theories and tools constitute the emerging field of Knowledge Discovery in Databases.**