

Introducción a MySQL y bases de datos modelo

Clara Isabel Bermúdez Santana.
Departamento de Biología
Universidad Nacional de Colombia

August 16, 2018

Principales bases de datos y navegadores genómicos

- Genbank: <https://www.ncbi.nlm.nih.gov/genbank/>
- Ensembl: <http://www.ensembl.org/index.html>
- UCSC Genome Browser: <https://genome.ucsc.edu/>
- DNA Data Bank of Japan: <http://www.ddbj.nig.ac.jp/>

Consulta general de un gen

- Ensembl: <http://www.ensembl.org/index.html>





The Ensembl Database Schema



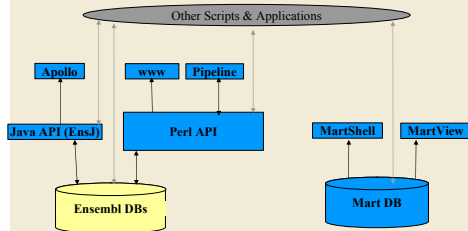
European Bioinformatics Institute



Requirements for the schema


- Store data for human genome
- ... and all the other genomes we have
- ... and all the genomes we might get
- Flexible to add more data
- Easy to adapt to new genome
- Responds fast enough for web site display and pipelined genebuild

System Context

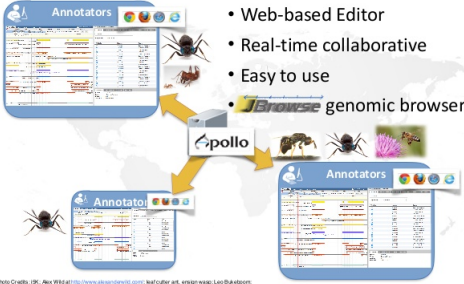


- Ensembl DB: Dos componentes:
 - Asociado con anotación de genes soportada por Apollo e implementada en JAVA
 - Asociada al manejo de toda la información genómica disponible, implementada en Perl
- Mart DB

Anotación genómica (JAVA): utiliza API (Application Programming Interface): desde las APIs las aplicaciones siguen reglas para permitir la comunicación con otras aplicaciones: por ejemplo la comunicación entre Ensembl DB y los programas de anotación. la base de datos se comunica con otros softwares por ejemplo los de anotación.

 genomearchitect.org

Apollo is a Tool for Collaborative Annotation




- Web-based Editor
- Real-time collaborative
- Easy to use
-  **JBrowse** genomic browser

Photo Credits: (l-r): Alex Wild at www.genomearchitect.org; Isotoular art; ensign-warp; Leo Bakaborn; Nascara xiliponja/pavel-novak; Wikimedia Commons; Apis mellea honey bee; Marc Martini USDA/ARS Fort Keogh LARRL; Blos laurus.com.

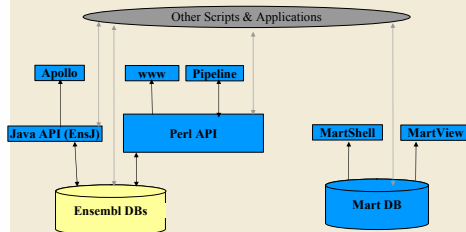
9

Estructura general del Ensembl: Componente Mart: Es una herramienta usada en el browser genómico para exportar data sin previo conocimiento de programación

e!
Ensembl



System Context



Este componente se basa en BioMart: las búsquedas de datos son virtualizadas en una especie de base de datos virtual "Data Federation Technology"

The screenshot shows the BioMart website interface. The browser address bar displays www.biomart.org. The left sidebar contains a navigation menu with the following items: HOME, TOOLS, WEB SERVICE, COMMUNITY, PUBLICATIONS, NEWS, CREDITS, DOCUMENTATION, VERSION 0.7, and CONTACT. The main content area features a header with the BioMart logo and a navigation bar with a "Download" button labeled "v 0.9". Below the header, a section titled "BioMart" describes the project as a community-driven effort to provide unified access to distributed research data. A paragraph explains that the project provides free software and data services to the international scientific community to foster collaboration and facilitate the scientific discovery process. Below this text are four icons representing different data services: BROWSE DATA, ID CONVERSION, SEQUENCE RETRIEVAL, and ENRICHMENT ANALYSIS. To the right, a "JOIN OUR COMMUNITY" section lists three steps: 1. Set up your own data source with a click of a button, 2. Expose your data to a world wide scientific community through BioMart Portal, and 3. Federate your local data with data from other community members. A "DOWNLOAD NOW" button is provided. At the bottom, a section titled "OUR GROWING COMMUNITY" features a world map with a yellow circle highlighting the UK, and a list of participating institutions: BCCTB Bioinformatics Portal, COSMIC, DAPPER, EMAGE, Ensembl, Ensembl Genomes, Europhenome, GWAS Central, HGNC, InterPro, Pancreatic Expression Database, Rfam, UniProt, VEGA, and WormBase ParaSite.

bio::mart

[4th January 2016] OASIS in Nature Methods

Download v 0.9

BioMart

is a community-driven project to provide unified access to distributed research data to facilitate the scientific discovery process.

The BioMart project provides free software and data services to the international scientific community in order to foster scientific collaboration and facilitate the scientific discovery process. The project adheres to the open source philosophy that promotes collaboration and code reuse.

JOIN OUR COMMUNITY

1. Set up your own data source with a click of a button
2. Expose your data to a world wide scientific community through BioMart Portal.
3. Federate your local data with data from other community members

DOWNLOAD OUR SOFTWARE TO JOIN [DOWNLOAD NOW](#)

OUR GROWING COMMUNITY

A large number of servers that provide access to a wide range of research data have been set up by the BioMart community. Using BioMart's unique data federation technology, a Central Portal was established to provide a convenient single point of access to all of these data, which is distributed worldwide.

UK

BCCTB Bioinformatics Portal, COSMIC, DAPPER, EMAGE, Ensembl, Ensembl Genomes, Europhenome, GWAS Central, HGNC, InterPro, Pancreatic Expression Database, Rfam, UniProt, VEGA, WormBase ParaSite

BioMart en el ensembl

The screenshot shows the Ensembl Genomes website with the BioMart section highlighted. The top navigation bar includes links for 'About us', 'Genomes', 'Data types', 'Data access', and 'FAQs'. The 'Data access' section is expanded, showing a list of links: Overview, Data export, Data download, Data archives, BioMart, Using your own data, Public Track Hubs, Programmatic access, Ensembl Perl API, Ensembl Genomes Perl API, REST service, MySQL database access, Linking to Ensembl Genomes, Building an Ensembl Genomes mirror, and Virtual Machine. The main content area features the BioMart logo and a paragraph explaining that Ensembl Genomes supports downloading correlation tables via the BioMart data mining tool, which is more user-friendly than extracting information from database dumps. It also notes that BioMart is not currently available for Ensembl Bacteria. Below this, there are links to Ensembl Protista BioMart, Ensembl Fungi BioMart, Ensembl Metazoa BioMart, and Ensembl Plants BioMart. A 'Help and documentation' section includes links to BioMart tutorials and videos, and Programmatic access with MartService (external link).

EnsemblGenomes

About us | Genomes | Data types | Data access | FAQs

Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Data access

- Overview
- Data export
- Data download
 - Data archives
- BioMart
- Using your own data
- Public Track Hubs
- Programmatic access
 - Ensembl Perl API
 - Ensembl Genomes Perl API
 - REST service
 - MySQL database access
- Linking to Ensembl Genomes
- Building an Ensembl Genomes mirror
- Virtual Machine

bio·mart

Ensembl Genomes supports downloading of many more correlation tables via the highly customisable [BioMart data mining tool](#). You may find exploring this web-based data mining tool easier than extracting information from our normalised database dumps. Note that BioMart is not currently available for Ensembl Bacteria.

- [Ensembl Protista BioMart](#)
- [Ensembl Fungi BioMart](#)
- [Ensembl Metazoa BioMart](#)
- [Ensembl Plants BioMart](#)

Help and documentation:

- [BioMart tutorials and videos](#)
- [Programmatic access with MartService](#) (external link)

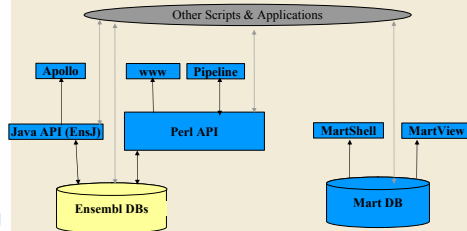
<https://www.ensembl.org/biomart/martview/13ca37370cf587a225d62c0365c72c7a>

Estructura general del Ensembl: Componente Perl y API para permitir comunicación entre la base de datos y otras aplicaciones no asociadas con anotación genómica

e!
Ensembl



System Context



Consulta básica de tablas en el Ensembl

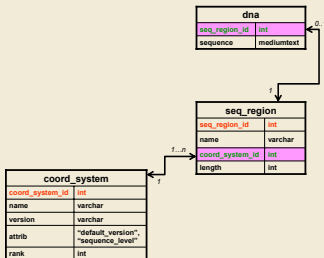
https://www.ensembl.org/info/docs/api/core/core_schema.html#coord_system



Sequence regions

- Everything which represents a length of nucleotide sequence is a sequence region.
 - chromosome, BAC-clone, supercontig, scaffold, contig ...
- Sequence regions of the same type belong to the same coordinate system.
 - “1”, “2”, and “3” are sequence regions with coordinate system “chromosome”
- Sequence regions have names and lengths.

Sequence regions





Example sequence region

- Chromosome, 1, 200MB
- Clone, AL123123.4, 132KB
- NT_contig, NT_1245675, 17MB
- Contig, AC332232.1.1.123223, 123223

Coordinate system

- The coord_system describes the type of the sequence region
 - Name (“chromosome”, “contig”,...)
 - Version (eg. NCBI35, ZFISH3)
 - Internal id (coord_system_id)
 - Attrib – (default, sequence_level)
 - rank (1..n)
- If you have 2 coordinate systems with the same name, choose a “default” one. They need to have different versions (NCBI34, NCBI35).
- The lower the rank, the bigger the sequence region. Choose 1 for your biggest regions (chromosomes).
- Only one coordinate system is allowed to contain sequence regions with actual sequence attached. Flag it with Attrib = sequence_level.

Coordinate system

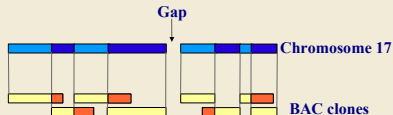
- “contig”
 - Contiguous sequence.
 - “N”s should be rare and of short length.
 - Can serve as your basic sequence holder
- “clone”
 - Should have a real BAC or PAC or maybe YAC behind it.
 - Might not be contiguous
- “supercontig”
 - Assembled from smaller contiguous sequences.
 - May have small gaps (eg between read pairs)
- “chromosome”
 - Use it only for real chromosomes.
 - or for alternative sequences of reference chromosomes.
- “chunk”
 - Artificial coordinate system to hold sequence regions for technical reasons.
 - Create, when none of the other coordinate systems can hold your sequence (eg. You only have full length chromosomes as coordinate system but they are to long to store)
 - or when you have 2 real sequence containing coordinate systems.



Assemblies

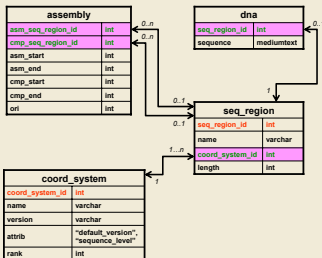
- An assembly defines how sequence regions in one coordinate system are made up of sequence regions from another coordinate system.
- For example human chromosomes are assembled from a “tiling path” of BAC clones.
- Assembly information stored in Ensembl makes it possible to obtain features or sequence from arbitrary sequence regions.

Assemblies



- A row in the assembly table references an assembled and component sequence region.
- How a piece of the assembled sequence region is made from a piece of a component region is defined by a pair of coordinates and an orientation.
- Gaps are represented by the absence of assembly information.

The assembly table





Sequence region attributes

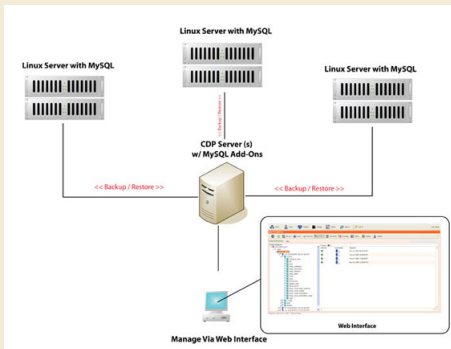
- Arbitrary attributes may be associated with a sequence region via the `seq_region_attr` table.
 - sanger ids for certain clones.
 - htg phases for clones.

MySQL: es un sistema de gestión de bases de datos relacional desarrollado bajo licencia dual: Licencia pública general/Licencia comercial por Oracle Corporation es el manejador de bases de datos mas popular y hace parte del área de los sistemas de manejo de bases relacionales RDBMs. El nucleo de lenguaje que lo conforma es SQL (Structured Query Language)



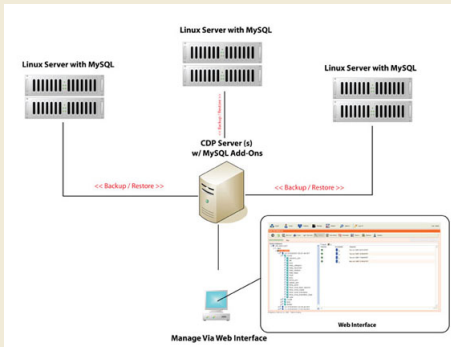
El modelo de bases relacionales fue creado por Edgar Frank Codd un investigador de la IBM en 1970

Qué permite el modelo de manejo de bases de datos relacionales?:
permite representar sofisticadas relaciones entre items de datos y
calcular estas relaciones con la velocidad necesaria para tomar decisiones en organizaciones modernas



El modelo de bases relacionales fue creado por Edgar Frank Codd un investigador de la IBM en 1970

Es sorprendente como se puede ir del diseño a la implementación en pocas horas y cuan facilmente uno puede desarrollar aplicaciones en la web para acceder a terabytes de datos y atender miles de usuarios web por segundo



Componentes básicos de instalación

- Server: que maneja los datos.
- Clientes: que solicitan al server que hacer con los datos. No son los usuarios o cuentas
- El cliente mas común es el programa MySQL monitor, le permite conectarse al servidor MySQL y consultar usando SQL
- Otros clientes son mysqladmin program, que permite como cliente gestionar labores de administración.
- Conclusión: cualquier programa que conoce como hablar a MySQL server es un cliente

Aspectos a tener en cuenta son:

- MySQL server
- Modulos especializados para manejo de bases de datos (DBMS or Data Base Managment System) en Perl.
- Lenguajes de programación de proposito general como Perl o PHP sirven de esquemas para comunicación con el lenguaje propio de la WENB que es HTML.
- Aplicaciones de bases de datos en la WEB.

Diseño de bases de datos; que no debemos hacer

Ejemplo: usted quiere crear una base de datos para almacenar las notas de sus estudiantes

for each of his courses.

GivenNames	Surname	CourseName	Pctg
John Paul	Bloggs	Web Database Applications	72
Sarah	Doe	Programming 1	87
John Paul	Bloggs	Computing Mathematics	43
John Paul	Bloggs	Computing Mathematics	65
Sarah	Doe	Web Database Applications	65
Susan	Smith	Computing Mathematics	75
Susan	Smith	Programming 1	55
Susan	Smith	Computing Mathematics	80

Diseño de bases de datos; que no debemos hacer

Ejemplo: usted quiere crear una base de datos para almacenar las notas de sus estudiantes, pero que pasa si hay nombres replicados

StudentID	GivenNames	Surname	CourseName	Pctg
12345678	John Paul	Bloggs	Web Database Applications	72
12345121	Sarah	Doe	Programming 1	87
12345678	John Paul	Bloggs	Computing Mathematics	43
12345678	John Paul	Bloggs	Computing Mathematics	65
12345121	Sarah	Doe	Web Database Applications	65
12345876	Susan	Smith	Computing Mathematics	75
12345876	Susan	Smith	Programming 1	55
12345303	Susan	Smith	Computing Mathematics	80

diseño de bases de datos; que no debemos hacer

el señor Jhon reprobó un primer examen y aprobó un segundo examen. Qué nos falta incluir? = el tiempo cuando se hizo la prueba

StudentID	GivenNames	Surname	CourseName	Year	Sem	Pctg
12345678	John Paul	Bloggs	Web Database Applications	2004	2	72
12345121	Sarah	Doe	Programming 1	2006	1	87
12345678	John Paul	Bloggs	Computing Mathematics	2005	2	43
12345678	John Paul	Bloggs	Computing Mathematics	2006	1	65
12345121	Sarah	Doe	Web Database Applications	2006	1	65
12345876	Susan	Smith	Computing Mathematics	2005	1	75
12345876	Susan	Smith	Programming 1	2005	2	55
12345303	Susan	Smith	Computing Mathematics	2006	1	80

diseño de bases de datos; que no debemos hacer

Ahora tenemos datos inflados en la base de datos. Podríamos crear una tabla alterna llamada Studentdetails

StudentID	GivenNames	Surname
12345121	Sarah	Doe
12345303	Susan	Smith
12345678	John Paul	Bloggs

diseño de bases de datos; que no debemos hacer

Ahora tenemos datos inflados en la base de datos. Podríamos crear una tabla alterna llamada Studentgrades

StudentID	CourseName	Year	Sem	Pctg
12345678	Web Database Applications	2004	2	72
12345121	Programming 1	2006	1	87
12345678	Computing Mathematics	2005	2	43
12345678	Computing Mathematics	2006	1	65
12345121	Web Database Applications	2006	1	65
12345876	Computing Mathematics	2005	1	75
12345876	Programming 1	2005	2	55
12345303	Computing Mathematics	2006	1	80

diseño de bases de datos; que no debemos hacer

Sin embargo aun resta incluir información como cuando ingresó al colegio, su dirección etc.

El diseño de una base de datos no es una tarea fácil.

Pasos principales para el diseño de una base de datos

- Análisis de requisitos: que se necesita en la base de datos y para que necesitamos la base de datos, hablar con los usuarios para saber que requieren y que tipo de interacciones se dan entre los datos.
- Diseño Conceptual: una vez sepamos cuales son los requisitos, los debemos describir en un formato de diseño conceptual.
- Diseño lógico: mapear el diseño conceptual en un código que permita el manejo de la base de datos.
- Diseñar las tablas .

Tipo de elementos requeridos para la construcción de una base de datos relacional

Entidades u objetos

Relaciones entre los objetos o asociaciones

Cual es el objetivo? establecer relaciones o asociaciones entre objetos que son también llamados "entidades "

Supongamos que una oficina de negocios está interesada en manejar su grupo de clientes con el objetivo de realizar negociaciones con ellos

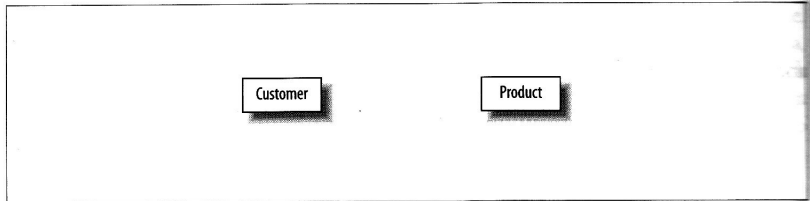


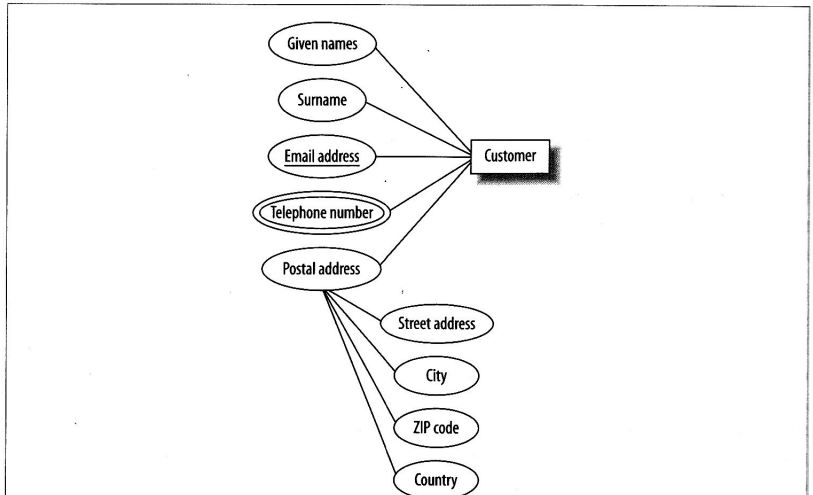
Figure 4-1. An entity set is represented by a named rectangle

Para describir cada una de nuestras entidades usamos atributos que las caracterizan o las describen y éstas a su vez permiten diferenciar las entidades

Los atributos se clasifican como: simples o compuestos. Por ejemplo la dirección postal es un atributo compuesto

Los atributos pueden ser univaluados o multivaluado. Por ejemplo: número telefónico puede poseer varios valores. Estos se representan como doble valor

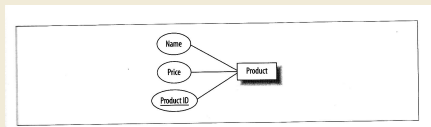
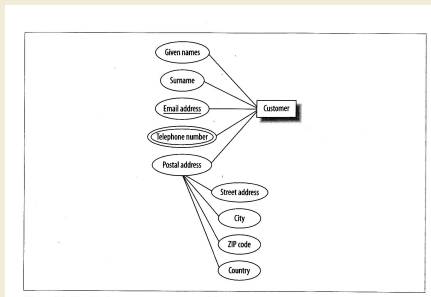
Para describir cada una de nuestras entidades usamos atributos que las caracterizan o las describen



Adicionalmente necesitamos un atributo o la combinación mínima de atributos que garanticen ser únicos para cada cliente. Eso se conoce como definir una clave o key.

Del conjunto total de atributos candidatas a ser claves, se escoge uno que va a ser nuestra clave primaria, ésta es muy importante ya que va a identificar de manera única los registros a través de toda la estructura de la base de datos. Este atributo será único para cada individuo de la entidad y no puede estar vacío. En el diagrama los atributos que son la clave primaria están sobresaltados por una línea.

Para describir cada una de nuestras entidades usamos atributos que las caracterizan o las describen



Los valores que pueden tomar los atributos se escogen de un dominio de valores permitidos

- Para nombres y apellidos: cadenas de un tamaño definido (por ejemplo una cadena de 100 caracteres)
- Número telefónico: cadenas de un tamaño definido (por ejemplo una cadena de 40 caracteres)
- Precio de un producto: valor real () .

Los valores que pueden tomar los atributos se escogen de un dominio de valores permitidos

- Los atributos puede tener valores vacios, por ejemplo un cliente puede no querer dar su número telefónico.
- Los atributos que son la clave primaria no pueden estar vacios. (NOT NULL)

Ejemplos de claves primarias: atributos artificiales que nos podemos inventar

- El código del estudiante.
- Números de seguridad social
- Número de la licencia de conducir
- Id del producto.
- Id del gen en el GenBank
- Id de la región en que se ubica un gen

Algunos ejemplos de relaciones entre entidades

- Relación comprar: Un cliente puede comprar un Producto.
- Relación tomar o asistir a un curso: Un estudiante puede tomar un curso
- Relación grabar: Un artista puede grabar un album.
- Relación activar: Un gen puede activar la expresión de un gen o varios genes
- Relación estar contenido: Un secuencia de DNA está contenida en una región de una secuencia
- Relación estar asociado: Una región de secuencia se asocia con sistemas de coordenadas

Clasificación de las relaciones

- One to One : 1:1. Ejemplo: número serial de un carro. Uno por carro y cada Carro tiene uno. Una secuencia de un gen tiene asociado una región secuencia
- One to many: 1:M. Un persona tiene varias tarjetas de crédito, pero cada tarjeta pertenece a una sola persona. Una región secuencia puede tener varios sistemas de coordenadas, pero un sistema de coordenadas le pertenece a una única región de la secuencia
- Many to Many: M:N. Cada cliente puede comprar diferentes productos y cada producto puede ser comprado por diferentes clientes

- El número de entidades sobre cada extremo de la relación define una especie de limitaciones de la relación.
- En el diagrama de modelo de relaciones de entidades (ER), el conjunto relaciones se representa como un diamante.
- Las relaciones tambien pueden tener atributos para identificarlas.

Representación del ER para una oficina de negocios

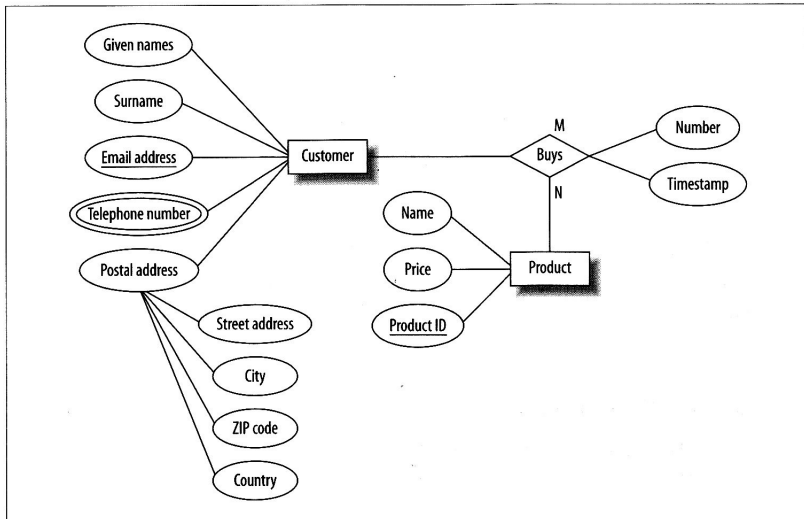
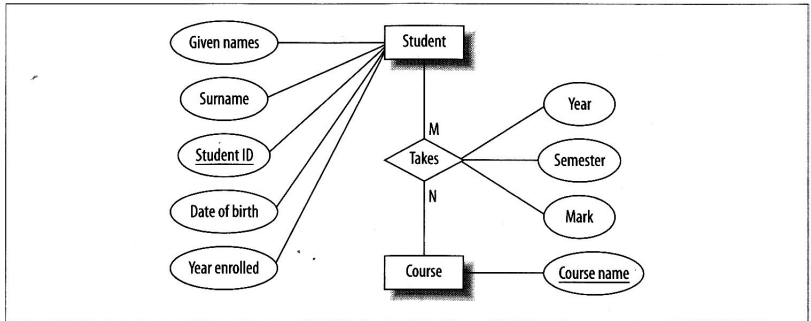


Diagrama ER parcial del caso de la Universidad:
identifique: entidades, la clave primaria, las relaciones, los
atributos de las relaciones y la cardinalidad de relación



Las relaciones entre las entidades pueden ser parciales o totales: restricciones de la participación en el ER

- Caso 1 Una persona es considerada un cliente solo si ha comprado un producto (Participación total)
- Caso 2 Una persona es considerada un cliente por que se tiene la esperanza de que algun dia compre algo (Participación parcial)
- Si son totales se representan como doble linea entre la entidad y la relación.

Aspectos generales para definir los roles en la bases de datos

Es entidad o atributo

Es entidad o relación

Aspectos generales para definir los roles en la bases de datos: es entidad o atributo

- El interés del objeto en la base de datos: que relaciono. Clientes-email?
- Tiene el item componentes de si mismo, buscar maneras de representarlo como otra entidad.
- Identificar si los atributos toman valores vacios "por ejemplo" solo para algunos cursos.

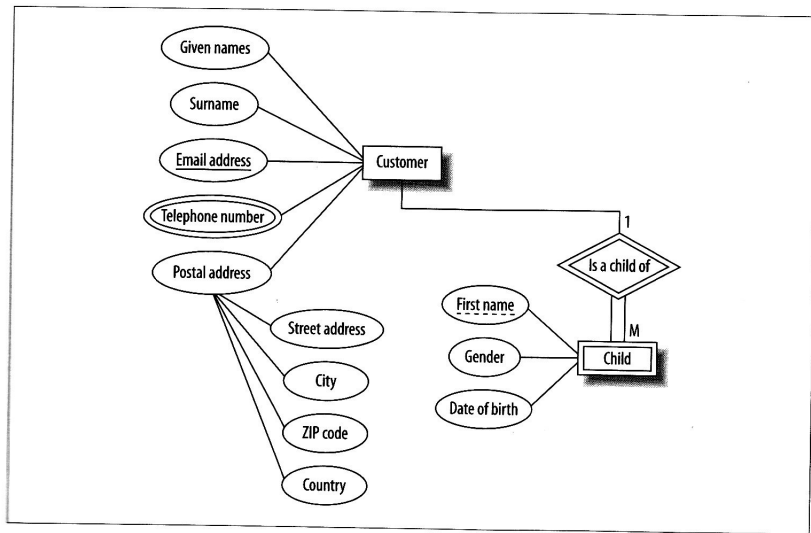
Aspectos generales para definir los roles en la bases de datos: es entidad o relación

- Mapear Sustantivos a Entidades y verbos a Relaciones
- Por ejemplo: un programa está constituido por uno o mas cursos
- Un estudiante se enrolla en un programa:

Estrategias para trabajar con menor información: construcción de entidades débiles que son dependientes de las entidades fuertes

- Si conocemos el contexto, podemos trabajar con una menor cantidad de información.
- Si queremos por ejemplo guardar la información de los hijos de los clientes, podemos crear una nueva entidad, que no puede existir independientemente de su entidad mayor o fuerte.
- La entidad debil tiene una relación con la entidad fuerte por medio de una relación llamada de identificación.
- Se combinan la clave parcial de la entidad débil con la clave de la entidad mayor de la cual depende.

Contruyendo entidades intermedias



Resumen del modelo general

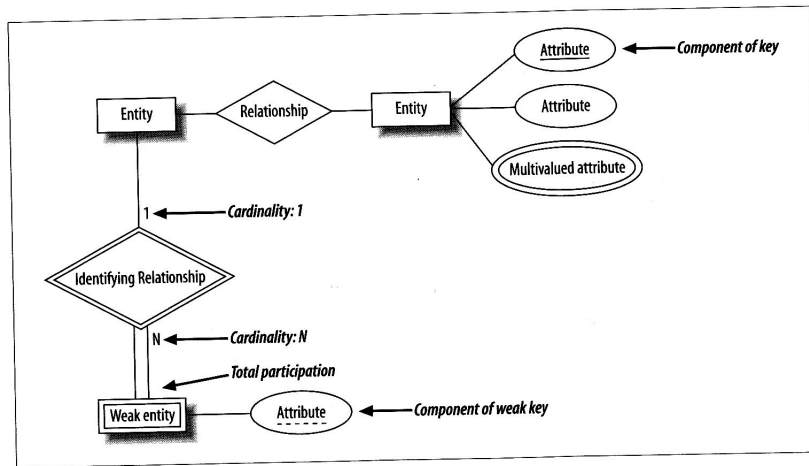
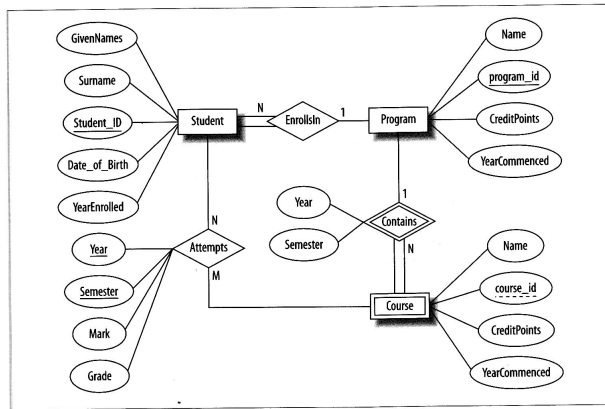
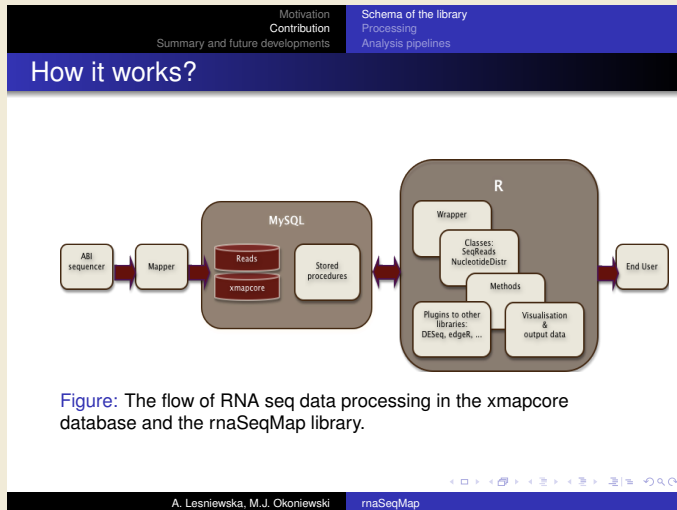


Diagrama ER para el caso de la Universidad: identifique: entidades, la clave primaria, las relaciones, los atributos de las relaciones y la cardinalidad de relación



Ejemplo de aplicabilidad en el almacenamiento de información de datos de RNAseq



Diseño base conceptual music.sql

