Nociones de bases de datos e informatización de colecciones

Busturia, 28-30 de octubre 2009

Francisco Pando, GBIF-ES

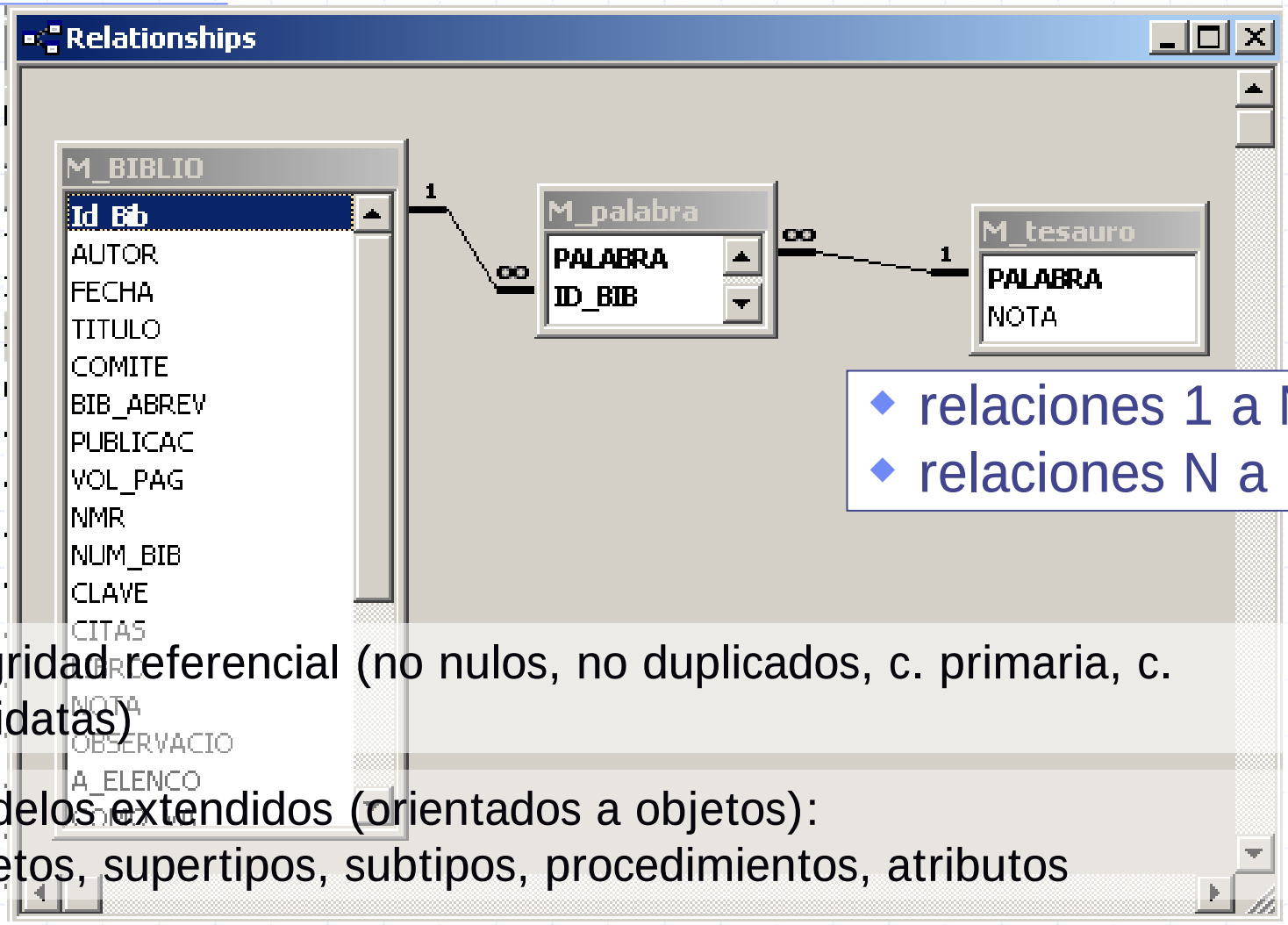
Informatización de colecciones: una visión general

- La elaboración de bases de datos es una labor costosa cuyos resultados no se obtienen sino a largo plazo.
- Es mejor proceder a informatizar las colecciones de manera completa por grupos taxonómicos, a informatizarlas parcialmente en su conjunto.
  - Contribuir a explotar de una manera más completa la información que contiene
  - Proteger el material conservado
  - Contribuir a la gestión del herbario
- Ligar la informatización de colecciones a proyectos de investigación en curso que, de alguna manera, precisen de la información contenida en las mismas.
- El crecimiento y mantenimiento de la base de datos debe formar parte de la rutina de trabajo del herbario.
- Mantener la información original y hacerlo de manera que se distinga claramente de la información derivada.
- El valor de la base de datos aumenta con la posibilidad de que los datos se puedan combinar e integrar con otras fuentes (nombre científico, coordenadas).
- No perder de vista el contexto y las prioridades: conservar, dar acceso, informatizar. La base de datos es el medio, no el fin.

Qué es una base de datos

- “conjunto de datos almacenados con una estructura lógica. Es decir, tan importante como los datos, es la estructura conceptual con la que se relacionan entre ellos. En la práctica, podemos pensar esto como el conjunto de datos más los programas (o *software*) que hacen de ellos un conjunto consistente”

Tablas y relaciones



- relaciones 1 a N
- relaciones N a M

Integridad referencial (no nulos, no duplicados, c. primaria, c. candidatas)

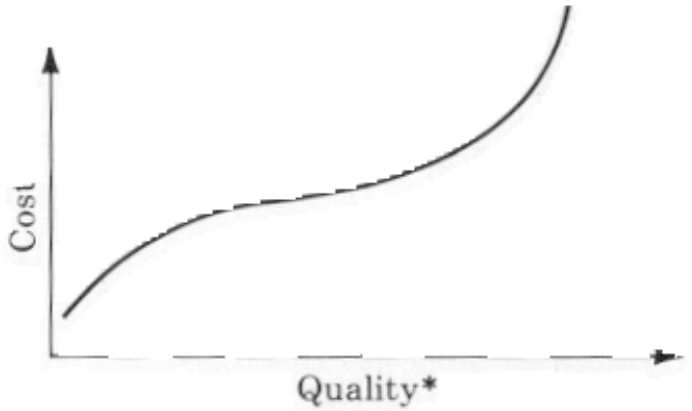
modelos extendidos (orientados a objetos):
objetos, supertipos, subtipos, procedimientos, atributos

Códigos abreviaciones y estándares

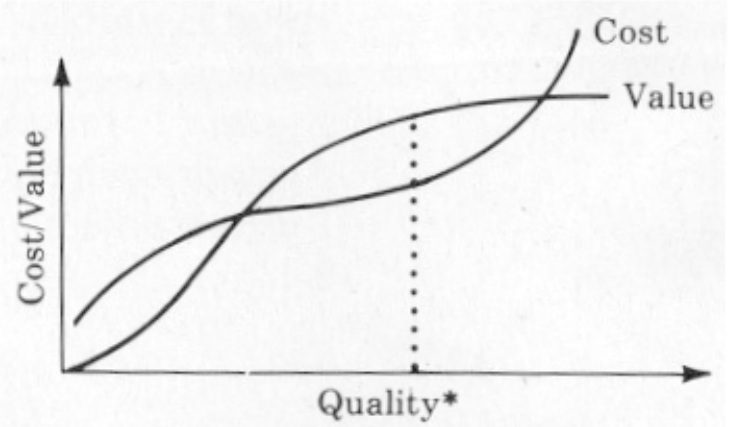
- nombres entendibles para campos y tablas; fáciles de recordar y usar; abreviar lo mínimo y de manera consitente.
- códigos arbitrarios => problemas de compatibilidad y fuente de errores
- el coste:
- Razones para códigos en las claves primarias: cuando la clave primaria es my larga o compleja (¿para salvar espacio en disco? quien se lo cree)
- estándares => compatibilidad => acceso unificado, interoperabilidad, soluciones comunes

Coste

(a) Information Quality Versus Value



(b) Information Quality Versus Cost



(c) Optimal Information Set

***Information quality = *F*(detail, age, accuracy, relevance)**

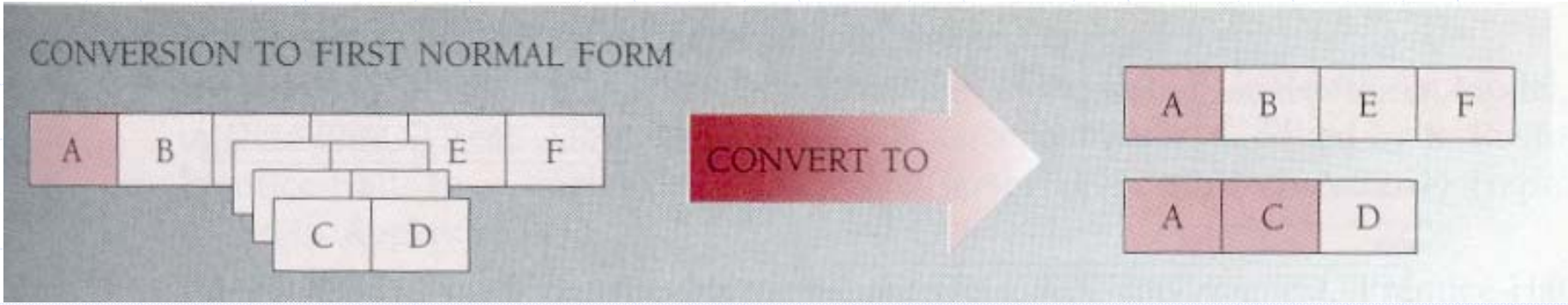
Normalización y modelo lógico de los datos

Al modelar una base de datos, desearemos evitar puntos que crean confusión, duplicación de la información y por ende, un mal funcionamiento y exploración de la información. Entre las propiedades indeseables en un diseño de bases de datos tenemos:

- Redundancia en la información.
- Incapacidad de representar cierta información.
- Registrar información que no sea identificable.

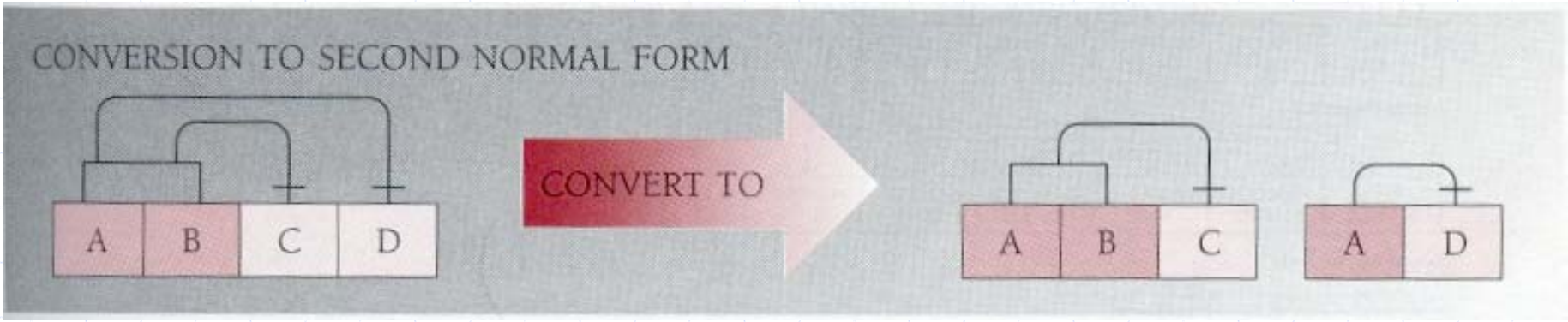
Primera forma normal

- Una relación está en primera forma normal (1FN) si y sólo si todos los dominios son atómicos. Un dominio es atómico si los elementos del dominio son indivisibles.
- Es decir, no tenemos grupos de repetición o un conjunto de valores asociados repetidos asociados a una misma tupla.
- Datos separados en tablas, cada tabla con su clave primaria, no hay grupos repetitivos
- ej. identificaciones separadas del resto -> tabla de identificaciones aparte



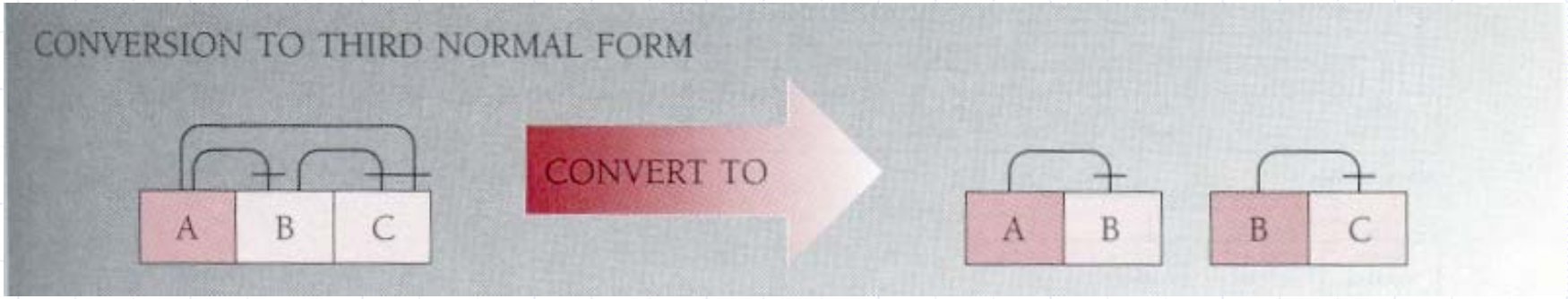
Segunda forma normal

- Una relación está en segunda forma normal (2FN) si y sólo si está en 1FN y todos los atributos que no sean llaves dependen por completo de llave primaria.
- Quita todos los campos que no dependen de la clave principal o que dependen solo de parte de la clave principal
- ej. nombre vernáculo en registro de pliego -> tabla de nombres vernáculos relacionada con el nombre científico



Tercera forma normal

- Una relación están en tercera forma normal (3FN) si y sólo si están en 2FN y todos los atributos no llave dependen de manera no transitiva de la llave primaria. Se dice que existe una dependecia transitiva cuando tenemos el par de dependencias funcionales: 78#78 y 79#79, porque de ellas se sigue que 80#80.
- Elimina todo aquello en las tablas que que no dependa únicamente de la clave principal
- Ej. En una base de datos de pliegos donde para todos ellos disponemos de las coordenadas, los datos del sitio de recolección (pais, provincia, localidad, municipio, ...) en la tabla de especímenes



Adaptado de:

F. Pando (2003). Introducción al taller de informatización de colecciones botánicas.

http://www.ahim.org/taller_valencia_2003/taller_valencia_2003.htm