

CALIDAD EN DATOS DURANTE LA DIGITALIZACIÓN Y EL ALMACENAMIENTO

**TALLER GBIF.ES
CALIDAD EN BASES DE DATOS DE
BIODIVERSIDAD**

Madrid, 2-4 Diciembre 2015



Francisco Pando
(sobre presentación de Katia Cezón)
Unidad de Coordinación – GBIF.ES
Nodo Nacional de Información en
Biodiversidad en España

DIGITALIZACIÓN Y EL ALMACENAMIENTO



1
Planificación



2 Recolección



3 Documentación



4
Digitalización



5
Control
de
calidad



6
Publicación
en internet

DIGITALIZACIÓN Y EL ALMACENAMIENTO

La forma de almacenar la información en una base de datos puede tener un efecto en la calidad de los mismos y tiene que ver con varios aspectos:

- Modo en el que se capturan electrónicamente los datos y *diseño de la base de datos*.
- **Archivo**: archivar datos en servidores accesibles para diversos responsables de la organización, y documentar dónde está cada base de datos y su contenido: incluyendo tanto datos obsoletos como actuales.
- **Backups**: la realización regular de copias de seguridad evita la pérdida de datos y garantiza unos niveles de calidad.





CAPTURA ELECTRÓNICA DE DATOS

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS

El nivel de error debido a la entrada de datos en la base de datos se puede disminuir a través de:

- Realización de un buen **diseño de la base de datos**. Interfaces de usuario que minimicen la entrada de errores: campos que se chequeen contra tablas de referencias, tablas de estándares, listas desplegables con valores predeterminados, etc.
- Informatización por parte de personal que haya tenido una **formación previa y supervisión de expertos** que lleven a cabo un testeo.

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS



Una base de datos debe estar diseñada de tal modo que permita comparar y analizar sus datos, para ello debemos buscar en la medida de lo posible, su **estandarización** y **normalización**. Además debe contar con una **buena interface** para facilitar la tarea de la introducción de datos.

Mediante la **estandarización** buscamos el manejar datos consistentes y de calidad.

Esto se puede conseguir:

1. Separando la información en diferentes campos.
2. Utilizando estándares, vocabularios controlados, estableciendo criterios y protocolos.

La **normalización** de bases de datos consiste en aplicar un modelo relacional en el que la información esté repartida en diferentes tablas.

1. Evitar la redundancia de los datos.
2. Disminuir problemas de actualización de los datos.
3. Protege la integridad de los datos.

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS



Syntrichia ruralis (Hedw.) F.
Weber D. Mohr.

subsp. *ruraliformis* (Bsch.)
Delonge

Género: *Syntrichia*

Especie: *ruralis*

Autor especie: (Hedw.) F. Weber D.
Mohr.

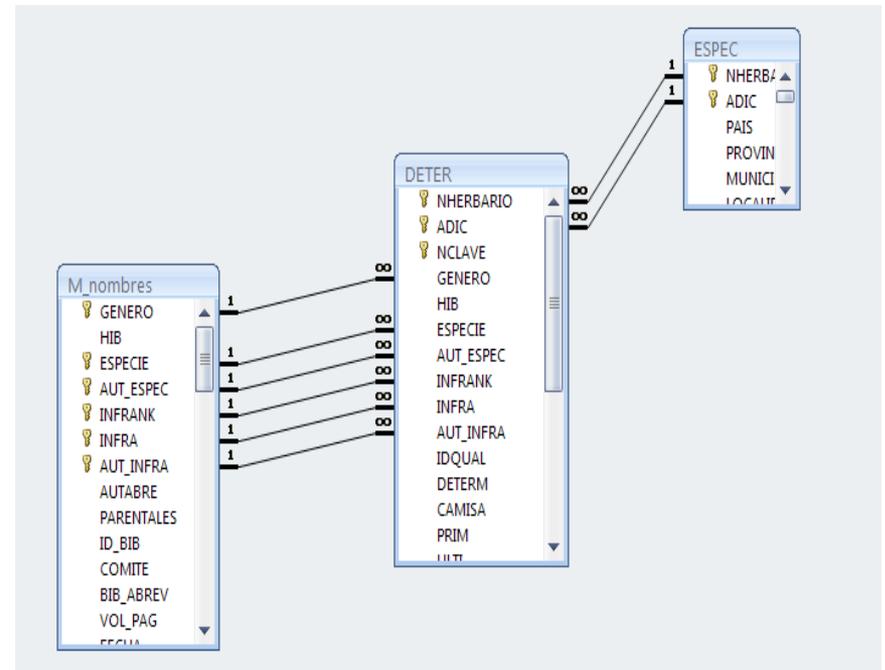
Rango infraespecífico: subsp.

Autor rango infraespecífico: (Bsch.)

Delonge

ATOMIZACIÓN DE LA
INFORMACIÓN

Un ejemplo con nombres científicos



NORMALIZACIÓN

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS

Un ejemplo con nombres científicos

NHERBARIO	<input type="text"/>	<input type="text" value="1"/>	<input type="text" value="1"/>							
GÉNERO	<input type="text" value="Grimmia"/>	<input type="button" value="N"/>		GRUPO	<input type="text"/>			Chequear Nombre	<input type="text" value="NO"/>	
ESPECIE	<input type="text" value="pulvinata"/>		HIB	<input type="text" value="-"/>						
AUT_ESPEC	<input type="text" value="Hedw."/>			INFRANK	<input type="text" value="-"/>			INFRA	<input type="text" value="-"/>	
AUT_INFRA	<input type="text" value="-"/>			IDQUAL	<input type="text" value="-"/>	<input type="button" value="+"/>		MESAN	<input type="text" value="-"/>	
DETERM	<input type="text"/>	<input type="button" value="v"/>		ES_TIPO	<input type="text"/>	<input type="button" value="+"/>		TAXONOMÍA	<input type="text"/>	
OBSERV	<input type="text"/>						CAMISA	<input type="text" value="S"/>	<input type="button" value="v"/>	
							PRIM	<input checked="" type="checkbox"/>	ULT	<input checked="" type="checkbox"/>

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS

Un ejemplo con localidades



ESPAÑA. Segovia: Basardilla, 2,5 Km. al SO por la Ctra. Sg-160, muy próximo al poblado de Don Toribio, en el cruce del puente sobre el río Pirón, 1470 m, 41°2'1.78"N 4° 1'0.56"W, WGS84.

ATOMIZACIÓN Y
ESTANDARIZACIÓN DE LA
INFORMACIÓN

País: España - ISO 3166-1 ES

Provincia: Segovia - ISO 3166-2 Sg

Municipio: Basardilla

Localidad principal: , 2,5 Km. al SO por la Ctra. ...

Altitud (punto o rango): 1470

Coordenadas: 41°2'1.78"N 4° 1'0.56"W

Datum: WGS84

Incertidumbre: 0

CAPTURA ELECTRÓNICA DE DATOS

DISEÑO DE LA BASE DE DATOS

- Información fácilmente consultable y accesible.
- Que se pueda sacar de la base de datos/aplicación.
- Que pase el “test del autobús”



CAPTURA ELECTRÓNICA DE DATOS

INFORMATIZACIÓN

El nivel de error debido a la digitalización de datos en la base de datos se puede disminuir a través de:

- 1 Formación del personal en el manejo de software y/o base de datos para la gestión de la colección de datos y familiarizados con la terminología.
- 2 Separación de tareas y actividades para optimizar y hacer más eficientes algunos procesos (ej. georreferenciación).
- 3 Procesos de testeo por parte personal especializado.
- 4 Feedback

CAPTURA ELECTRÓNICA DE DATOS

INFORMATIZACIÓN

Pautas o modelos de error (CONABIO): un ejemplo con fechas

- Omisión (información ausente e incompleta)
 - Ausencia de total de información: campo vacío.
 - Expresiones que indican falta del dato: 0, “_”, “-”, “unkown”, “n.d.”, “none”
 - Dato incompleto por falta del año: “Septiembre”, “4 Mayo”, etc.
- Tipográfico (errores tipográficos y ortográficos)
 - Cambios de letras y números: “o4 Feb 19o3”
 - Espacio al principio y/o al final del campo. Errores de ortografía: “ 14 Avril 1981”
- Contexto (Dato que no corresponde a la definición del campo)
 - Información que no corresponde al campo: “2050 m.” , “M.B.G 830 – 12-08-1987”

CAPTURA ELECTRÓNICA DE DATOS

INFORMATIZACIÓN

Pautas o modelos de error (CONABIO): un ejemplo con fechas

- Redundancia
Mismo datos capturado más de una vez: “1983-8 Mar 1983”, “29-29 Feb 1975”
- Convención (Datos capturados sin utilizar convenciones establecidas ni estándares)
“17 ? 1963” “17 00 1963” “s.d.[1931-1932]”
- Uniformidad:
La misma descripción escrita de forma diferente: “Verano 2001” “Spring 96” “Mayo-Agosto 1989”
La separación entre números se realiza por distintos signos:
– “10-7-1992” “12/10/1993” “10.5.1981”
- Congruencia: fechas inexistentes de colecta, años imposibles, etc.
“31 Abril 1997” “21/15/2030” “21/11/1050”



ALMACENAMIENTO DE DATOS

ALMACENAMIENTO DE DATOS

COPIAS DE SEGURIDAD

- 1 Tener **dos ordenadores** distintos para realizar las copias de seguridad.
- 2 Realizar **alternativamente** copias en cada ordenador.
- 3 Una de las máquinas conviene que esté en un lugar **separado físicamente** del lugar de trabajo.
- 4 Establecer la **frecuencia** de copias (diario, semanal...)
- 5 Usar “la nube”
- 6 Realizar copias **acumulativas** que **no se reemplacen** unas a otras, para tener distintas fases del 6 crecimiento de la BBDD y no arrastrar errores.
- 7 Hacer **periódicamente** copias en DVD o discos externos, y documentar sus metadatos: contenido, fecha, versión del programa/s.
- 8 En bases de datos de MS-Access, compactar y reparar la base de datos antes de cada copia, para reducir el tamaño.
- 9 Chequear las copias de seguridad y **comprobar** su correcto funcionamiento.



Francisco Pando

Unidad de coordinación, GBIF España
Real Jardín Botánico - CSIC
Claudio Moyano 1, 28014 Madrid, Spain
pando@gbif.es

www.gbif.es , twitter.com/GbifEs , datos.gbif.es



<http://creativecommons.org/licenses/by-sa/3.0/es/>

GBIF-ES es el Nodo Nacional de Información sobre Biodiversidad patrocinado por el [Ministerio Español de Economía y Competitividad](#), gestionado por el [Consejo Superior de Investigaciones Científicas](#).



*Argema
mittrei*