CALIDAD EN TODO EL PROCESO

Isabel Ortega Maqueda
Unidad de Coordinación GBIF-ES
ortega@gbif.es

Taller de calidad de datos en Bases de datos de Biodiversidad
Real Jardín Botánico de Madrid (España)
25-26 Noviembre 2008

Calidad en todo el proceso

Fuentes disponibles en la Web de Gbif:

- *Principles of Data Quality (http://www.gbif.org/prog/digit/data_quality/DataQuality.pdf)*
- *BioGeomancer Guide to Georeferencing: (http://www.gbif.org/prog/digit/data_quality/BioGeomancerGuide.pdf)*
- *Principles and Methods of Data Cleaning (http://www.gbif.org/prog/digit/data_quality/DataCleaning.pdf)*
- *Uses of Primary Species-Occurrence Data (http:/www.gbif.org/prog/digit/data_quality/UsesPrimaryData.pdf)*

Captura de datos en el campo

La captura de los datos puede ser realizada de diversas maneras, y de ellas dependerá también la calidad de los datos:

- **De forma oportunista.**
  Especímenes de colecciones como datos observacionales son capturados **de manera no sistemática**, lo que genera **sesgos espaciales** (correlación con carreteras, ríos, etc.)

- **Estudios de campo**
  Generalmente incluyen coordenadas geográficas o UTM.

- **Observaciones en áreas de gran escala.**
  La precisión de sus coordenadas suele ser baja debido a la gran extensión de la zona de estudio (Ejemplo: Estudio dentro de un parque nacional)

Usando el GPS

- La exactitud de un GPS suele tener un rango de error **menor de 10 metros.**
- La exactitud puede mejorarse, si se realiza **la media de los resultados** de múltiples capturas o colectas en un mismo punto.
- El uso de **GPS Diferencial** proporciona una exactitud de **1 a 2 metros** (Sistema que proporciona, a los receptores de GPS, correcciones a los datos recibidos de los satélites GPS, a partir de un receptor GPS de referencia **fijo en tierra**)
- Los **GPS Diferenciales en tiempo real**: Tienen una alta precisión y dan una exactitud en un rango de **1 a 2 centímetro**. Son muy caros y pocas veces se necesita tanta precisión en los registros de las colecciones biológicas.

Captura de coordenadas con GPS

Los requisitos para una buena toma de coordenadas con el GPS son:

- Se recomienda la recepción de **al menos 7 satélites** (son necesarios 4 satélites como mínimo para determinar la localización de un punto sobre la tierra)
- El GPS receptor debe estar en una **zona despejada de obstáculos** sobre nuestras cabezas y **lejos de superficies reflectoras**
- Tener una **visión despejada sobre el horizonte** (estar bajo una fuerte cubierta forestal **no** ayudaría a la toma de coordenadas)
- El GPS debe configurarse para usar el **Datum** apropiado para el área. El 30 de agosto de 2007 entró en vigor el **REAL DECRETO 1071/2007 de 27 de julio**, por el que se regula el **sistema geodésico** de referencia **oficial en España**:
  - **ETRS89** en el ámbito de la **Península Ibérica y las Islas Baleares**
  - **REGCAN95** en el caso de las **islas Canarias**.

Captura electrónica de datos

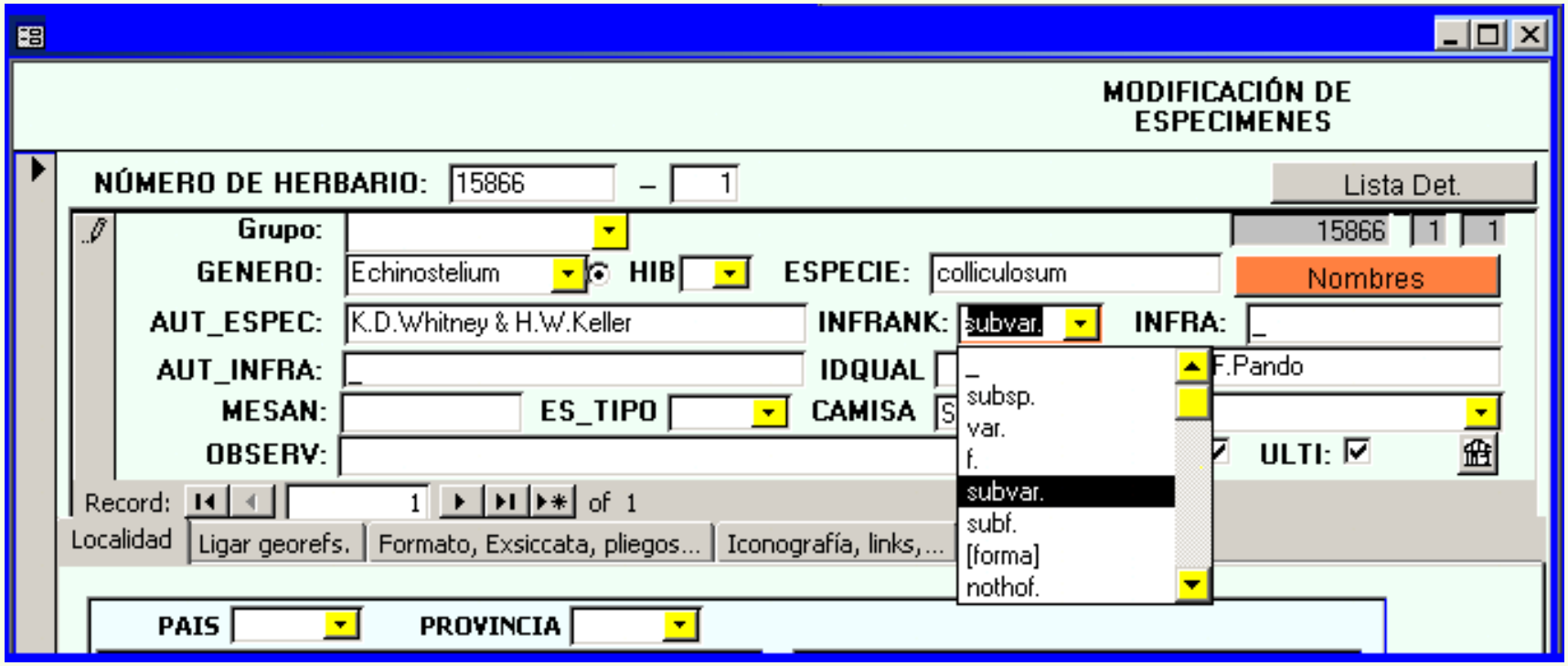
Captura básica de datos

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
- Uso de software del que se haya tenido una formación previa
- Supervisión de expertos que lleven a cabo un testeo.
- Desarrollo 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.

Diseñando la interface de usuario

Una buena interface debe facilitar la tarea de la introducción de datos:



Separación de tareas...

Es a menudo más rápido y eficiente realizar la georreferenciación como una tarea separada de la actividad de digitalización de la información de la etiqueta. De esta manera, la propia base de datos nos puede facilitar el proceso de georreferenciación:

- Ordenando por colector, localidad y fecha de recolección, etc.
- Permitiendo un uso más eficiente de los mapas o programas GIS utilizados para la obtención de las coordenadas.
- Ahorra la duplicación de esfuerzos, a la hora de georreferenciar múltiples registros de la misma localidad.

Datos espaciales

Herramientas geográficas

http://www.gbif.es/HerramGeo.php

Guías para una buena georreferenciación

Georeferencing Guidelines

http://manisnet.org/manis/GeorefGuide.html

MaPSTeDI

Georeferencing in MaPSTeDI

http://mapstedi.colorado.edu/georeferencing-howto.html

Almacenamiento

La forma de almacenar y conservar los datos puede tener un efecto en la calidad de los datos, y tiene que ver tanto con el diseño de la base de datos como con el resto de pasos dentro de la **cadena de obtención de la calidad**.

- **Diseño** de la base de datos
- **Backups** – la realización regular de **copias de seguridad** evita la pérdida de datos y garantiza unos niveles de calidad.
- **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 y evitando la dispersión, el difícil acceso o el olvido de muchas bases de datos en Universidades, ONG's, etc.

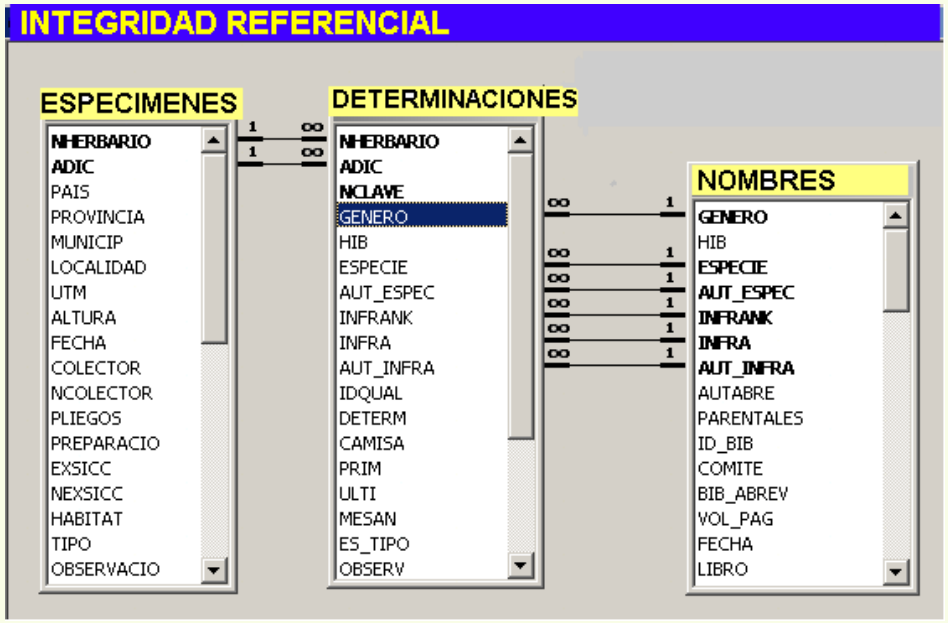
Copias de seguridad

Establecer un sistema de backups:

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. Realizar copias acumulativas que no se reemplacen unas a otras, para tener distintas fases del crecimiento de la BBDD y no arrastrar errores.
6. Hacer periódicamente copias en DVD o discos externos, y documentar sus metadatos: contenido, fecha, versión del programa/s.
7. En bases de datos de MS-Access, compactar y reparar la base de datos antes de cada copia, para reducir el tamaño.
8. Chequear las copias de seguridad y comprobar su correcto funcionamiento.

Integridad de datos

La integridad de los datos se refiere a la condición en la cual los datos **no han sido alterados ni destruidos sin autorización**, ni han sido **maliciosamente modificados o destruidos** (por ejemplo, por un virus).



En el mantenimiento de la integridad de los datos influyen: una buena **gestión de los datos**, un buen **diseño de la base de datos**, los **backups** y el **archivo correcto.**

Pautas o modelos de error

La Conabio (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad de Méjico) clasifica los tipos errores en las bases de datos biológicas según estos criterios:

1. **Omisión**
   - Información ausente
   - Información incompleta
2. **Tipográfico**
   - Error de tipografía
   - Error de ortografía
3. **Contexto**
   - Dato que no corresponde a la definición del campo
4. **Redundancia**
   - Redundancia del dato en un celda
5. **Convención**
   - Datos capturados sin utilizar convenciones establecidas ni estándares
6. **Uniformidad**
   - Registros con datos sin uniformidad
7. **Congruencia**
   - Más de un dato del mismo tipo capturado en una celda
   - Fechas imposibles
   - Números ilógicos. Colectores cuyo intervalo de colecta es mayor de 70 años.

Diferentes tipos de errores con fechas

Ejemplos con la **Fecha de Recolección o Captura**:

Omisión

- **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:

- **Cambios** de letras y números: "o4 Feb 19o3"
- **Espacio** al principio y/o al final del campo. Errores de **ortografía**: " 14 Avril 1981"

Contexto:

- Información que **no corresponde** al campo: "2050 m." , "M.B.G 830 – 12-08-1987"

Redundancia

- Mismo datos capturado **más de una vez**: "1983-8 Mar 1983", "29-29 Feb 1975"

Diferentes tipos de errores con la fecha

Convención:

- Datos capturados **sin estándares** ni convecciones establecidas:

  “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”

Integración de datos

La integración de datos provenientes de diferentes bases de datos puede generar inconsistencias si…

- Se utilizan diferentes técnicas de medición.
- Diferencias de resolución (medidas de distancias-->Km., millas, medidas de tiempo).
- Diferentes interpretaciones de la terminología y la nomenclatura (uso de diferentes taxonomías)
- Diferencias en la configuración del GPS (datum, sistemas de coordenadas (decimales / UTM)

La integración de datos conlleva una mayor calidad si en la grabación de los datos se han usado estándares.