Generalización de datos sensibles

Isabel Ortega Maqueda
ortega@gbif.es

I TALLER SOBRE GESTIÓN DE DATOS SENSIBLES
Madrid, 15-16 de Octubre 2008

Imagen de portada: Alberto González Talaván

Generalización de datos sensibles

1) Guía de buenas prácticas
2) Datos técnicos
3) Métodos de generalización
4) Generalización sobre la localidad
5) Generalización de información textual
6) Generalización sobre la georreferenciación
7) Extensión geoespacial de Darwincore1.4
8) Documentación y metadatos

Guía de buenas prácticas

- La segunda etapa en el proceso ha sido el desarrollo de un informe cuya finalidad es la formalización de recomendaciones sobre una buena práctica en esta materia:

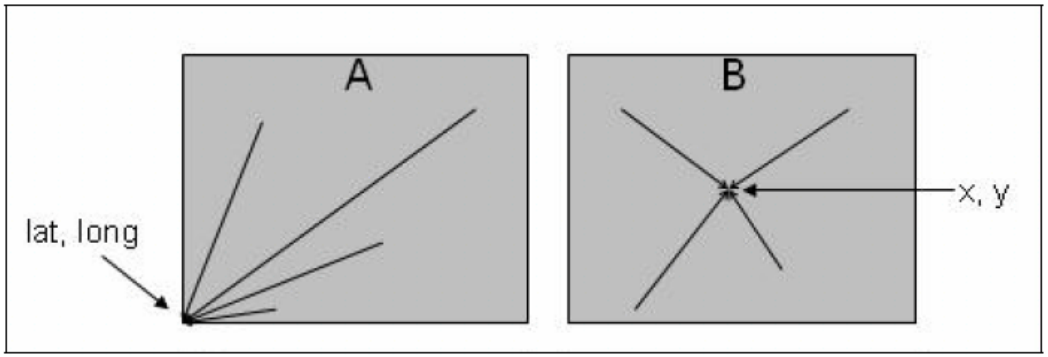
  Guide to best practices for generalising primary species occurrence data

- Este documento deber una guía primordial para instituciones, proveedores de datos y Nodos de GBIF para desarrollar sus propias guias de buenas prácticas, adaptadas a cada institución y organización y que incorporen también la guía de buenas prácticas para georreferenciar de Chapman and Wieczorek (2006), accesible en el enlace:

  Guide to Best Practices in Georeferencing

Datos técnicos

- Generalización: se refiere a cualquier modificación llevada a cabo en los datos originales para ocultar o difuminar información sensible, normalmente por reducción de la precisión de los datos.
- En términos geográficos se refiere a la conversión a una representación geográfica con menos resolución y menos contenido de información, asociado con un cambio en la escala.
- Randomizar: se refiere a deliberadas asignaciones de valores al azar, para desorientar sobre la verdadera localización. Este proceso tiende a una falsificación de los datos.



A -> Rejilla geográfica: los registros son referenciados a la esquina inferior izquierda

B -> Rejilla métrica: los registros son referenciados al centro de la cuadrícula.

.

Métodos de generalización

- Dependiendo del método/s que se utilicen, los datos técnicos serán sencillos o complejos.
  - La generalización utilizando una cuadrícula geográfica de 10 minutos, 1 minuto, 30 segundos es muy sencilla de implementar.
  - Las cuadrícula de 100m, 500m, 1 Km., 10 Km., etc. son algo más complicadas
  - Generalizar a una región biogeográfica o política es sencillo de implementar desde un campo de texto.

A tener en cuenta...

- Si se suministran coordenadas en los casos en que se generalice a una región biogeográfica, los resultados pueden ser equívocos - por ejemplo, una especie costera puede quedar con unas coordenadas que la sitúan cientos de kilómetros hacia el interior - reduciendo su uso para análisis y chequeo de datos.
- Poner accesibles los datos sin una adecuada documentación puede llevar a resultados desastrosos para los usuarios. En estos casos es mejor no suministrar coordenadas geográficas.
- Generalizar la información georreferenciada va a depender de tener los registros con coordenadas geográficas, y actualmente se estima que sólo el 1% de 2.5 billones de registros de colecciones sobre biodiversidad contienen coordenadas geográficas.

Recomendaciones

- La Generalización es preferible a la randomización a la hora de proteger las localidades exactas de los taxones sensibles y sus atributos cuando estos datos son compartidos, y cuando la información sobre estas localidades puede conducir a daños medioambientales.

- Tres son los niveles de generalización recomendados:
  - 0.1 degrees (~10-12 km),
  - 0.01 degrees (~1-1.2 km),
  - 0.001 degrees (~100-120 m).

Generalización frente randomización

- La generalización crea/mantiene datos “verdaderos” , mientras que la randomización crea deliberadamente datos falsos. La generalización puede ser implementada (o no) caso por caso, dependiendo del uso intencionado que se quiera dar a los datos.
- La generalización es fácilmente implementable quitando precisión a los datos y suministrando datos redondeados a cuadrículas.
- La generalización es sencilla y suministra información que sigue siendo útil en escalas medias, sin perder la localización exacta de las poblaciones.
- La generalización conducirá a mejorar la credibilidad de los estudios basados en datos suministrados a través de GBIF.

Qué generalizan los proveedores de datos

Campos generalizados por los proveedores encuestados:

CAMPO	ENCUESTADOS	COMENTARIOS
Localidad	42	Eliminada o alterada
Georreferenciación	42	Eliminadas o disminuida la precisión
Exactitud	1	
Nombres de Colector/Observador	16	Restringido por numerosas razones: privacidad de las personas, trazado de itinerarios, etc. Otros nunca suprimen esta información
Fechas	8	Pueden ser usadas para rastrear recolecciones antes y después del taxon sensible
Información taxonómica	4	
Hábitat	1	

El control del proceso

Las respuestas fueron de forma abrumadora de la opinión de que los custodios de los datos deben mantener el control sobre qué datos pueden ser generalizados y qué información debe ser ocultada antes de su puesta en Internet.

La responsabilidad de la información sobre exactitud y fiabilidad de los datos, y las restricciones de acceso a los datos sensibles, reside en el proveedor de datos (GBIF 2004a).

Generalización de la localidad

- La mayoría de las instituciones que generaliza la descripción de la localidad lo hace de varias formas:
  - Hacer completamente inaccesible este dato en(60%)
  - Referenciar a un área geográfica mayor (20%)
  - Cambiar la localidad por una explicación del hecho (20%):

    “Este especimen representa a una especie en peligro de extinción o amenazada. La localidad específica ha sido borrada del registro on-line para proteger a esta especie de sobre-recolección. Estos datos pueden ser suministrados a los investigadores que lo soliciten.”

Restricción sobre determinadores

Hay una tendencia en contra de la ocultación de los nombres de los determinadores, no así en ocultación de los colectores (por razones de privacidad). Las recomendaciones son:

1. Cuando sea posible, el nombre y la fecha de la determinación serán citadas.
2. Será citada la precisión con la que se realizó la determinación:
   - Holotipo o forma parte de los tipos de la colección.
   - Ha sido comparado con algún holotipo, isotipo, etc.
   - Ha sido utilizada una taxonomía determinada, especificandola.
3. Nivel de habilidad y certeza en la determinación:
   - Si la identificación fue llevada por un experto mundial, un experto regional, por alguién no experto, por un colector.
   - Si la identificación puede clasificarse de “alta certeza”, de “razonable certeza”, o “con dudas”.
4. Las razones por las que no ha podido ser determinada con “alta certeza”: especimen dañado, mal conservado, estéril, etc.

Generalización del colector

- El nombre del colector será restringido si así lo dictan las leyes de privacidad de cada legislación, pero no en ningún otro caso.
- Habrá que tomar medidas para evitar las posibilidades de análisis co-relacionales y deducir así las coordenadas de un registro sensible basándose en los siguientes datos:
  - Colector
  - Número de colector
  - Fecha de recolección
  - Hábitat
- Las medidas a tomar no restringirán la información en los registros relacionados, sino que será eliminada la información en el propio registro sensible, lo que evitará pérdidas de calidad de datos, mejorará la eficiencia de los procesos de validación y data Cleaning.
- La información de registros relacionados con un registro sensible (aunque sí en el propio registro sensible) no debería ser restringida a menos que sea absolutamente necesario.

Generalización de las coordenadas

CATEGORÍA	SENSIBILIDAD	GEORREFERENCIACIÓN
Categoría 1	Extrema	Georreferenciación no proporcionada, se eliminan las coordenadas del registro publicado en la Web.
Categoría 2	Alta	Georreferenciación redondeada a 0.1 grado
Categoría 3	Media	Georreferenciación redondeada a 0.01 grado
Categoría 4	Baja	Georreferenciación redondeada a 0.001 grado
No sensible	No sensible	Georreferenciación sin restricciones

En cada registro cuyas coordenadas se generalicen, debe especificarse su rango de incertidumbre, que se corresponde con el concepto de **RadioPunto**.

Extensión geoespacial de Darwincore1.4

La nueva extensión geoespacial propuesta por el TDWG (Biodiversity Information Standards) se puede consulta en:

Geospatial Extension Concept List

- DecimalLatitude
- DecimalLongitude
- GeodeticDatum
- CoordinateUncertaintyInMeters
- PointRadiusSpatialFit
- VerbatimCoordinates
- VerbatimLatitude
- VerbatimLongitude
- VerbatimCoordinateSystem
- GeoreferenceProtocol
- GeoreferenceSources
- GeoreferenceVerificationStatus
- GeoreferenceRemarks
- FootprintWKT
- FootprintSpatialFit

RadioPunto y Ajuste espacial

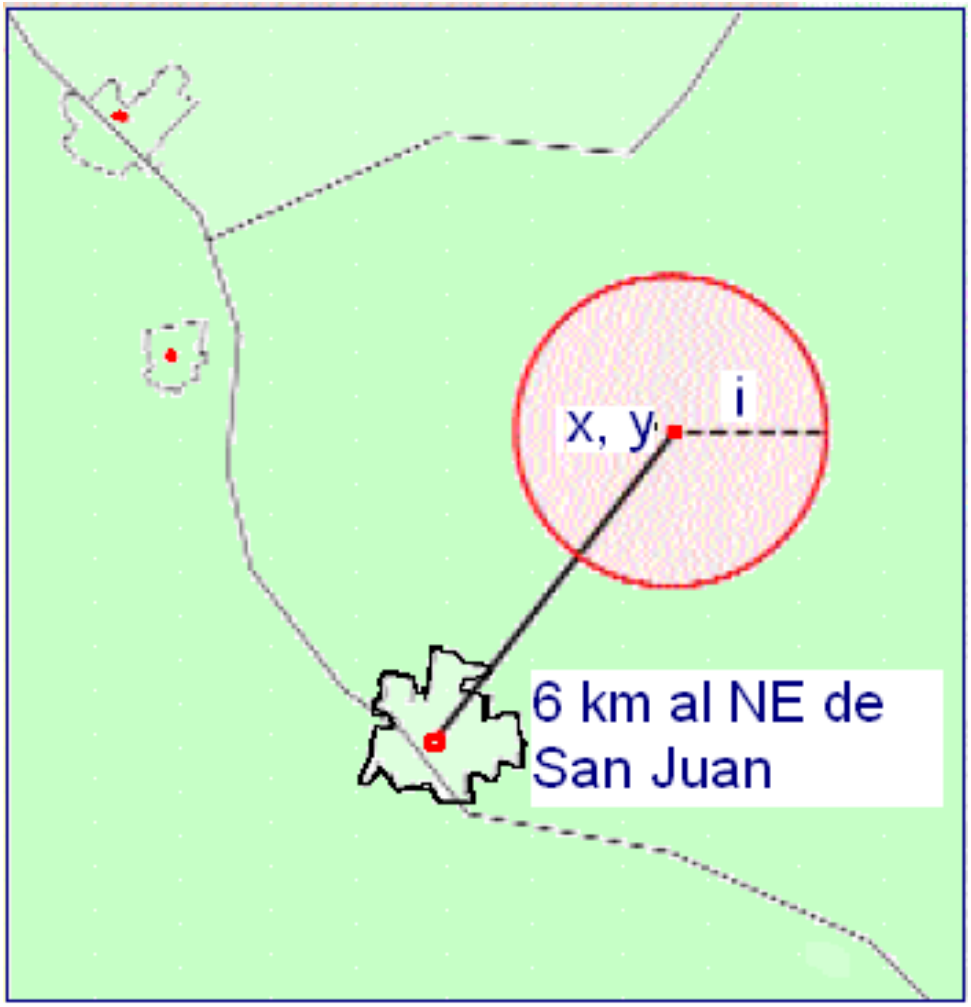
- **CoordinateUncertaintyInMeters = RadioPunto.**

  La distancia mayor expresada en metros, medida desde una latitud y longitud dadas, describiendo un círculo dentro del cual se encuentra la localidad descrita.

- **PointRadiusSpatialFit = Ajuste espacial.**

  Expresa lo bien que encaja el círculo definido por las coordeadas y su radio de incertidumbre respecto de la representación espacial original

El Radiopunto



A partir de la descripción de la localidad obtenemos un par de coordenadas **(x,y)** asociadas a una medida de longitud que será su **incertidumbre (i):**

Esta distancia define **el radio del área más probable** en donde se encontraría el sitio de recolección.

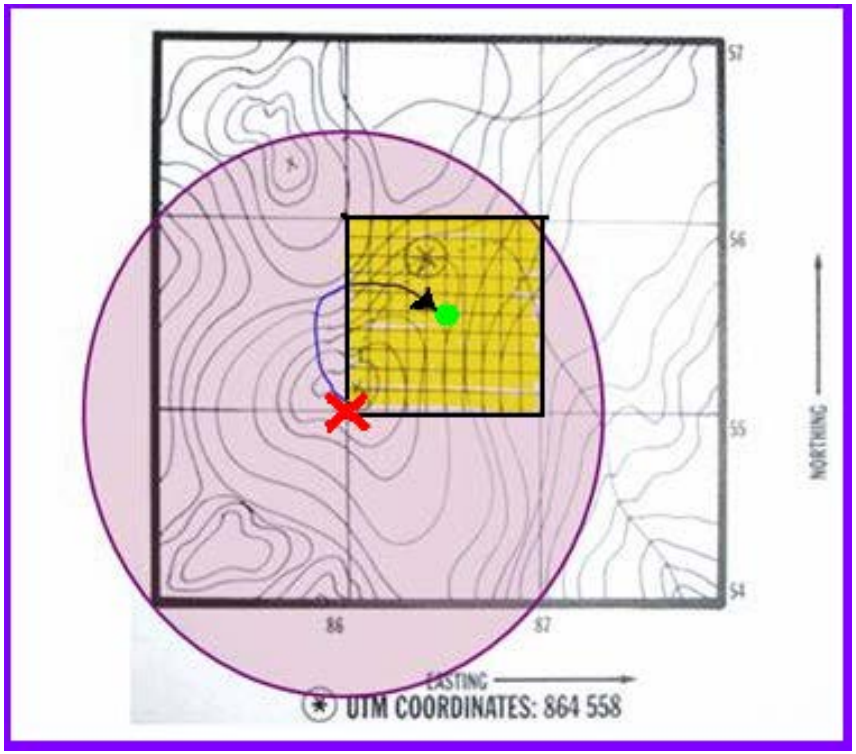
Coordenadas y Datum

Precision	0 degrees Latitude	30 degrees Latitude	60 degrees Latitude	85 degrees Latitude
1.0 degree	156,904 m	146,962 m	124,605 m	112,109 m
0.1 degree	15,691 m	14,697 m	12,461 m	11,211 m
0.01 degree	1,570 m	1,470 m	1,246 m	1,121 m
0.001 degree	157 m	147 m	125 m	112 m
0.0001 degree	16 m	15 m	13 m	12 m
0.00001 degree	2 m	2 m	2 m	2 m
1.0 minute	2,615 m	2,450 m	2,077 m	1,869 m
0.1 minute	262 m	245 m	208 m	187 m
0.01 minute	27 m	25 m	21 m	19 m
0.001 minute	3 m	3 m	3 m	2 m
1.0 second	44 m	41 m	35 m	32 m
0.1 second	5 m	5 m	4 m	4 m
0.01 second	1 m	1 m	1 m	1 m

Table 4. Table showing metric uncertainty due to precision of coordinates based on the WGS84 datum at varying latitudes. Uncertainty values have been round up in all cases. From Wieczorek (2001).

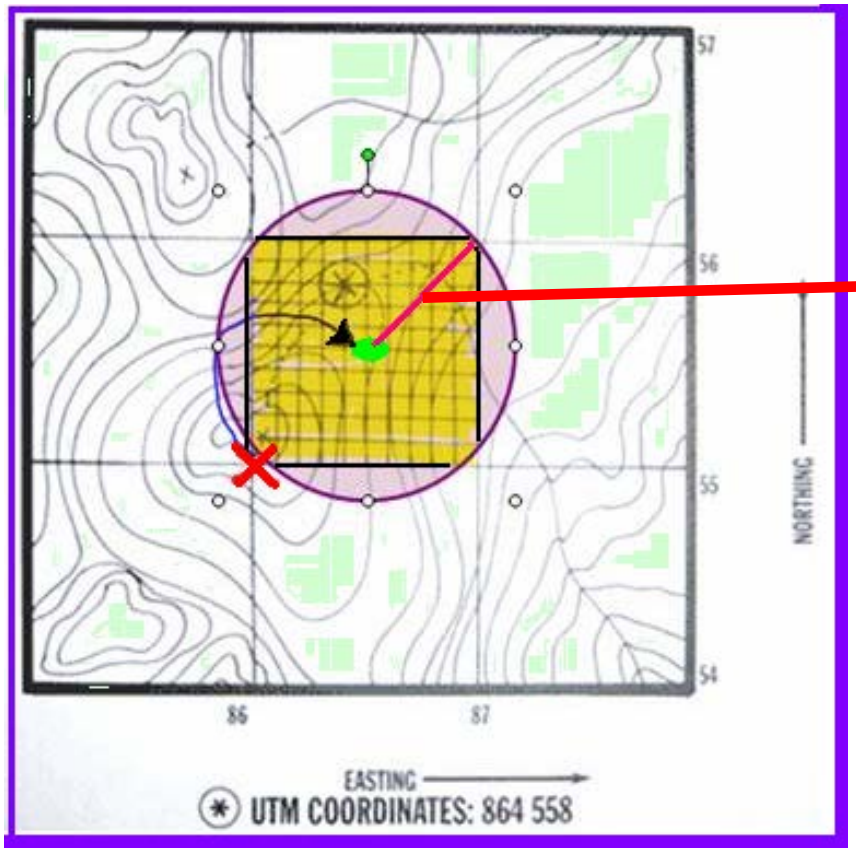
Recentrando la UTM

Para calcular el **radio** de incertidumbre, **recentramos** la coordenada UTM: 30TUN63 -> 30TUN6**5**3**5**



Calculando la incertidumbre

coeficiente de escala = lado del cuadrado de la UTM



Incertidumbre =

(coeficiente de escala * $\sqrt{2}) / 2$

Teorema de Pitágoras:

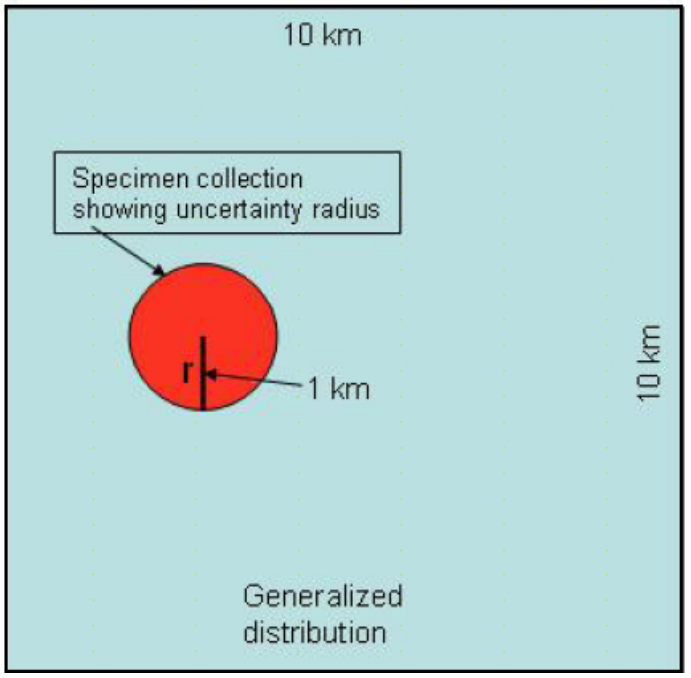
Ajuste espacial

El ajuste espacial (Spatial Fit) es un nuevo concepto de georreferenciación que proporciona una medida de cómo una representación geométrica modificada encaja con la representación geométrica original

Puede tener los valores

- **0**: la nueva representación no comprende la totalidad de la original.
- **1**: coincide al 100%
- **> 1**: la nueva engloba a la totalidad de la original, y el valor es la proporción del área nueva respecto de la original.
- **Indefinido**: la representación original es un punto y la nueva representación es un área.

Generalization Fit = 31.8 – i.e. (10^2/(PI*r^2))



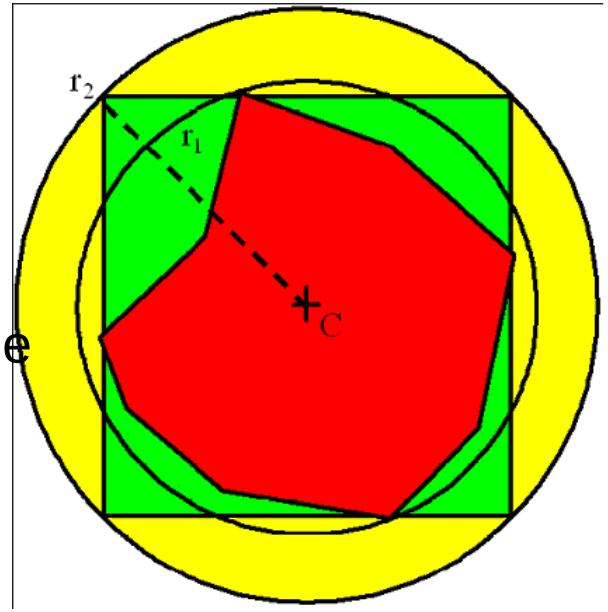
Determinando el ajuste espacial

El área original de la localidad es el polígono rojo con un área **A**:

- El ajuste del círculo amarillo: $\pi * r2^2/A$
- El ajuste del cuadrado verde: $2 * r2^2/A$
- El ajuste del polígono rojo: **1**

El área original de la localidad es el cuadrado verde con un área de $2 * r2^2$

- El ajuste del círculo amarillo: $\pi * r2^2/ 2 * r2^2$
- El ajuste del cuadrado verde: **1**
- El ajuste del polígono rojo: **0**



Documentación de la generalización de los datos

- La mayoría de usuarios de los datos entienden la necesidad de los proveedores de restringir cierta información sobre taxones sensibles.
- A su vez, remarcan la necesidad de una buena documentación para conocer:
  - qué taxones tienen restringida su información
  - cómo ha sido restringida o generalizada la información
- Esta documentación les permitirá tomar decisiones sobre el valor y/o la utilidad de los datos para su particular uso y análisis.
- Actualmente hay datos que están siendo generalizados, pero sin documentación asociada que informe a los usuarios. Los datos así presentados pueden ser usados de manera inapropiada y producir falsos resultados.

Documentación y metadatos

Metadatos: datos que documentan a los datos.

Los metadatos cumplen una función esencial atendiendo a la comunicación con terceros, restricciones de acceso y condiciones de uso, que los proveedores de datos dan a sus datos.

Pueden ser considerados como una ayuda en la protección de datos, ya que permitirán a los usuarios visualizar las condiciones establecidas por el proveedor de datos para el acceso y uso de la información (Llinás, 2005).

Documentación (cont.)

- La recomendación es que allí donde el datos es restringido, esta circunstancia sea documentada con las adecuadas expresiones:
  - "*nombre suprimido por razones de privacidad*";
  - "Esta especie representa a una especie amenazada o en peligro *. La localidad específica ha sido eliminada del registro on-line para proteger a este taxon. Esta información puede ser suministrada a investigadores bajo petición"*
  - Esta especie representa a una especie amenazada o en peligro. La localidad específica ha sido generalizada a una presencia dentro de una cuadrícula de 0.1 grado de resolución.
- Los campos así tratados no deberán dejarse en blanco.

Niveles de documentación

Metadatos a nivel de dataset (colección): documentación referida a todo el conjunto de datos.

Metadados a nivel de registro: se refiere a la documentación a nivel de registro que no puede especificarse con detalle en los metadatos de la colección. Es recomendable la documentación del status de sensibilidad del registro (de la especie o de algún atributo) además de las restricciones de acceso a ciertos datos y los detalles de la generalización de los datos.

El TDWG está considerando añadir cinco nuevos campos a los elementos de la Extensión Geoespacial de DarwinCore:

- *DataSensitiveIndicator;*
- *DataSensitiveReason;*
- *DataSensitiveComment;*
- SensitiveDateForReview
- *PrecisionDataProvided;*
- *PrecisionDataStored*

Documentando datos sensibles

- DataSensitiveIndicator
  - *SI / NO que indique si el registro es sensible.*
- DataSensitiveReason
  - *Razones por las que el registro es sensible.*
- DataSensitiveComment
  - *Otra información de las razones de su inclusión como sensible.*
- SensitiveDateForReview
  - *Fecha de revisión del estado de sensibilidad.*
- PrecisionDataProvided
  - *Escala o precisión de los datos publicados:*
    - *0 = 1 grado*
    - *1 = 0.1 grado*
    - *2 = 0.01 grado, etc*
- PrecisionDataStored
  - *Escala o precisión de los datos almacenados en origen.*

Concluyendo con la documentación

- El conocimiento de los metadatos es esencial por muchas razones, y donde los datos han sido restringidos o generalizados es importante que esta información sea registrada a nivel de registro.
- Para los metadatos asociados con la generalización de la información espacial, sería recomendable que la Extensión Geoespacial de Darwin Core pueda ser modificada.
- GBIF promoverá el uso de metadatos para describir adecuadamente los recursos de datos y en particular, cualquier restricción que se haya realizado sobre la disponibilidad de los datos.
- Se ha sugerido la idea de Icono para hacer los metadatos más accesibles desde el portal de GBIF, que informaran de los metadatos a nivel de registro y enlazaran con los metadatos de la colección.

Preguntas, dudas, sugerencias?