Transferencia de capacidades técnicas asociadas a la publicación de datos de biodiversidad a través de GBIF

CESP: Regional Capacity Enhancement to Latinamerica by establishing Chile's node

23 al 25 abril 2018

Santiago - Chile

Leonardo Buitrago
Administrador Nodo GBIF Colombia,
Representante (S) Regional de Nodos GBIF (Latinoamérica y Caribe)
Líder Administración de contenidos SiB Colombia
albuitrago@humboldt.org.co

Conceptos básicos sobre calidad y curación de datos

Ministerio del Medio Ambiente
Gobierno de Chile

Gbif CHILE
Nodo Nacional de Información sobre Biodiversidad

Gbif.es

SiB Colombia

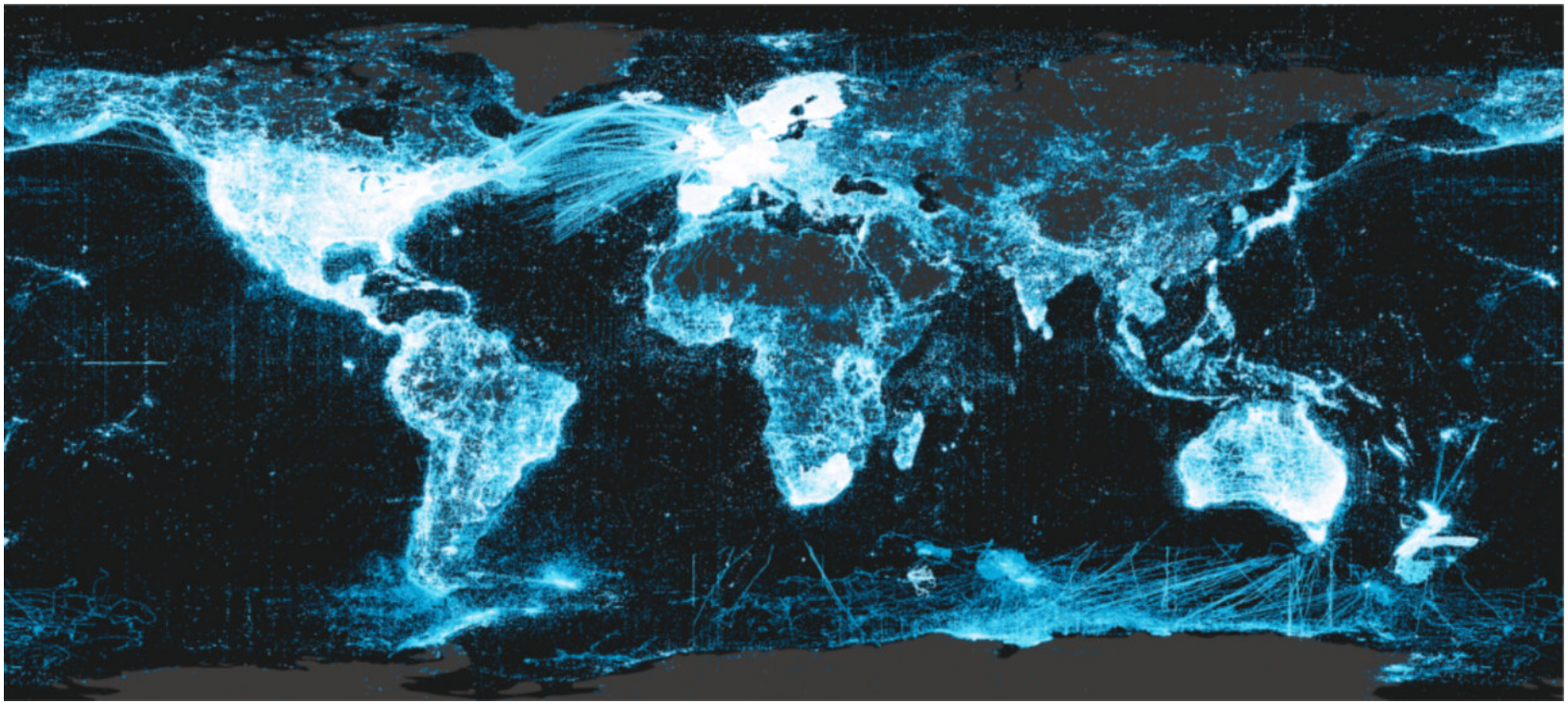
SNDB
Sistema Nacional de Datos Biológicos

Ministerio del Ambiente

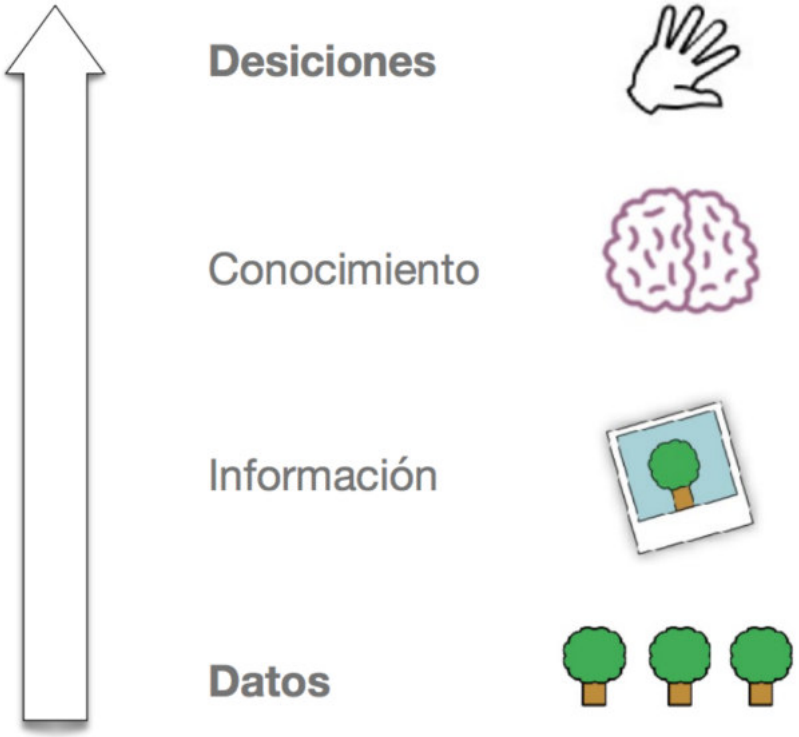
GBIF

#GBIF1billion

Resources for the GBIF network to use in association with the upcoming big data milestone



DE LOS DATOS A LA TOMA DE DECISIONES



PERO...EXISTEN VACÍOS DE INFORMACIÓN

¿Cuántas y cuáles especies existen?

¿Cuál es el tamaño de sus poblaciones y dinámicas?

¿Cuál es su distribución temporal y espacial?

¿Cuántas están siendo afectadas por condiciones bióticas y abióticas?

NECESITAMOS MÁS DATOS !

- Regiones pocos estudiadas o representadas
- Trabajo de campo y laboratorio
- Apoyo y financiación científica

NECESITAMOS USAR MEJOR LOS DATOS EXISTENTES !

- Una gran cantidad de datos no están disponible para su uso
  - No digitalizados, no compartidos
  - No fácilmente accesibles
  - **Problemas de calidad**

http://goo.gl/pnjg82

REFERENTES TEÓRICOS

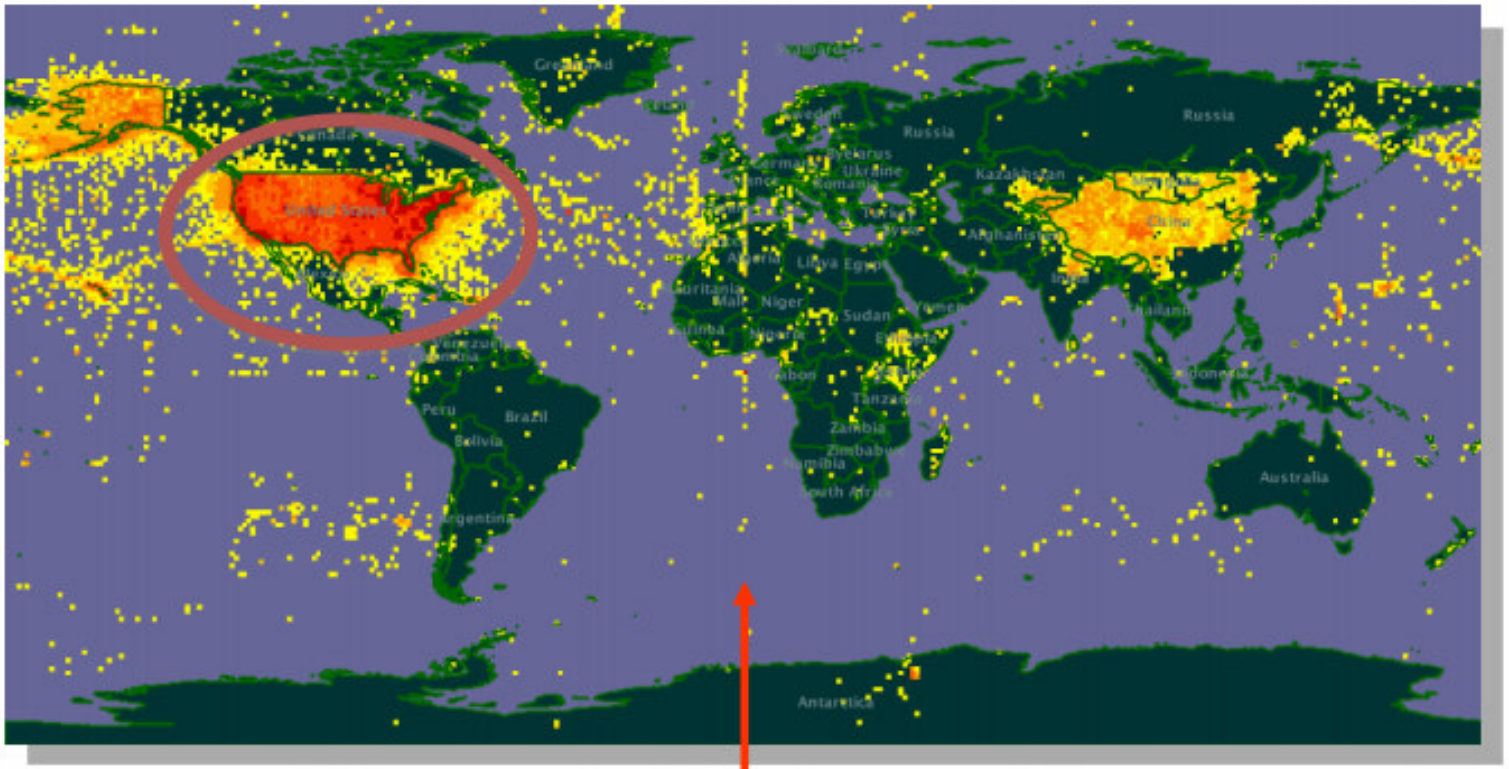
Saraiva A. 2014

Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

BASURA ENTRA → BASURA SALE

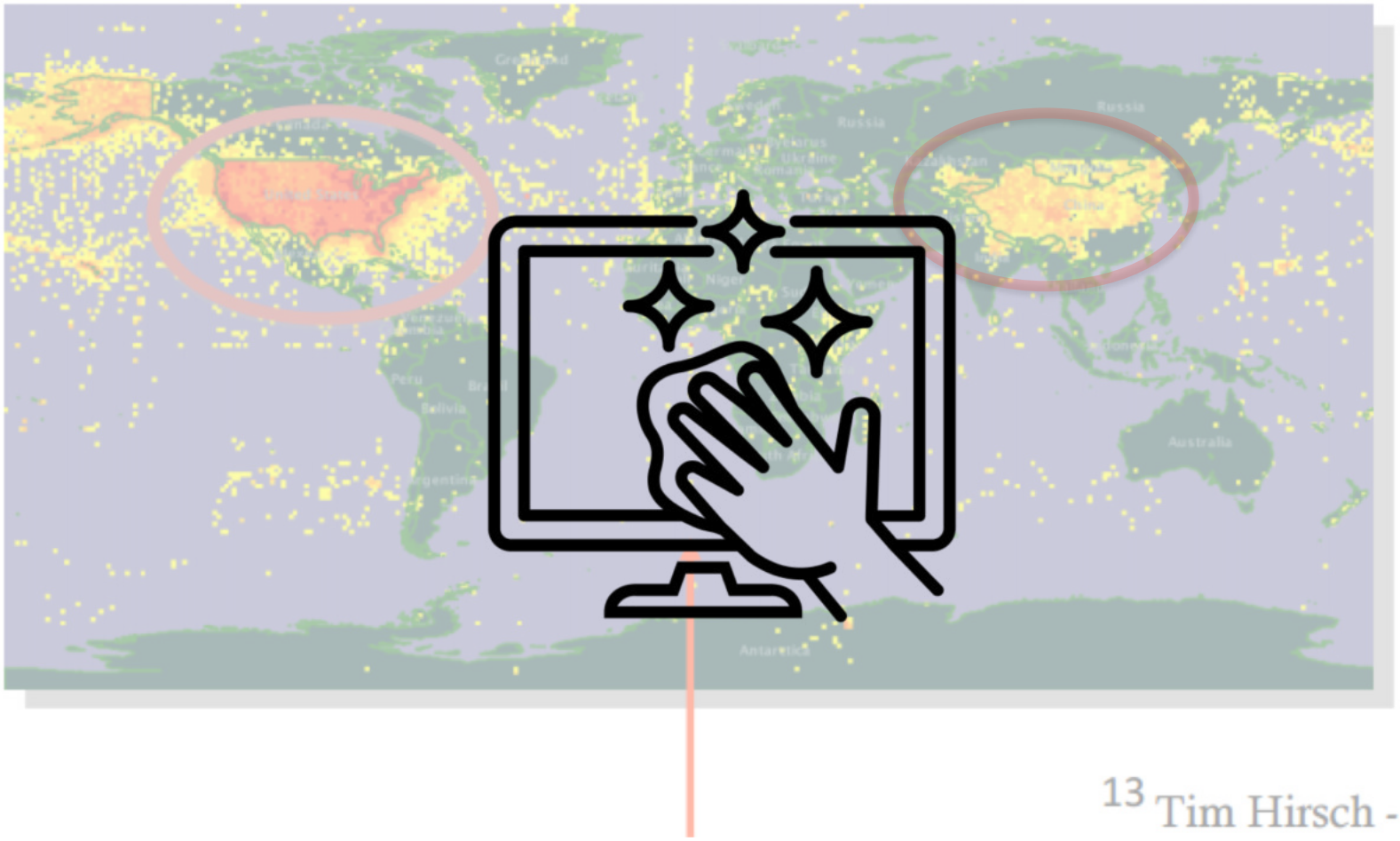
- **Problemas de calidad**: conllevan a resultados de mala calidad: análisis, decisiones, etc.
- **Los problemas surgen de**: toma de datos, digitalización, falta de metadatos, ausencia de estándares.
- **Hay mucho por hacer**: limpieza de datos (corrección), prevención y políticas de calidad de datos

EJEMPLO



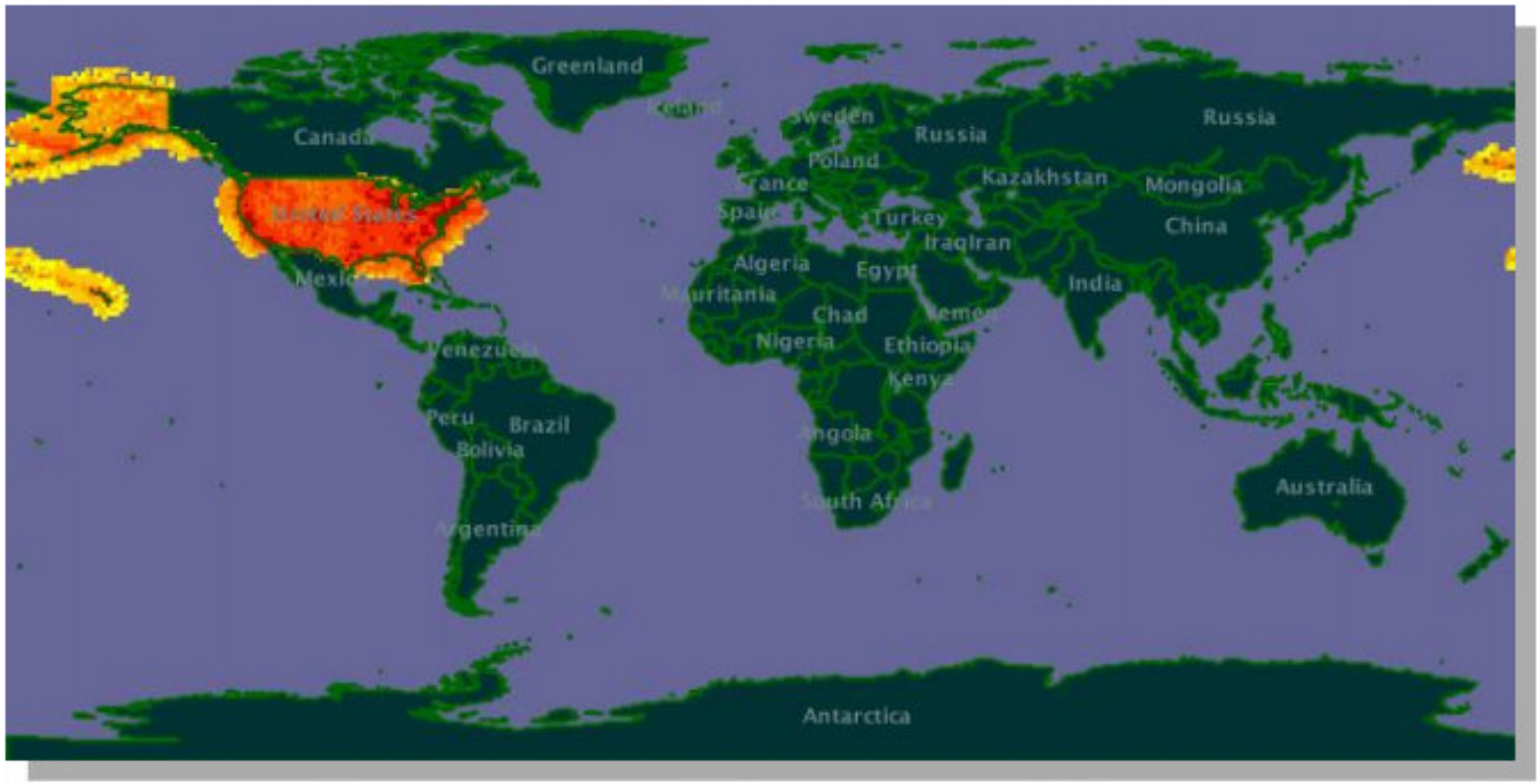
13 Tim Hirsch - GBIF

EJEMPLO

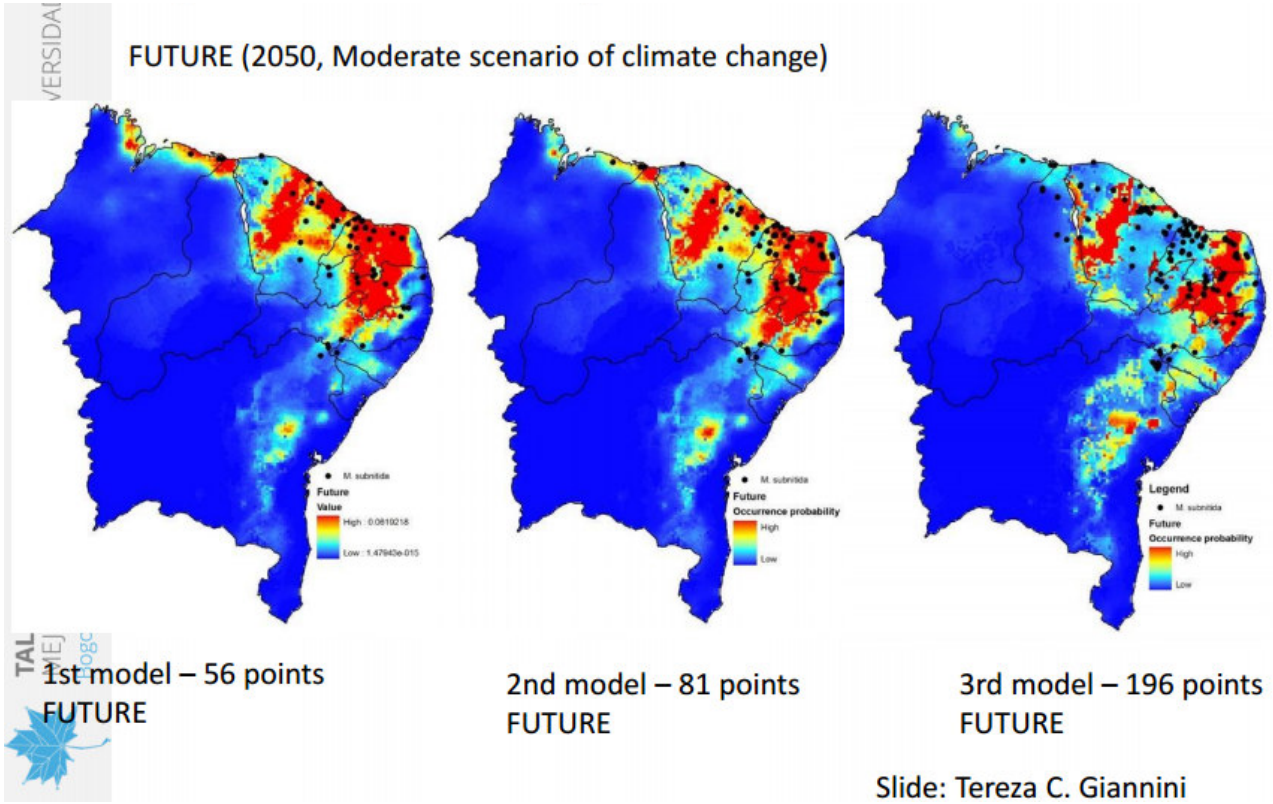


13 Tim Hirsch - GBIF

- Wrong longitude (missing minus signal)



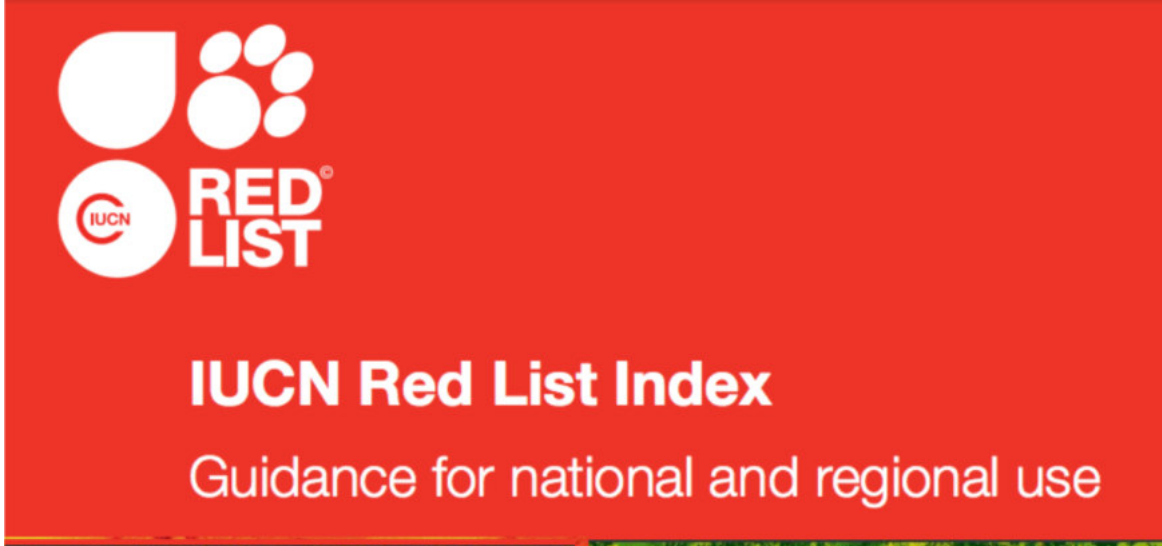
IMPACTO DE LOS DATOS EN LA GENERACIÓN DE MODELOS



Slide: Tereza C. Giannini

CALIDAD DE DATOS

La calidad de datos puede afectar los indicadores, análisis, toma de decisiones y políticas



...y ahora quién podrá ayudarnos?

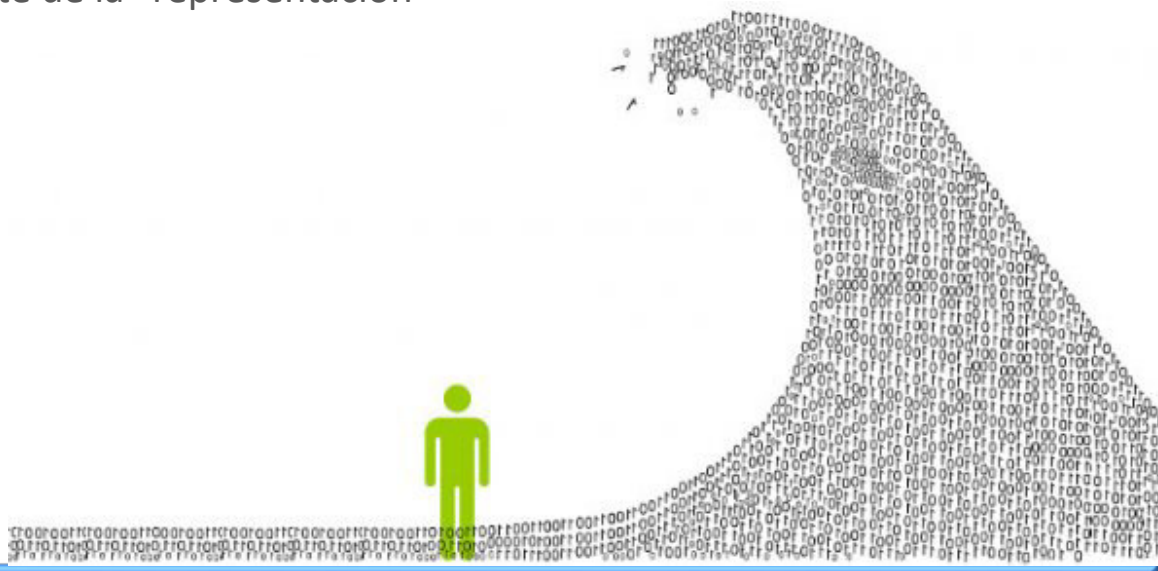
CALIDAD DE DATOS

CD

Calidad de Datos

ALGUNOS CONCEPTOS

- **Información**: morfè (forma) / éidos (concepto)
- Es la **representación** de la **realidad**
- La realidad es diferente de la “representación de la realidad”



CALIDAD DE DATOS

Calidad de datos

Definición #1

Los datos tienen calidad si la información derivada de estos representa correctamente el mundo real (hechos).

Información

Datos

Mundo real

EJEMPLO

Dato1: *Saguinus*

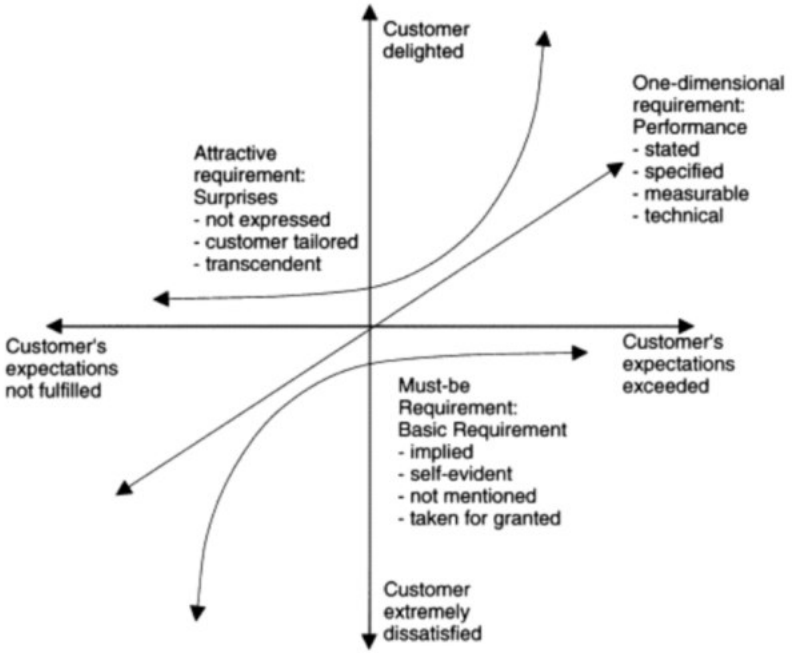
Dato 2: Mico tití

¿tienen la misma calidad?

Calidad de datos

Definición #2

Satisfacción del consumidor. Si un consumidor está satisfecho con un servicio producto, este servicio o producto tiene calidad para este consumidor.



EJEMPLO

Requerimiento: el dato debe tener nombre científico y debe ser suministrado a nivel de especie

Nombre: *Saguinus*
***Categoria**: Genero*

¿este dato tiene calidad?

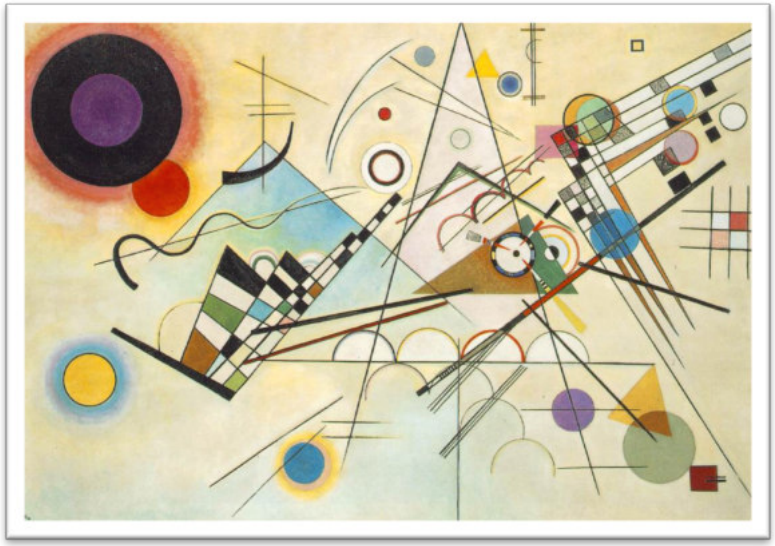
¿puede ser usado en el estudio de la distribución de primates en Suramérica?

Calidad de datos

Definición #3

Usabilidad. Un dato tiene calidad si es adecuado para ser usado. Si el dato no sirve para el propósito del que lo usa, **puede ser útil para otros**.

ALGUNOS CONCEPTOS



La calidad de datos es un concepto idiosincrásico

”La idiosincracia es algo distintivo y propio de un individuo”

Definir calidad de datos es similar a definir qué es bonito, bueno divertido o valioso.

PALABRAS CLAVE

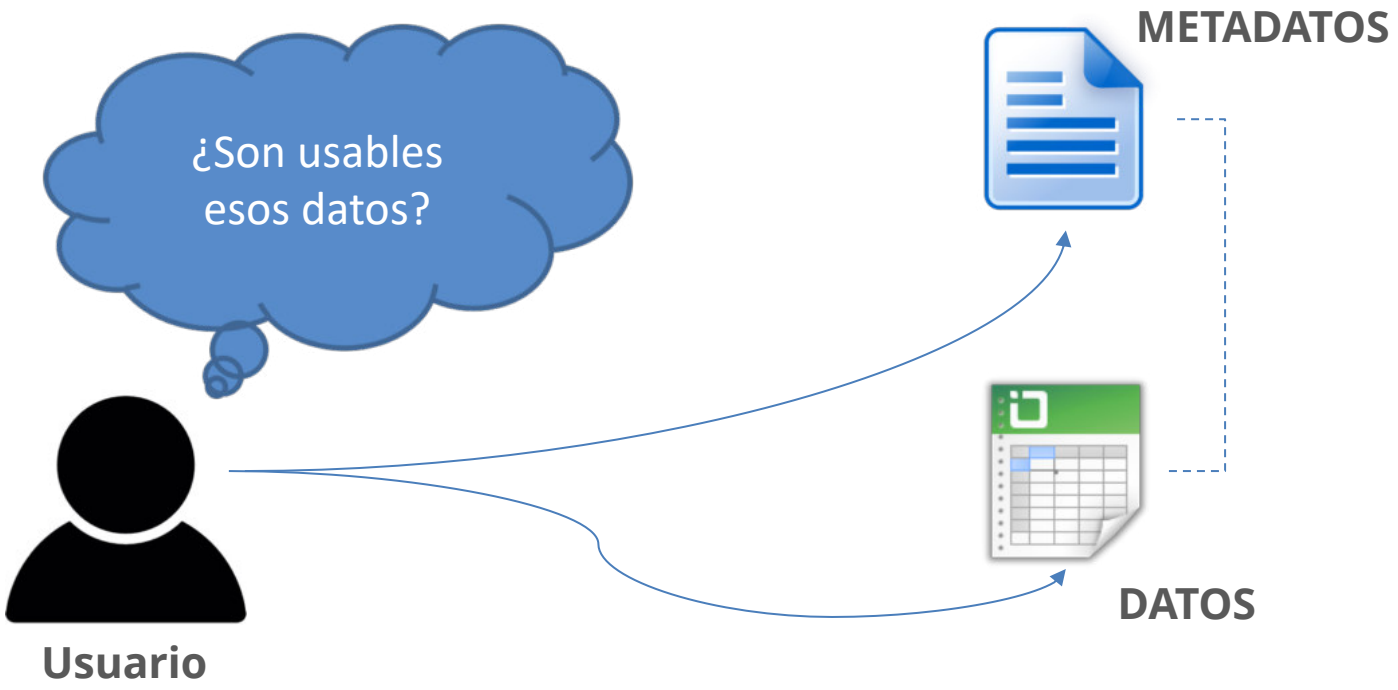
La palabra clave y la definición mas aceptada:
Usabilidad de los datos

USO: Calidad en relación a un propósito.
- *Modelos de distribución, lista nacional de especies, etc.*

DATOS: Para cada propósito existe un tipo de datos.
- *Modelamiento: coordenadas y nombres de especies.*

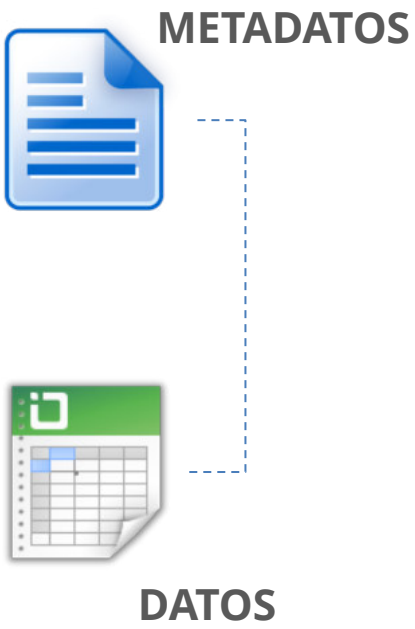
USABILIDAD: Para cada tipo de dato existen atributos a cumplir.
- *Completitud, consistencia, precisión, exactitud, etc.*

USABILIDAD



USABILIDAD

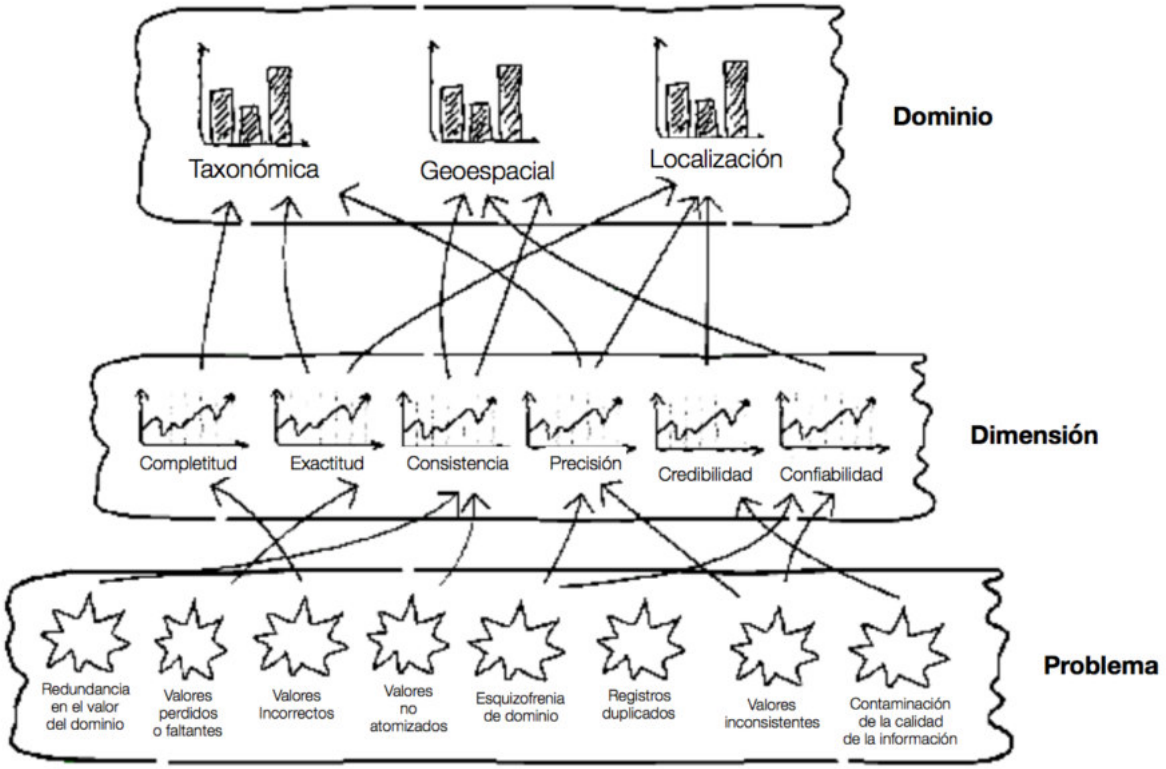

LO SOSPECHE DESDE
UN PRINCIPIO!!
memegenerator.es


METADATOS
DATOS

EVALUACIÓN DE LA CALIDAD

VALIDACIÓN DE CALIDAD

EVALUACIÓN DE LA CALIDAD



DOMINIOS

Geo-espacial: Datos con coordenadas o georereferenciados y regiones político-administrativas documentadas.

Taxonomía (nomenclatura): Nombres científicos, niveles taxonómicos

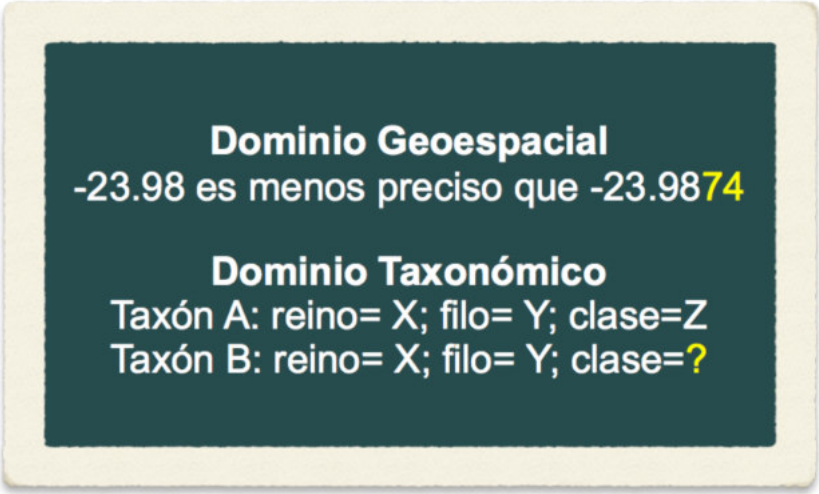
Formato (Estandarización): Errores de tipeo, formatos de fecha, formatos de coordenadas, caracteres especiales y demás.

DIMENSIÓN

Es el aspecto **medible** de la calidad del dato.

Ejemplo: ***Precisión***

* La calidad de datos es un concepto multidimensional



PROBLEMAS

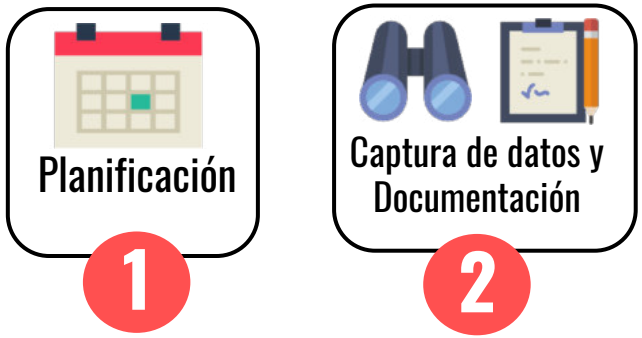
Todo lo que pueda **degradar** la calidad para una o mas **dimensiones**.

- **Domain value redundancy**: Ex.: Brazil x Brasil, São Paulo x Sao Paulo.
- **Missing data value**: absence of latitude or longitude.
- **Incorrect data values**: mispelling errors.
- **Nonatomic data values**: Ex.: São Paulo, SP.
- **Domain schizophrenia**: latitude and longitude equal "0" (zero) when they are not known.
- **Duplicate occurrences**: More than one record representing the same fact.
- **Inconsistent data values**: Ex.: lat=43; long=44; country=Brazil.
- **Information quality contamination**: create a new data based on an already existent wrong data.

CADENA DE LA INFORMACIÓN

CADENA DE LA INFORMACIÓN

CADENA DE LA INFORMACIÓN

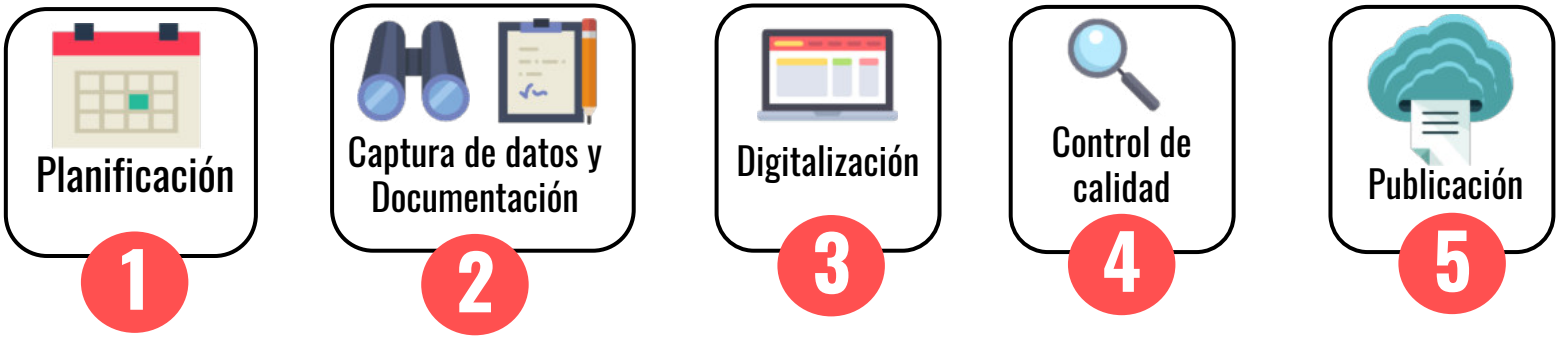


CADENA DE LA INFORMACIÓN

CADENA DE LA INFORMACIÓN

CALIDAD DE DATOS

CADENA DE LA INFORMACIÓN



CADENA DE LA INFORMACIÓN

COSTO DE LA CORRECCIÓN DE ERRORES

MECANISMOS DE MEJORA

PREVENCIÓN: Evitar que se presenten errores previo a la creación de los datos

DETECCIÓN Y LIMPIEZA: Detectar errores en el conjunto de datos y corregirlos

DETECCIÓN Y RECOMENDACIONES: Detectar errores en el conjunto de datos y generar recomendaciones de limpieza

CESP: Regional Capacity Enhancement to Latinamerica by establishing Chile's node
23 al 25 abril 2018
Santiago - Chile

Muchas gracias por su atención

Preguntas?

Leonardo Buitrago
Administrador Nodo GBIF Colombia,
Representante (S) Regional de Nodos GBIF (Latinoamérica y Caribe)
Líder Administración de contenidos SiB Colombia
albuitrago@humboldt.org.co