

# Regional capacity enhancement to Latinamerica by establishing Chile's node



*Project ID: CESP2017-0007*

Primera reunión de mentoring

14- 17 NOV 2017

Bogotá D.C. - Colombia



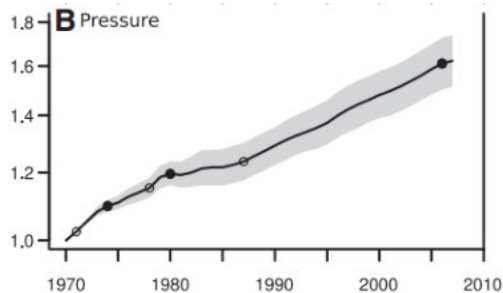
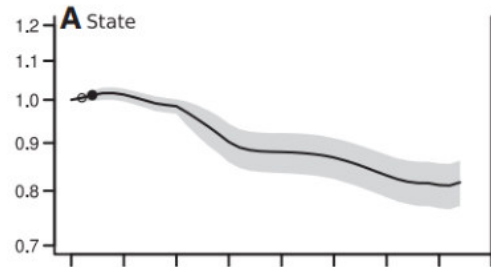


# Conceptos básicos de la calidad y limpieza de datos de biodiversidad

Leonardo Buitrago

Basada en: Saraiva & Koch, 2016. Koch, 2017

# ESTADO GLOBAL DE LA BIODIVERSIDAD



(Stuart H. M. et al 2010)

*CBD Report (2014)*



*Global Biodiversity Outlook (GBO)*



La **presión sobre la biodiversidad** seguirá aumentando al menos hasta el 2020 y el estado de la biodiversidad seguirá deteriorándose!

Hay **avances**, pero, en la mayoría de los casos, estos **no han sido suficientes** para alcanzar los objetivos de 2020.

# PLAN ESTRATÉGICO DE BIODIVERSIDAD 2011-2020 Y METAS AICHI



**Compartir datos**, desarrollar **indicadores y medidas**, fomentando la **generación y uso** de información científica. **Para sostener** la nueva plataforma intergubernamental científiconormativa sobre diversidad biológica y servicios de los ecosistemas (IPBES)



# DE LOS DATOS A LA TOMA DE DECISIONES



**Decisiones**



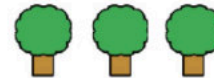
Conocimiento



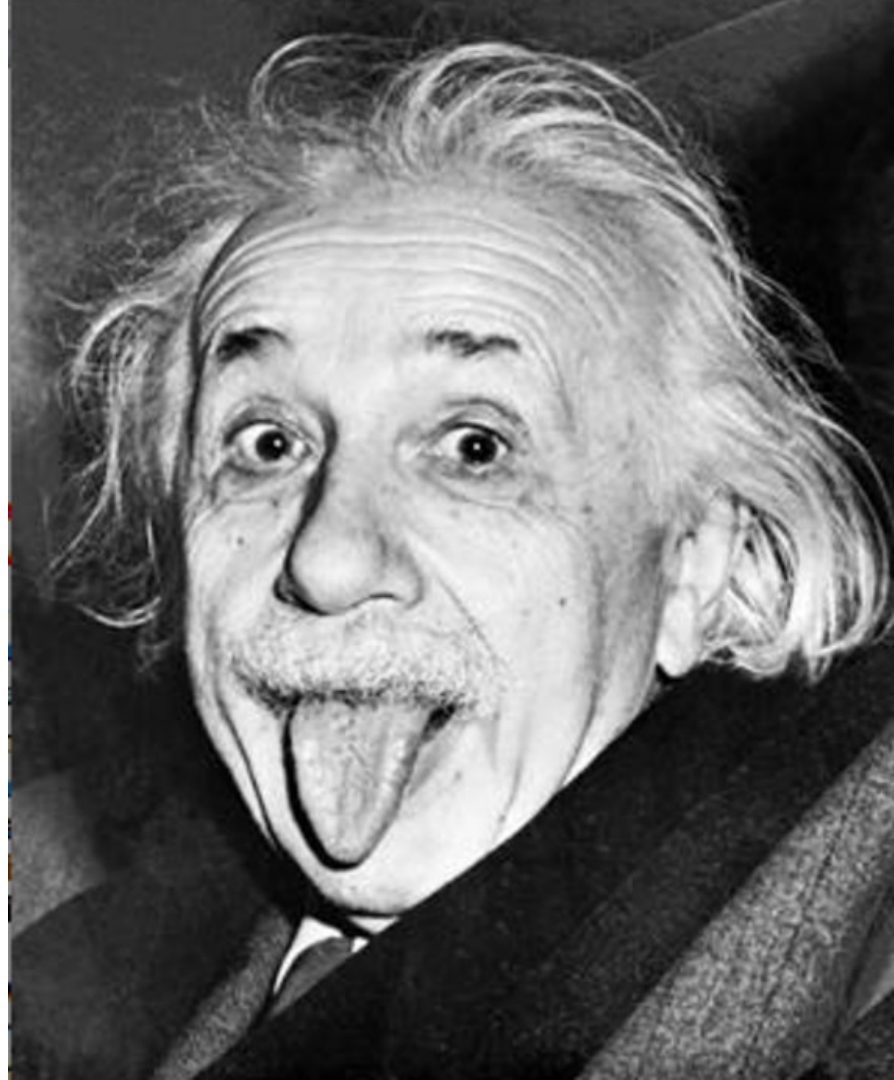
Información



**Datos**



**YA CONTAMOS CON  
BASTANTES DATOS!**





**GBIF**

Global Biodiversity  
Information Facility

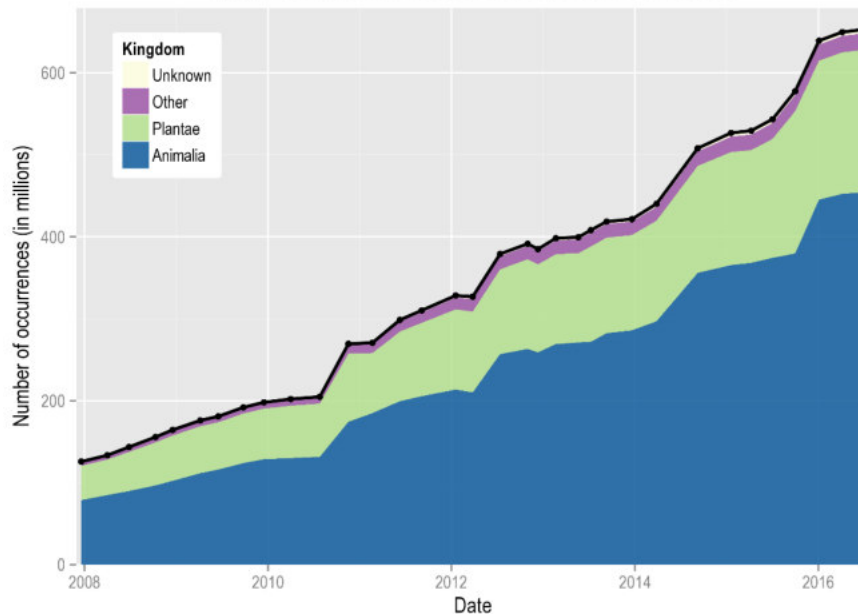
**874.743.999** Registros

**4.564.238** Especies

**37.024** Conjuntos de datos

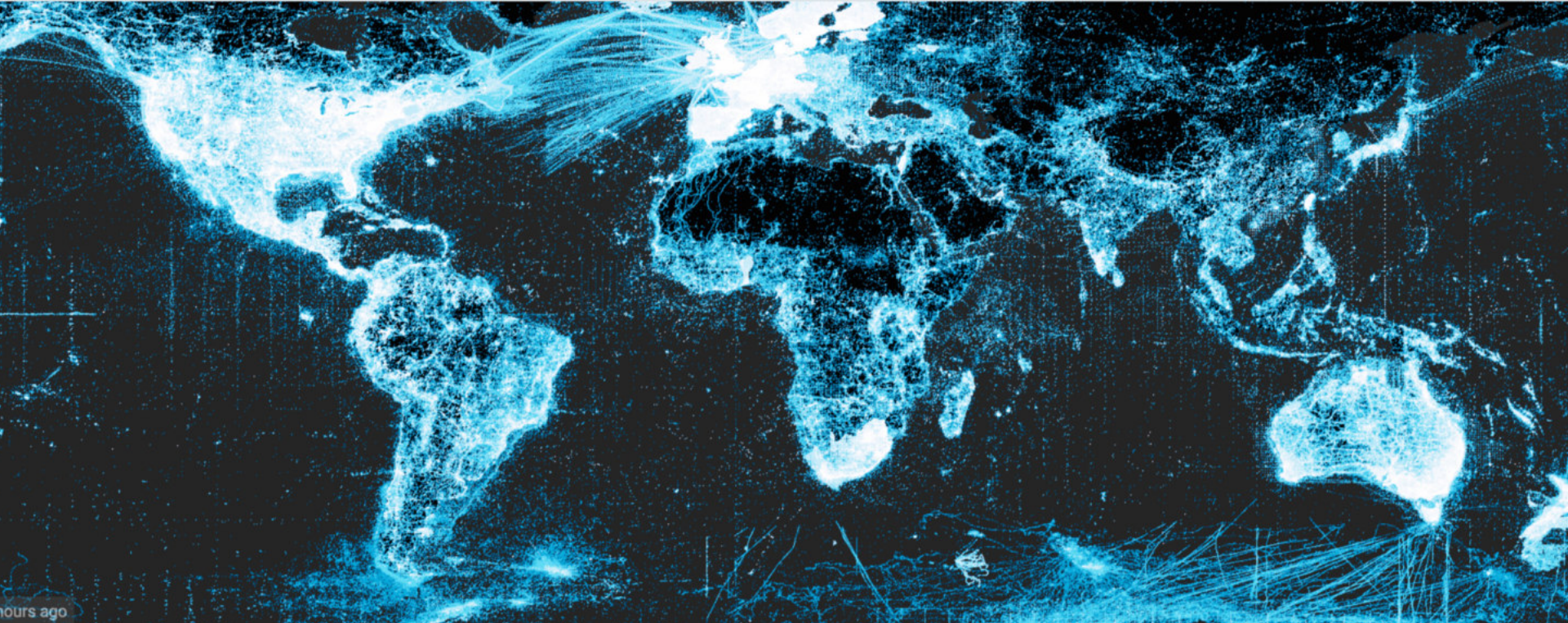
**1.433** Publicadores

Species occurrence records accessible through GBIF over time





# GBIF 2017





# 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?





**KEEP  
CALM  
AND  
COLLECT  
MORE DATA**

## **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/prnjg82>

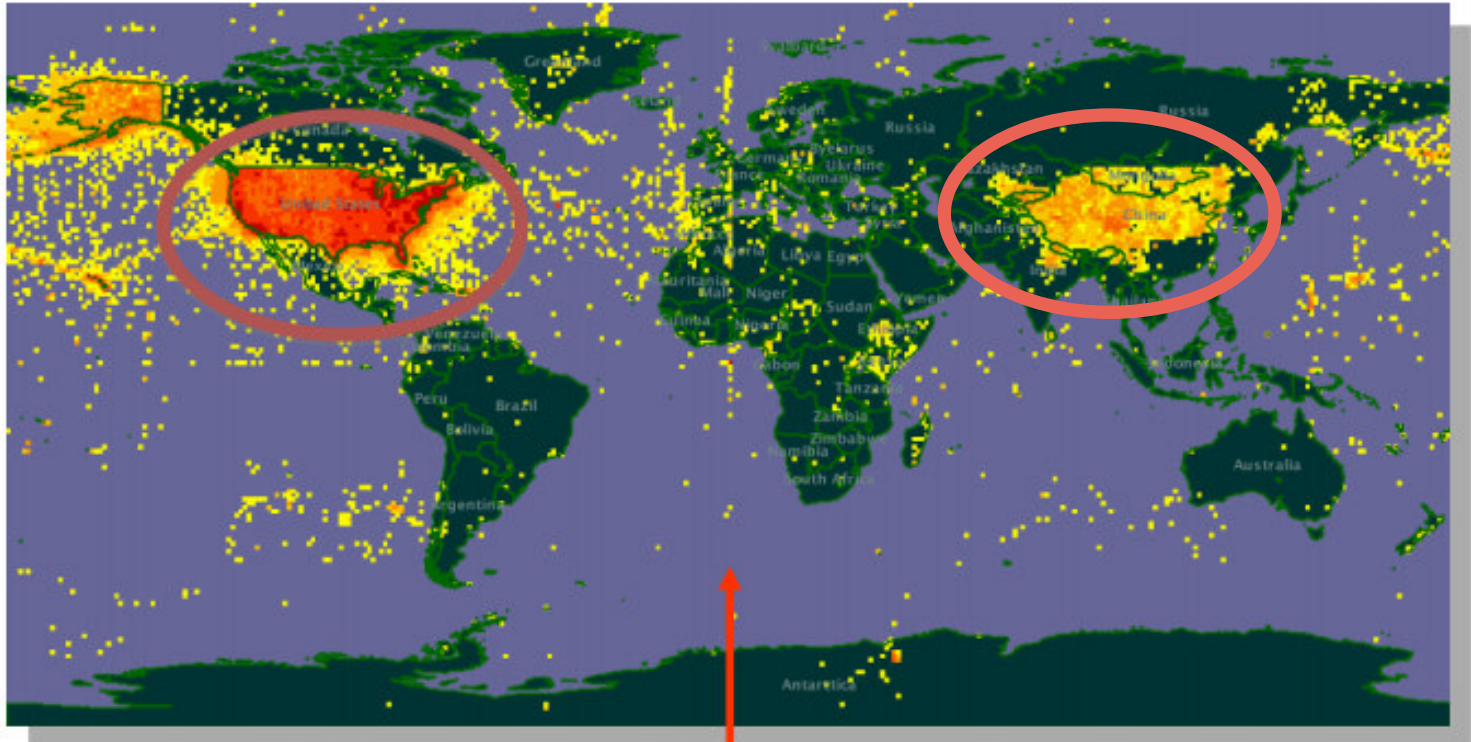
# BASURA ENTRA → BASURA SALE

- ❑ **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



- ☑ Artículos científicos
- ☑ Modelamiento y análisis
- ☑ Políticas de conservación

# EJEMPLO

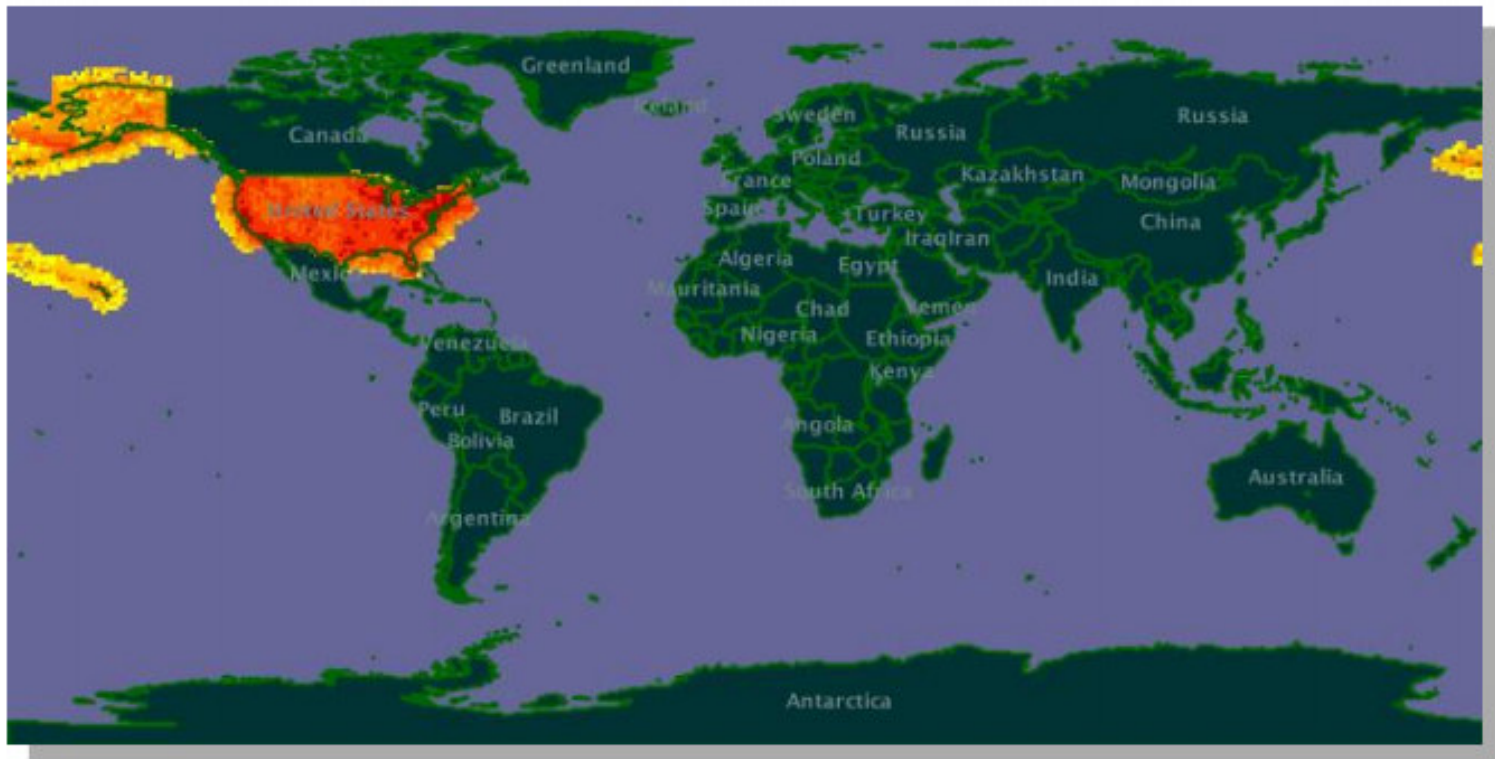




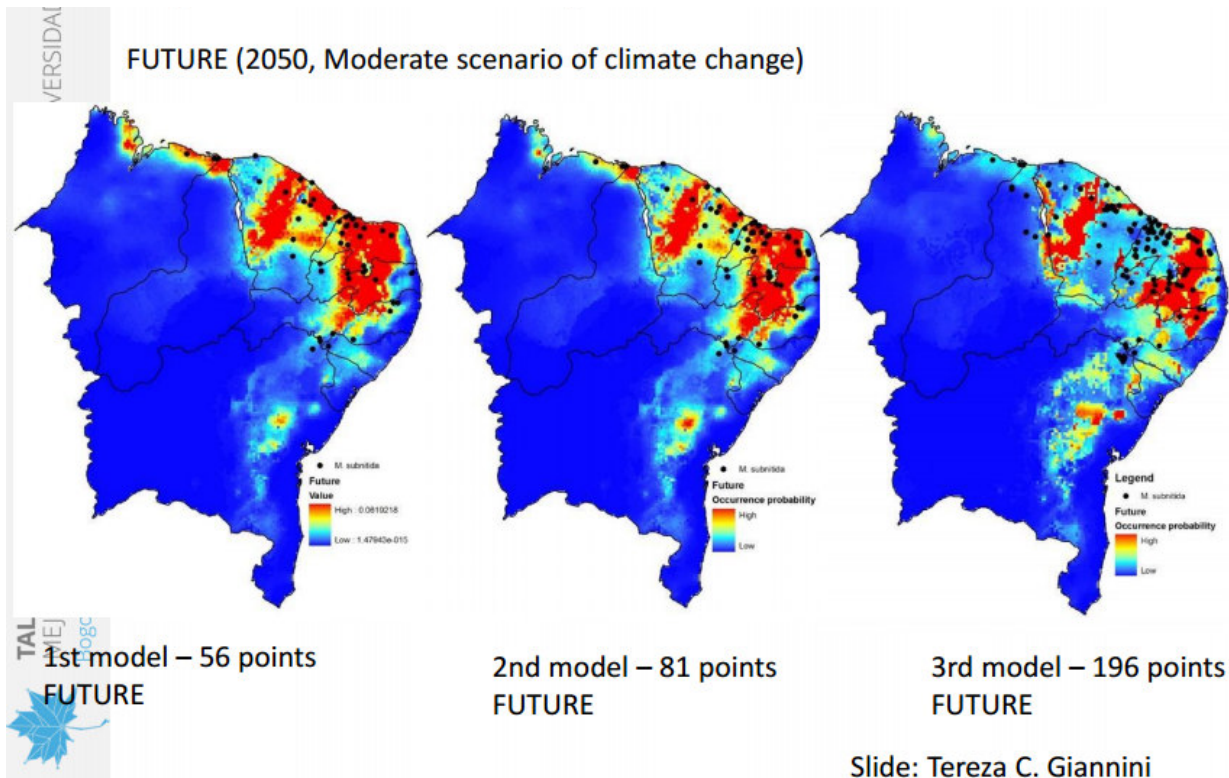
# EJEMPLO



- Wrong longitude (missing minus signal)



# IMPACTO DE LOS DATOS EN LA GENERACIÓN DE MODELOS



Slide: Tereza C. Giannini

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



## IUCN Red List Index

Guidance for national and regional use

**BIO**  
DIVERSIDAD  
**2015**

Estado y tendencias de la biodiversidad continental de Colombia



**BIO**  
DIVERSIDAD  
**2016**

Estado y tendencias de la biodiversidad continental de Colombia



**...y ahora quién podrá ayudarnos?**





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”



# ALGUNOS CONCEPTOS

- Existe una **brecha** entre la **representación de la realidad** y la **realidad misma**, la cual se puede medir en ciertas dimensiones:
  - Completitud
  - Precisión
  - Consistencia
  - Exactitud
  - Etc.



# Calidad de datos

## Definición #1

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



Información

11010101:01111101001011010101:0101010101010101  
10101010110101010110110101010110101010101010101  
11010110:01300010101010130010101011010101010101  
1140110000010110101011031011001010101010101010101  
0001101010101000110100011010101011011010101010101  
01101010:1010110010010010101010101101010101010101010101  
11010101:011110100101010110101010101010101010101010101  
1010101011010101011011010101011010101011010101010101  
1101010101:013000101010101300101010101010101010101010101  
110001000001010101011001011001010010101010101010101010101  
000110101010101010101010001101001101010101011010101010101  
01101010:1010110010010101010101010101010101010101010101  
1101010110101010101011011010101010101010101010101010101  
11010110:01300010101010130010101010101010101010101010101  
11010110:01300010101010130010101010101010101010101010101  
000110101010100011010001101010101101101010101010101010101  
01101010:101011001001010101010101010101010101010101010101

Datos



Mundo real



# EJEMPLO



**Dato1:** *Saguinus*

**Dato 2:** Mico tití

---

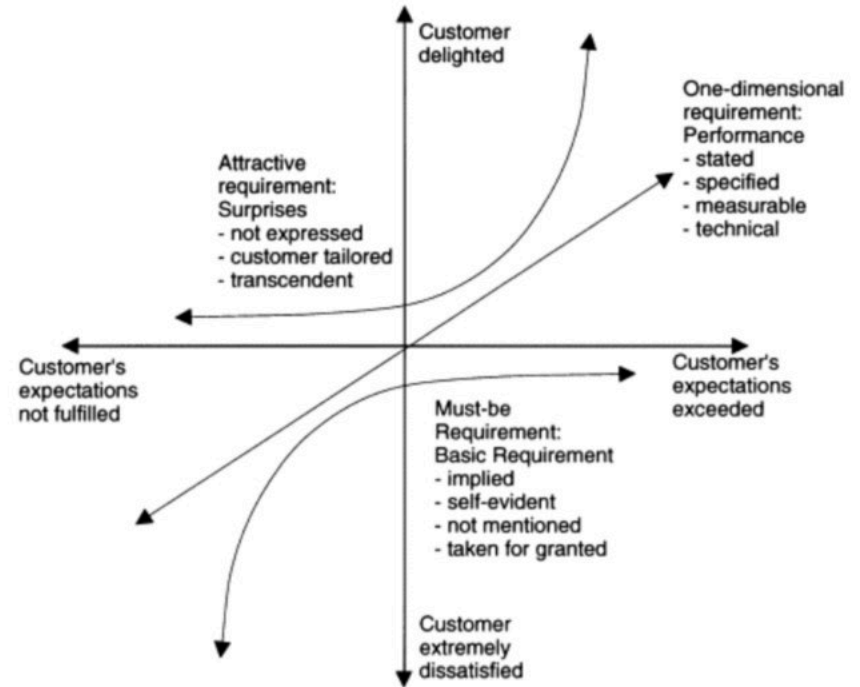
¿tienen la misma calidad?



# 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*

**Categoría:** *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.

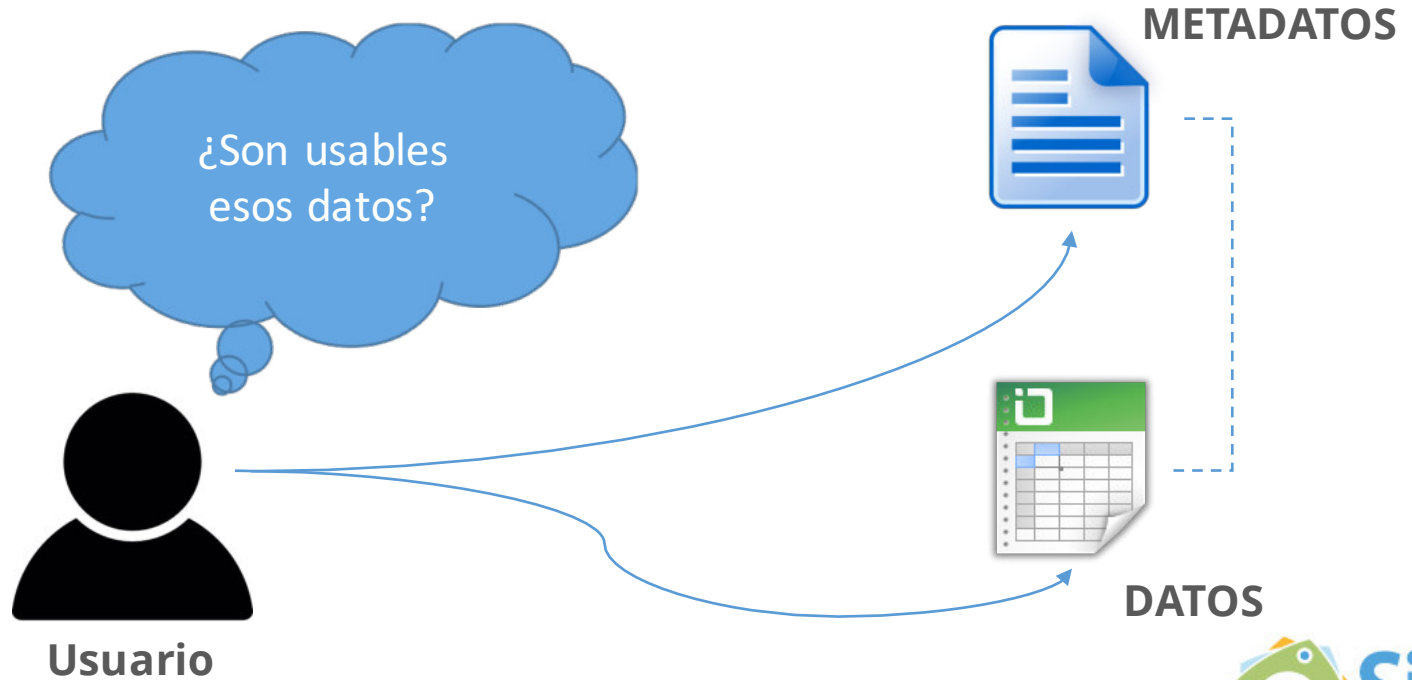
- *Modelamiento: coordenadas y nombres de especies.*

**USABILIDAD**: Para cada tipo de dato existen atributos a cumplir.

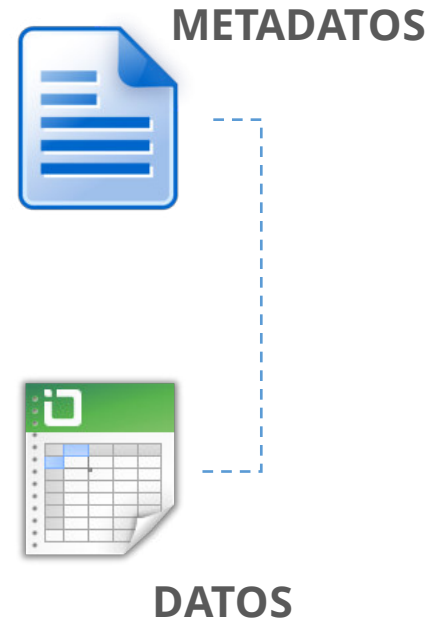
- *Compleitud, consistencia, precisión, exactitud, etc.*



# EVALUACIÓN DE LA CALIDAD



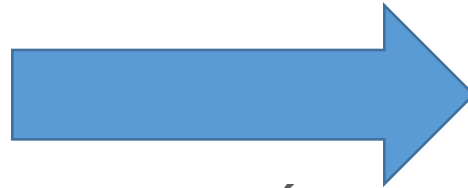
# EVALUACIÓN DE LA CALIDAD



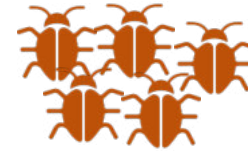
# EVALUACIÓN DE LA CALIDAD



DATOS



*EVALUACIÓN*



**NO USABLES**



**USABLES**

# EVALUACIÓN DE LA CALIDAD



DATOS



*VALIDACIÓN*

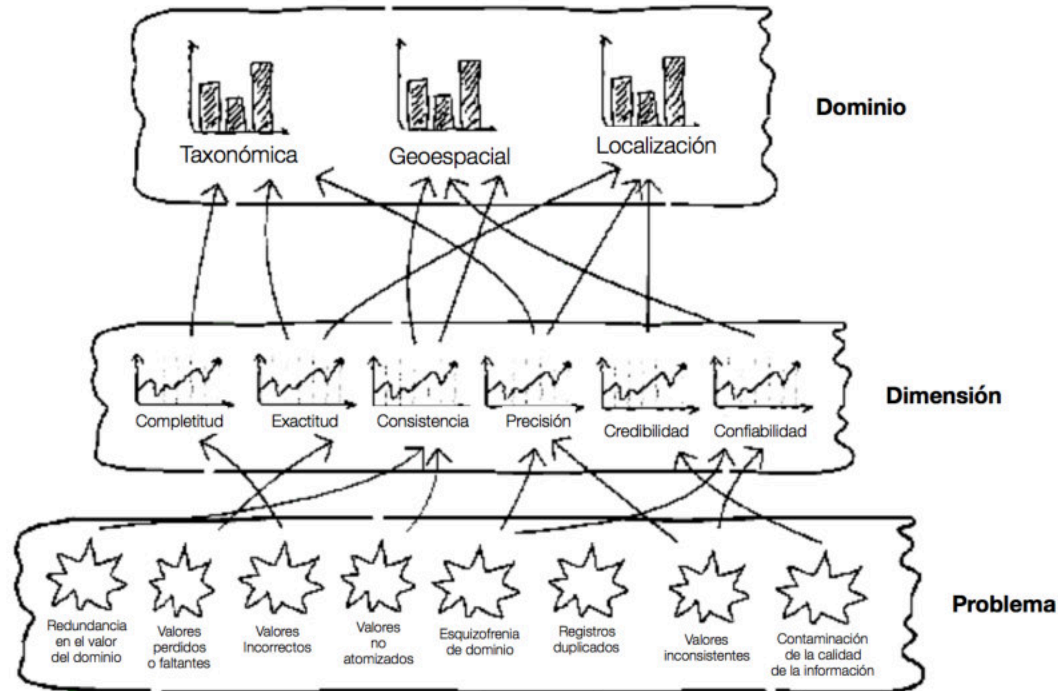


**NO USABLES**



**USABLES**

# EVALUACIÓN DE LA CALIDAD





# DOMINIOS



***Geo-espacial:*** Datos con coordenadas o georreferenciados 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

## **Dominio Geoespacial**

-23.98 es menos preciso que -23.9874

## **Dominio Taxonómico**

Taxón A: reino= X; filo= Y; clase=Z

Taxón B: reino= X; filo= Y; clase=?

# 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:** misspelling 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.

# 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

# CADENA DE LA INFORMACIÓN

COSTO DE LA CORRECCIÓN DE ERRORES





# CADENA DE LA INFORMACIÓN

COSTO DE LA CORRECCIÓN DE ERRORES



Planificación

1

# CADENA DE LA INFORMACIÓN

COSTO DE LA CORRECCIÓN DE ERRORES



1



2

# CADENA DE LA INFORMACIÓN

COSTO DE LA CORRECCIÓN DE ERRORES



1



2



3

# CADENA DE LA INFORMACIÓN

## COSTO DE LA CORRECCIÓN DE ERRORES



1



2



3






4

# CADENA DE LA INFORMACIÓN

## COSTO DE LA CORRECCIÓN DE ERRORES



# ¡Gracias!

facebook/sibcolombia   
twitter/sibcolombia   
youtube/sibcolombia 



 Instituto Humboldt

**Leonardo Buitrago**  
Administración de contenidos

albuitrago@humboldt.org.co  
www.sibcolombia.net

