PRINCIPIOS GENERALES SOBRE LA CALIDAD DE DATOS

Cristina Villaverde – GBIF España

Katia Cezón - GBIF España

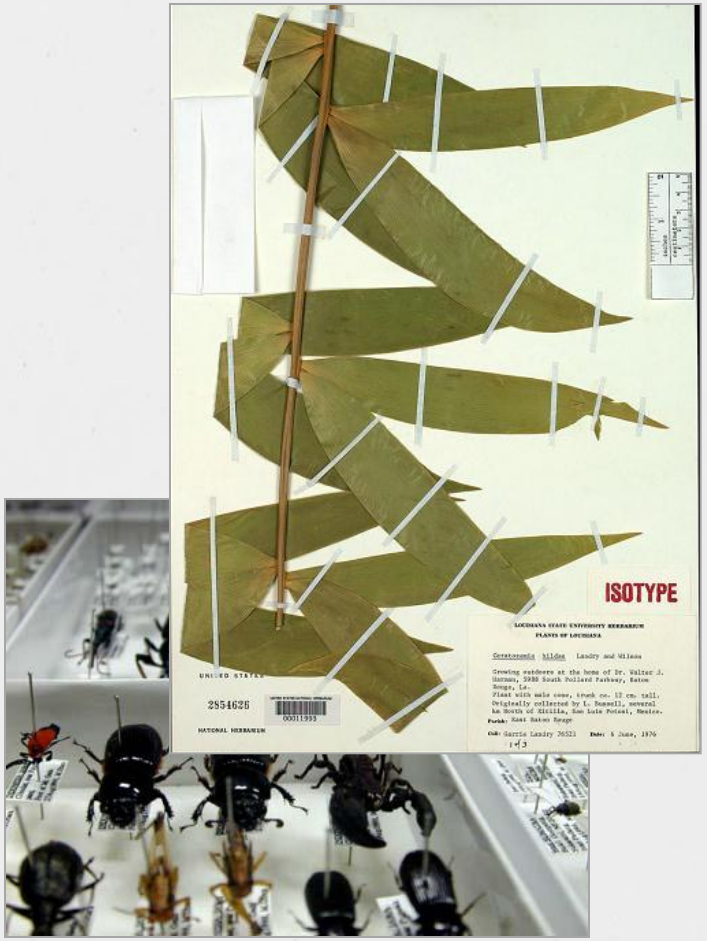
Colecciones Biológicas 3,0
Villa de Leyva, Colombia 2012

CHAPMAN, A.D. 2005. Principles of Data Quality, version 1.0.

Report for the GBIF, Copenhagen.

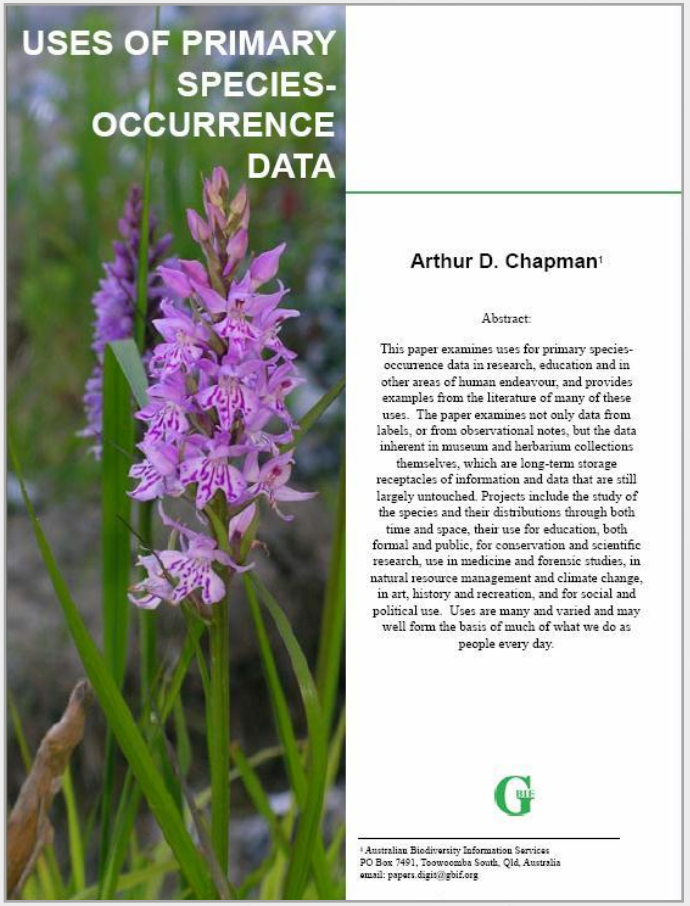
Usando los datos sobre ejemplares

- Estudios taxonómicos, ecológicos, biogeográficos, filogenéticos.
- Estudios de población y distribución de especies.
- Estudios sobre especies amenazadas.
- Sobre migración de especies.
- Planificación sobre la conservación de espacios protegidos.
- Gestión de recursos naturales.
- Modelado de datos de especies.
- Impacto del cambio climático.



Usando los datos sobre ejemplares

- Agricultura, Montes y Pesca
- Perspectivas basadas en productos biológicos.
- Salud y seguridad públicas.
- Medicina forense.
- Ecoturismo.
- Arte e Historia, Ciencias y política
- Planificación de infraestructuras humanas.
- Etc.



¿Qué es la calidad de datos?

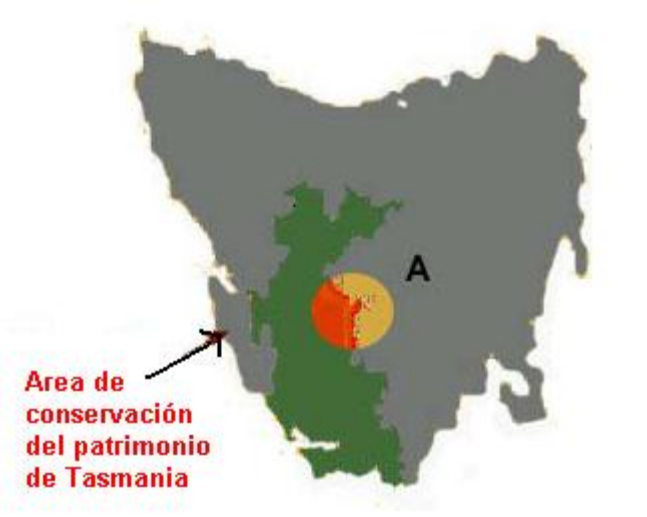
*Una característica esencial y necesaria para que los datos sean "**adecuados para el uso**".*

El propósito general al describir la calidad de los datos de un registro concreto es describir la **adecuación** del registro para **un uso particular** que el usuario tenga en mente para dichos datos.

Chrisman, 1991

¿Qué es la calidad de datos?

¿Este dato es de buena calidad?:



- ¿La especie 'A' se encuentra en Tasmania?
- ¿La especie 'A' se encuentra en el Área de conservación del patrimonio de Tasmania'?

CADENA DE INFORMACIÓN Y PÉRDIDA DE LA CALIDAD

Cadena de información de la calidad de datos

COSTE DE LA CORRECCIÓN DE ERRORES

PLANIFICACIÓN | RECOLECCIÓN | DOCUMENTACIÓN | DIGITALIZACIÓN | CONTROL DE CALIDAD | PUESTA EN INTERNET

- No Planificación

- Información incompleta

- Poca experiencia del personal
- Mala interpretación
- Base de datos
- Copias

- Mala exportación
- Conversiones
- Uso incorrecto de los datos

Institución

VISIÓN INSTITUCIONAL

- Reconocimiento de la información como fundamental en los procesos institucionales
- Se busca maximizar interoperabilidad
- Orientación de la calidad de datos a largo plazo

POLÍTICA DE CALIDAD

- Definir qué se va a hacer respecto a la calidad

ESTRATEGIA

- Definición de normas y procedimientos para obtener la calidad que se busca

PRINCIPIOS DE LA CALIDAD

Principios de la calidad

Prevenir es mejor que curar, y es mucho más barato…

La prevención de errores nada tiene que hacer con los datos que **ya existen** en la base de datos. En estos casos, la **validación** y la **corrección** serán muy **importantes** en el proceso hacia la calidad.

Detectar las causas del error nos ayudará a prevenirlas

Corregir los datos y no hacer nada para prevenir los errores significa que los errores seguirán apareciendo sistemáticamente y no los reduciremos nunca.

Principios de la calidad

EXACTITUD – Debemos tender a que el dato esté lo más próximo posible al valor real.

CONSISTENCIA

Datos presentados siempre de la misma manera y se mantienen en el tiempo de forma clara, consistente y sin ambigüedad:

- **Consistencia semántica:** la información que se almacena
- **Consistencia estructural:** la forma en que se almacena

Genus	Species	Infraspecies
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

Table 1. *Showing semantic inconsistency in the Infraspecies field.*

Genus	Species	Infrasp_rank	Infraspecies
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

Table 2. *Showing semantic consistency in the Infraspecies field by addition of a second ("Infrasp_rank") field.*

Principios de la calidad

DEPURACIÓN – Detección y exclusión de los datos que no sean correctos ni consistentes.

EFECTIVIDAD – La probabilidad de que una tarea alcance los objetivos deseados.

- Ej: el porcentaje de registros para los cuales la latitud y la longitud pueden ser determinados exactamente.

EFICIENCIA – Producir los máximos resultados con los mínimos recursos.

- Ej: optimizar los procesos de georreferenciación ordenando por localidad y georreferenciándolos utilizando los mismos mapas para este conjunto de registros.

Principios de la calidad

ACCESIBILIDAD – cómo de accesibles son los resultados para los usuarios/el público.

- Ej: la facilidad con la que los usuarios acceden a la georreferenciación de una localidad particular que acaba de ser georreferenciada.

TRANSPARENCIA – hacer públicos los procedimientos y la documentación para el manejo de la colección, los análisis realizados, los informes y las actualizaciones.

- Ej.: conocer los métodos con que han sido georreferenciados un grupo de registros y disponer de la documentación asociada a esta tarea.

Principios de la calidad

ACTUALIDAD DE LOS DATOS – Se refiere a la frecuencia de actualización del conjunto de los datos de la colección.

- ¿Cuándo fueron los datos actualizados por última vez?
- ¿Con qué frecuencia se actualizan y son puestos a disposición de los usuarios?

La frecuencia de actualización deben ser concretada y documentada.

También debemos tener en cuenta…

Otras características que deben ser observadas son:

- **Documentación:** es un principio clave. Permite a los usuarios verificar si los datos se ajustan al uso que necesitan en ese momento.
- **Feedback:** mecanismo de retroalimentación a través del cual los usuarios informan a cerca de errores, y hacen que esta información se refleje en la calidad de los datos.
- **Formación y entrenamiento del personal:** debe incluir desde los colectores, hasta los operadores de digitalización de los datos y los gestores de las bases de datos.
- Crear **protocolos de actuación** que sirvan de base para la formación del personal, y para las tareas del día-a-día.

Errores en los datos

En general, un buen entendimiento de los errores y su propagación conduce a un control activo de la calidad.

Burrough and McDonnell, 1998

Ya que el error es algo ineludible, debería ser reconocido como una dimensión fundamental del dato. Y necesita ser detectado, registrado y documentado.

Chrisman, 1991

CALIDAD DE DATOS EN GBIF-ESPAÑA

GBIF.ES Y LA CALIDAD DE LOS DATOS

Formación

TALLERES DE CALIDAD

Herramienta de validación

DARWIN TEST

Repositorio

BIODIVERSITY DATA QUALITY HUB (BDQ)

GBIF.ES Y LA CALIDAD DE LOS DATOS

Formación

Desde 2007

Talleres presenciales y online

- III Taller GBIF sobre calidad en bases de datos sobre biodiversidad (2009)
  http://www.gbif.es/formaciondetalles.php?IDForm=60
- Entorno Virtual de Formación GBIF.ES: III Taller de Calidad en bases de datos sobre biodiversidad (2012)
  http://elearning.gbif.es/AContent/home/course/content.php?_cid=77

Vídeos de las sesiones teóricas

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

GBIF.ES Y LA CALIDAD DE LOS DATOS

Formación presencial

GBIF.ES Y LA CALIDAD DE LOS DATOS

Formación en línea

ATUTOR

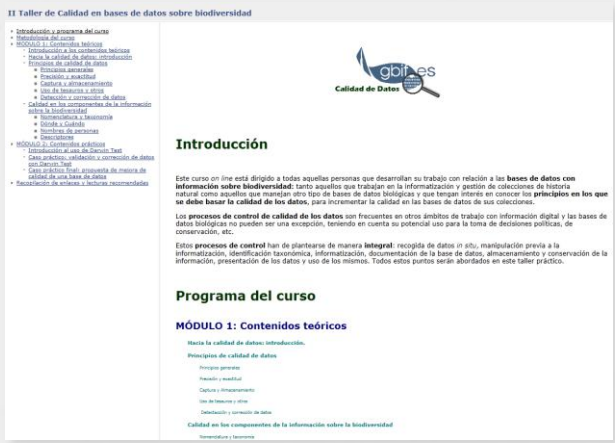
http://elearning.gbif.es/login.php

GBIF.ES Y LA CALIDAD DE LOS DATOS

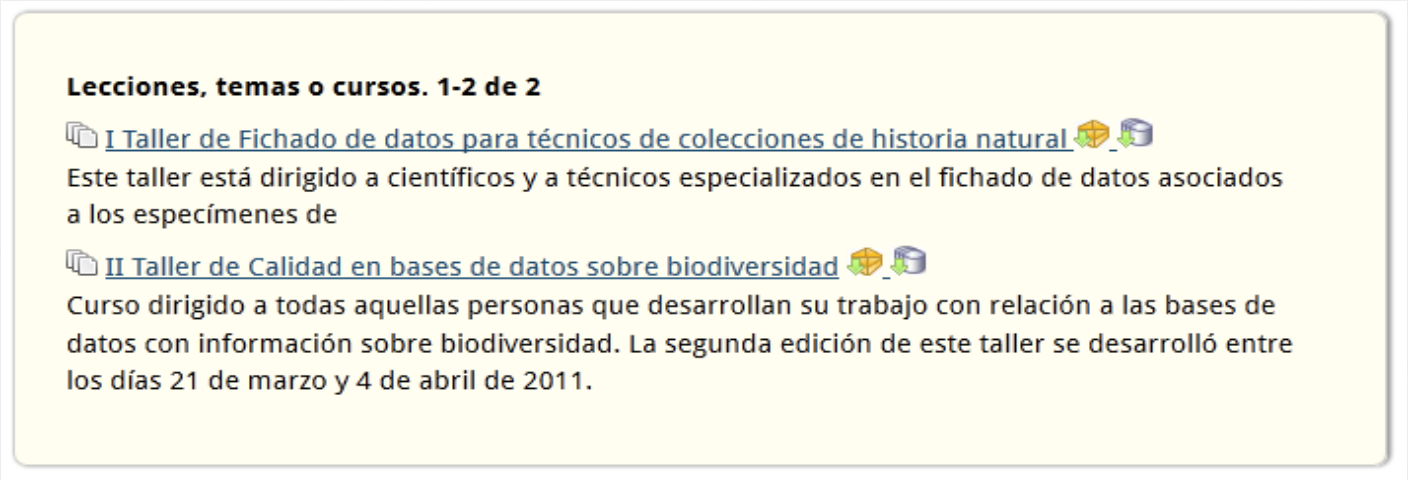
Formación en línea

ACONTENT

http://elearning.gbif.es/AContent/home/index.php



Paquete SCORM



GBIF.ES Y LA CALIDAD DE LOS DATOS

Herramienta de validación

- **Darwin Test**
- Software MS Access Windows
- Interfaz gráfico de fácil manejo a través de formularios
- Validación y chequeo de los datos en formato Darwin Core
- Detecta errores de varios tipos:
  - Errores de omisión
  - Errores tipográficos
  - Errores de convención
  - Errores de congruencia
- Permite corregirlos de una manera sencilla desde los propios formularios de validación

GBIF.ES Y LA CALIDAD DE LOS DATOS

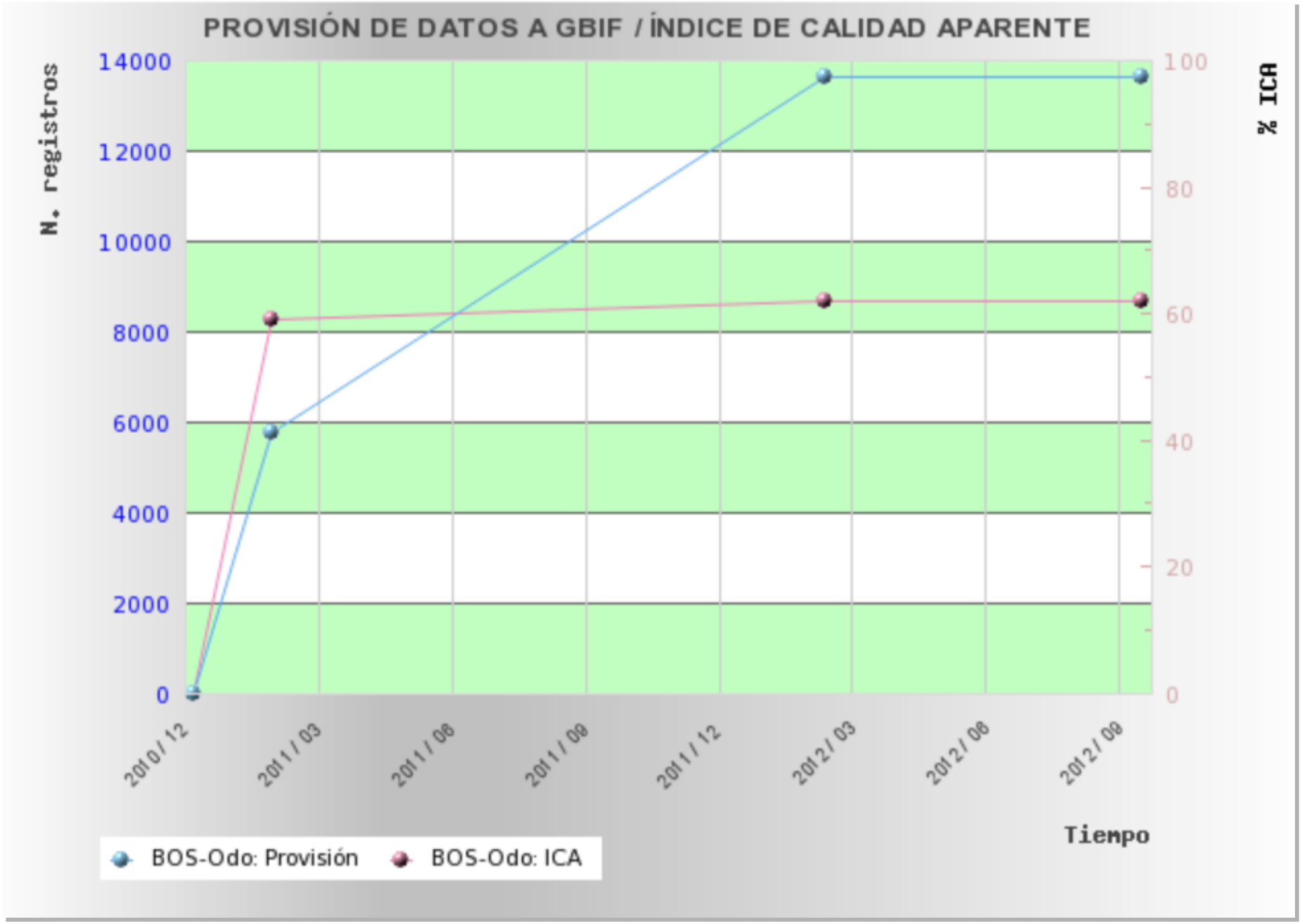
CARACTERÍSTICAS I

- V**alidar** y **chequear** las tablas en formato Darwin Core 1.2 y 1.4.
- **Activación/desactivación** de las consultas existentes.
- **Creación de nuevas consultas.**
- **Corregir los errores** detectados.
- **Importación/exportación de archivos con formato** Darwin Core Archive**.**
- **Importación de datos procedentes de archivos eml y xml**.

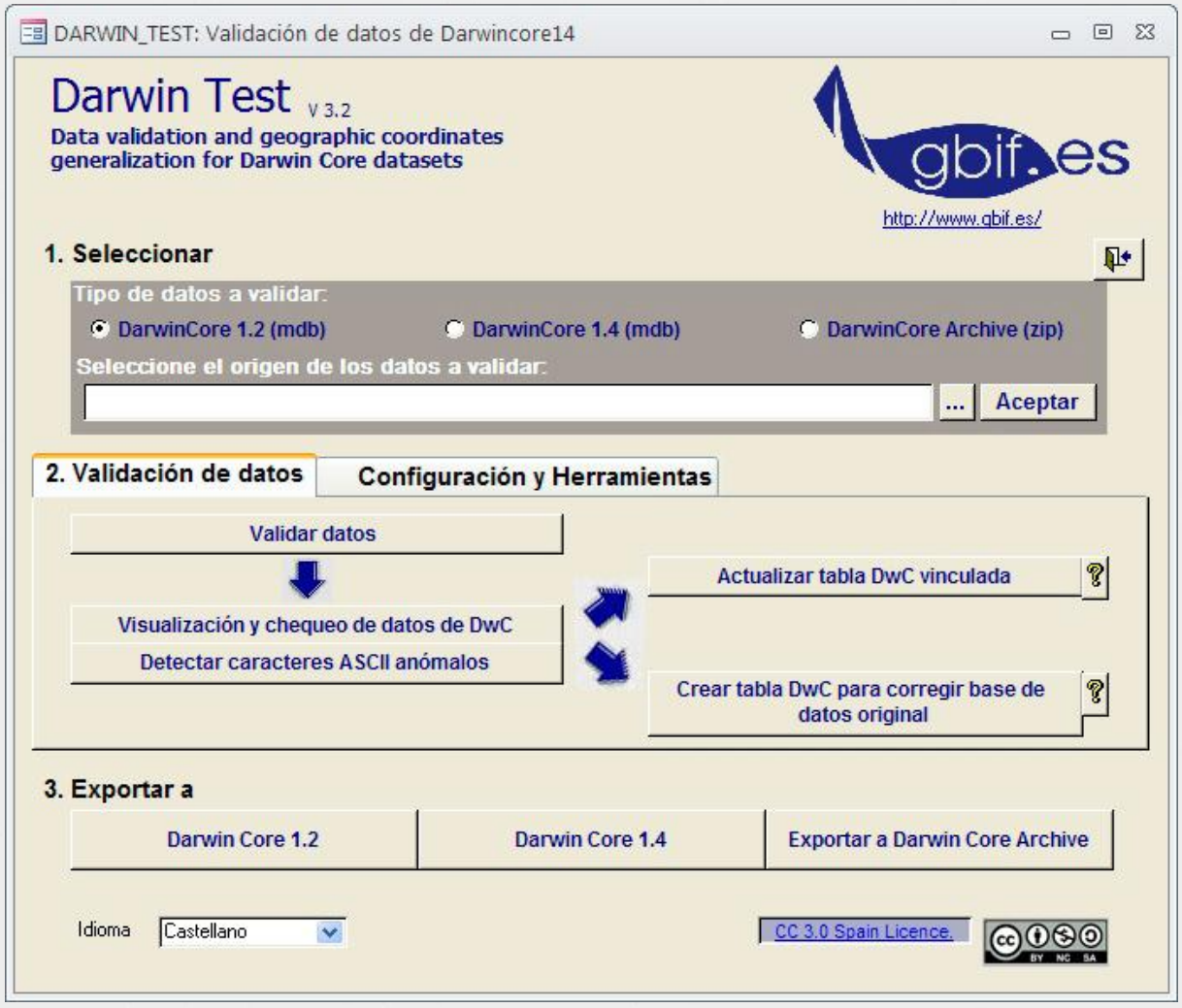
GBIF.ES Y LA CALIDAD DE LOS DATOS

CARACTERÍSTICAS II

- Chequeo de los **nombres científicos**:
  - The Catalogue of Life/Species 2000
  - Tabla Archivos de Autoridad Taxonómica (AAT) del SiB Colombia
  - Otras base de datos con nombres científicos.
- **Conversión de coordenadas** a geográficas en grados decimales.
- Detecta y elimina **caracteres ASCII anómalos.**
- Permite la **creación y gestión de filtros** de usuario para eliminar registros o generalizar coordenadas selectivamente de las tablas **DARWINCOREV2 y DARWINCOREV14**.
- Creación del **Índice de Calidad Aparente (ICA)** para el seguimiento de la mejora de la calidad de los datos**.** http://www.gbif.es/ICA.php


PROVISIÓN DE DATOS A GBIF / ÍNDICE DE CALIDAD APARENTE
N. registros
% ICA
14000
12000
10000
8000
6000
4000
2000
0
100
80
60
40
20
0
2010/12
2011/03
2011/06
2011/09
2011/12
2012/03
2012/06
2012/09
Tiempo
BOS-Odo: Provisión
BOS-Odo: ICA

GBIF.ES Y LA CALIDAD DE LOS DATOS



1

2

3

4

5

GBIF.ES Y LA CALIDAD DE LOS DATOS

DESCARGA

- Desde la página web: http://www.gbif.es/darwin_test/Darwin_Test.php
- MS Access 2003 o 2007
- **Código fuente** accesible desde la **página del proyecto en Sourceforge.net** bajo licencia Creative Commons
- Última versión julio 2012
- Manual 3,2

Descarga de la versión: Darwin Test 3.2		
Versión **para usuarios** MS Access **2003**	Manual 3.2	Julio 2012
Versión **para usuarios** MS Access **2007**		
Programa fuente **para desarrolladores** desde Sourceforge		
Programa fuente **para desarrolladores** desde GBIF-ES (MS Access 2003)		
Programa fuente **para desarrolladores** desde GBIF-ES (MS Acess 2007)		

GBIF.ES Y LA CALIDAD DE LOS DATOS

Biodiversity Data Quality (BDQ):

- Localizador de recursos relacionados con la calidad de los datos de biodiversidad
- Reunión nodos europeos de GBIF 2011
- Compatible con el *Centro de Recursos en Línea* de GBIFS
- Estructura: herramientas, tesauros, formación y procesos y experiencias
- Formulario para añadir recursos
- http://www.gbif.es/BDQ
- Presentación flash

GBIF.ES Y LA CALIDAD DE LOS DATOS

Biodiversity Data Quality (BDQ):

>> Please fill out this form to include your biodiversity data quality information or item
click **here** to send us feedbak

Biodiversity Data Quality Hierarchy

Detection Tools

- Darwin Test -- [metadata]
- GEOLocate -- [metadata]
- Species Link - infoXY -- [metadata]
- Species Link - SpOutlier -- [metafata]
- Georeferencing Calculator -- [metadata]
- Diva GIS -- [metadata]
- QuantumGIS -- [metadata]
- The R-project -- [metadata]
- ECAT name parser -- [meta
- Name Finder -- [metadata]
- Taxon Tagger -- [metadata]
- Google Fusion Tables -- [me
- Google Refine -- [metadata]
- OpenStreetMap -- [metadat
- NBD Record Cleaner -- [me

Validation Tools

- Darwin Test -- [metadata]
- GBIF's Darwin Core Archive

Thesauri: Checklists

- Animals in general: *Index t*
- Mammals: Mammal Species
- Birds: AviBase
- Fish: FishBase
- Vascular plants: *Internation*
- Mosses: W3Most and also in
- Fungi and lichens: Index Fu
- Algae: AlgaeBase, Index No
- Bacteria: *List of Bacteria with Standing in Nomenclature* (LBSN)
- Virus: *The Universal Virus Database of the International Committee on Taxonomy of Viruses* (ICTVdB)
- All organisms: Catalogue of Life

Thesauri: ISO Codes

- Countries ISO 3166-1 codes as MS Access
- Country subdivisions ISO 3166-2 codes as MS Access

Training

- GBIF.ES Online Workshop on Data Quality -- [metadata]
- GBIF.ES Sensitive Data Management Workshop -- [metadata]
- GB18 GBIF Biodiversity data quality and fitness-for-use training sessions -- [metadata]

Procedures and Best Practices: Guides

- Principles of Data Quality - GBIF
- Uses of Primary Species Occurrence Data - GBIF
- Guide to Best Practices for Georeferencing - GBIF
- Principles and Methods of Data Cleaning - GBIF
- Digitisation and Data Quality Control of Mexican and Central American Botanical Specimens... - CONABIO
- Improving Wildlife Data Quality - National Biodiversity Network Trust

Procedures and Best Practices: Experiences

- Practical approaches to data quality: training and tools and video
- Atlas of Living Australia - Data Quality Portal -- [metadata]
- GBIF Online Resource Centre -- [metadata]

Cristina Villaverde

Unidad de Coordinación de GBIF
Real Jardín Botánico -CSIC
Claudio Moyano 1
28014 Madrid, Spain

villaverde@gbif.es
www.gbif.es
Telf: + 34 91 420 3017

GRACIAS