

GLOBAL  
BIODIVERSITY



# INFORMATION FACILITY

Data quality, cleaning and  
dealing with sensitive data

Francisco Pando, GBIF Spain



NATIONAL MUSEUMS OF KENYA  
WHERE HERITAGE LIVES ON

KENBIF TRAINING WORKSHOP

Nairobi, 6th to 7th June 2011

# Part 1: Principles

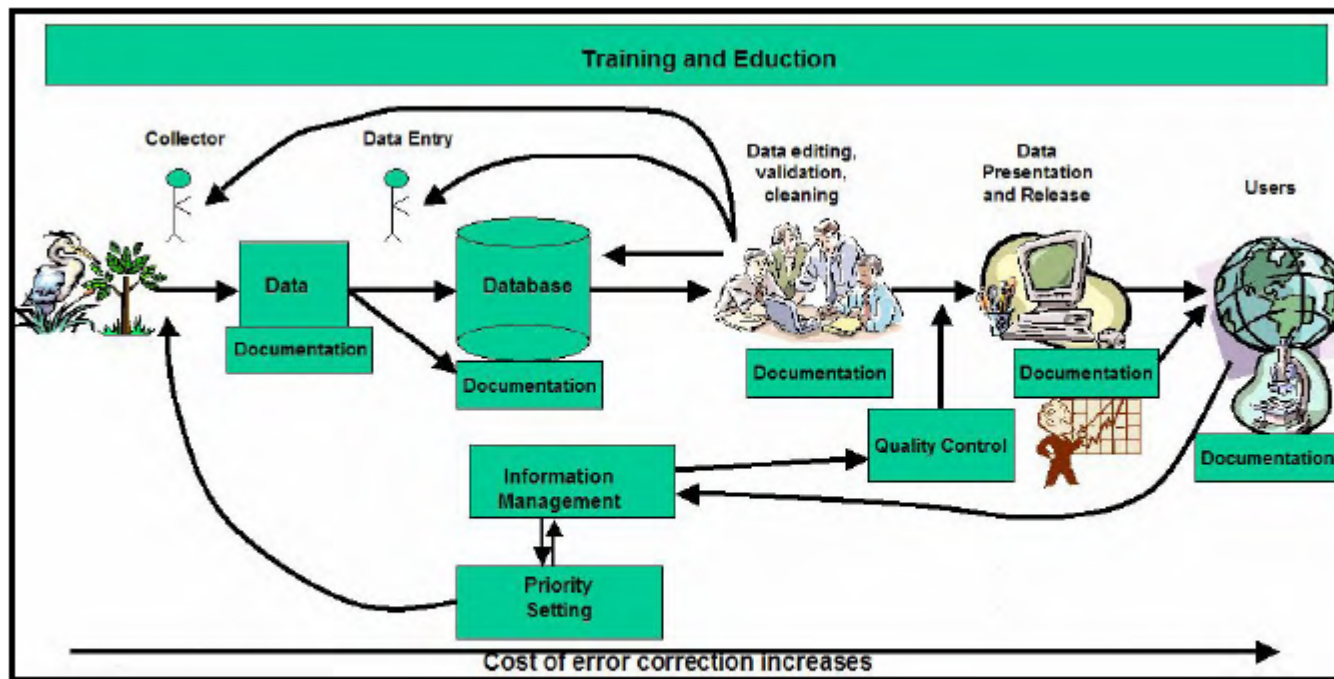
---

Based on presentations by Vishwas Chavan 'Senior Programme Officer for DIGIT - GBIFS, in turn based on A. Chapman's "Principles of data Quality"

- **Primary biodiversity data can be used for multiple purposes by various user communities worldwide.**
  - **Assessing and enhancing fitness-for-use of data is therefore critical for the scientific and social relevance of biodiversity science.**
    - **Fitness-for-use varies from one use case to another.....**
    - **Data quality assessment and quality control are important components of 'fitness-for-use' regime**
-

# Loss of Data Quality

- At the time of collection
- During digitisation
- During documentation
- During storage and archiving
- During analysis and manipulation
- During dissemination and presentation
- Through the use to which they are put



**Fig. 1.** Information Management Chain showing that the cost of error correction increases as one moves along the chain. Education, Training and Documentation are integral to all steps (from Chapman 2005a).

**It is important for a organisation to have**

- **a vision with respect to having good quality data**
- **a policy to implement that vision, and**
- **a strategy for implementation**

## A Vision may involve

- **Not reinventing** information management wheels
- Looking for **efficiencies** in data collection and quality control procedures
- **Sharing** data, information and tools
- **Using** existing **standards** or develop new, robust standards
- **Fostering** the development of networks and **partnerships**
- Presenting a **sound business case** for data collection and management
- **Reducing duplication** in data collection and data quality control
- Looking **beyond immediate use** and examining requirements of users
- Ensuring that **good documentation** and metadata procedures

# Issues influencing data quality

- Accuracy and precision
- Completeness
- Currency and Timeliness
- Update frequency
- Consistency
- Flexibility
- Transparency
- Performance measures and targets
- Data cleaning
- Outliers
- setting targets for improvement
- Truth in labelling
- Error and bias
- Uncertainty
- Auditability
- Edit Controls
- Minimise duplication and reworking of data
- Maintenance of original (or verbatim) data
- Categorisation can lead to loss of data and quality
- Documentation
- Feedback
- Education and Training
- Accountability

- **Collectors**
  - **Custodian or Curator**
  - **Aggregator**
  - **Publisher**
  - **Users**
-

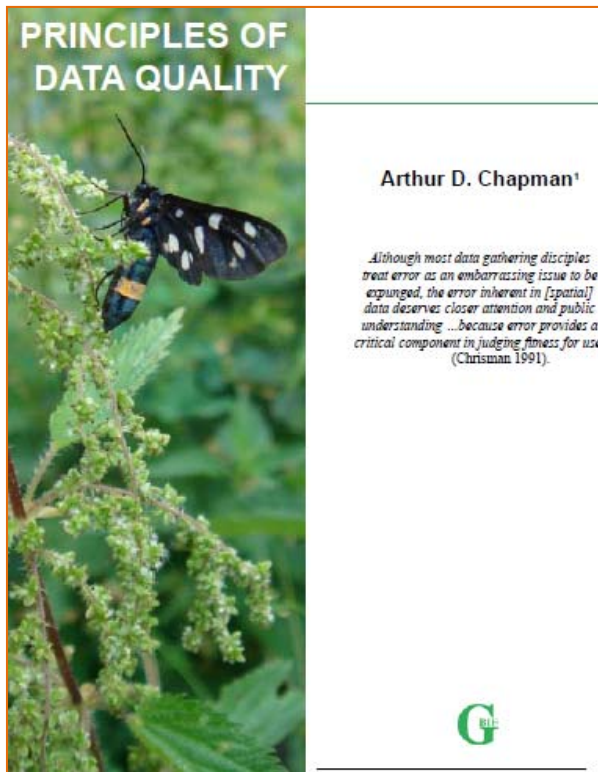
**A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions**

## **General framework for data cleaning**

- **Define and determine error types**
  - **Search and identify error instances**
  - **Correct the errors**
  - **Document error instances and error types; and**
  - **Modify data entry procedures to reduce future errors**
-



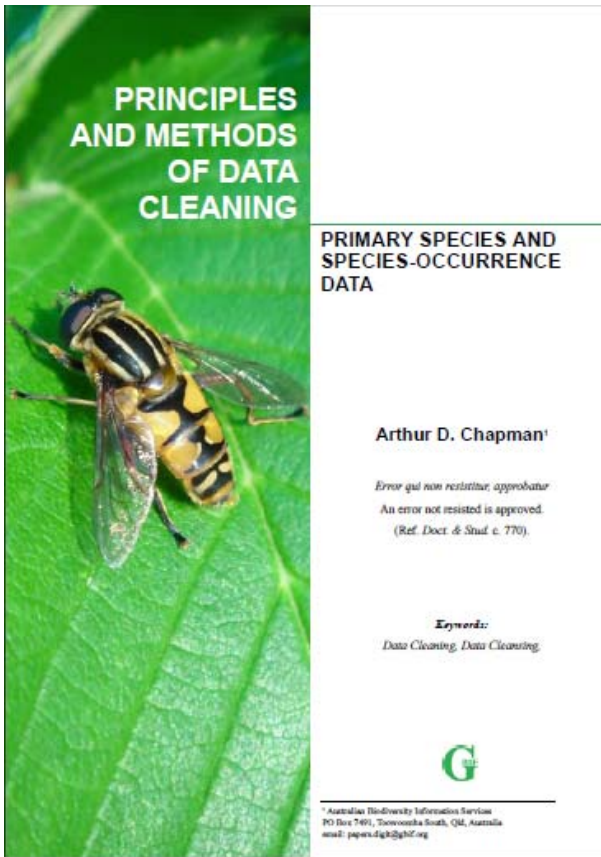
- **Accuracy**
- **Effectiveness**
- **Efficiency**
- **Reliability**
- **Accessibility**
- **Transparency**
- **Timeliness**
- **Relevance**



## Principles of Data Quality

The rapid increase in the exchange and availability of taxonomic and species-occurrence data has made data quality principles important, as users of the data begin to require more and more detail on the quality of this information.

<http://www2.gbif.org/DataQuality.pdf>



## Principles and Methods of Data Cleaning

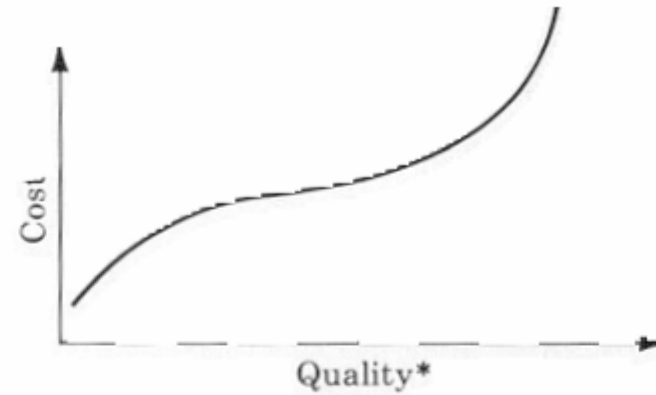
Error prevention is far superior to error detection and cleaning, but no matter how efficient the process of data entry, errors will still occur. Therefore, data validation and correction cannot be ignored, especially when dealing with legacy biodiversity data and this manual helps to correctly face these issues.

<http://www2.gbif.org/DataCleaning.pdf>

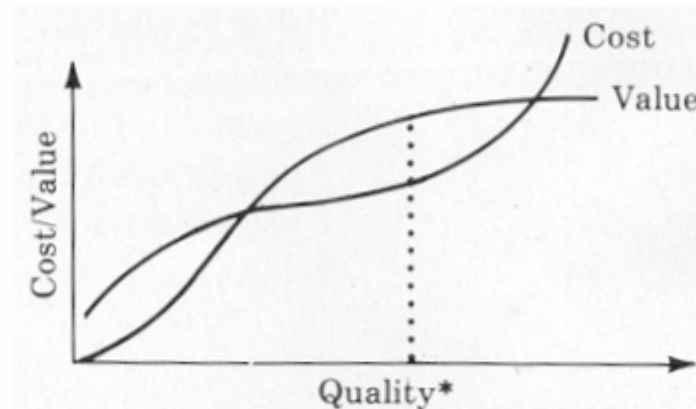
# Keep approaches practical



(a) Information Quality Versus Value



(b) Information Quality Versus Cost

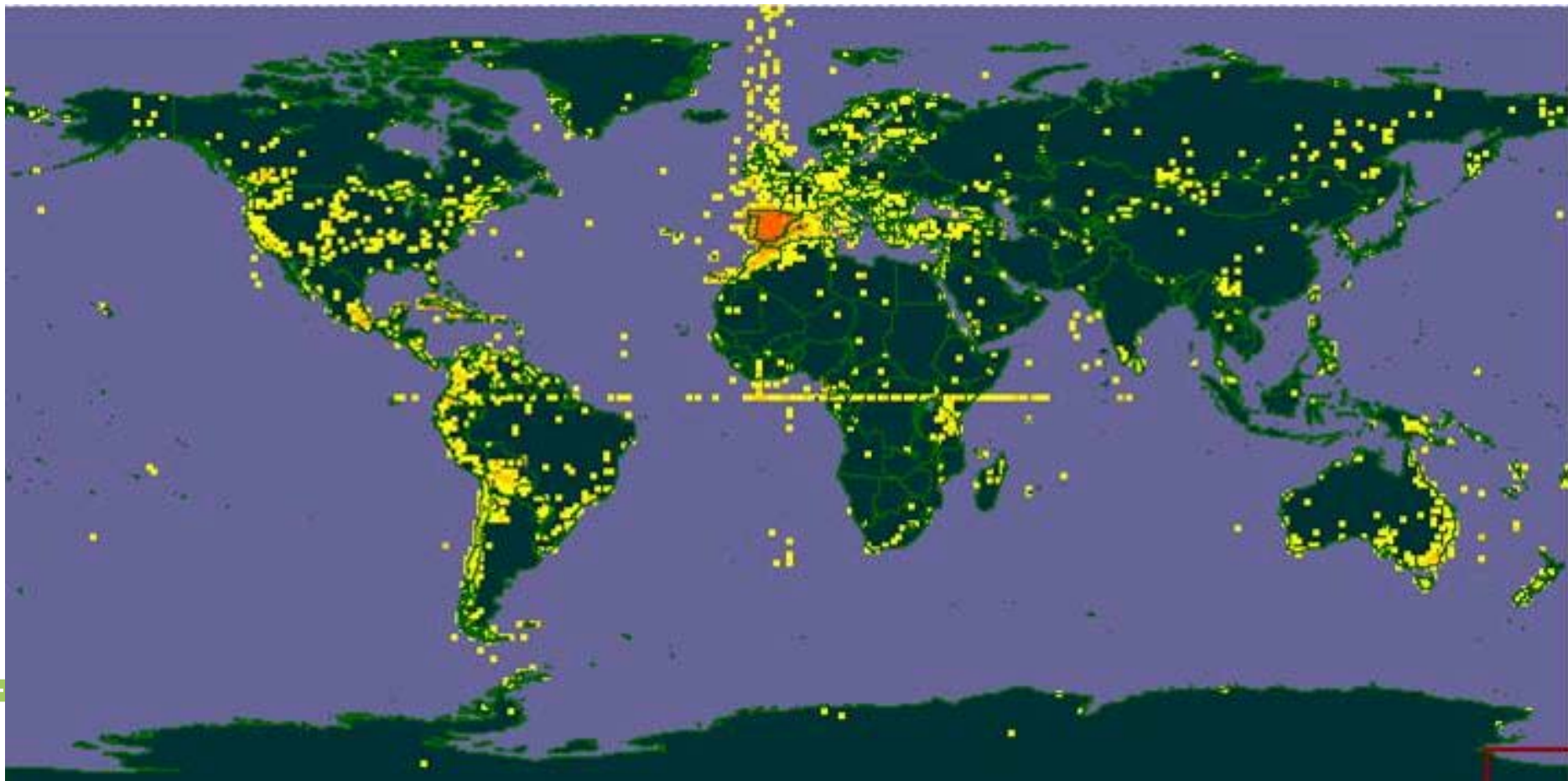


(c) Optimal Information Set

\*Information quality =  $F(\text{detail, age, accuracy, relevance})$

# Part 2: Situation in GBIF Spain

- 110 out of 122 datasets using “hosting service”
- DIGIR crashes
- Room for improvement



# We put in place some corrective measures

- Training
- Validation tools

# Training

3rd Workshop on Data Quality in Biodiversity Databases - Oct. 2009

Status: **open**

Venue and date: **Last one Feb. 2011 (the 5<sup>th</sup>) and delivered online using eLearning Platform**

Description: his databases: both those working in natural history collections and those working with other biodiversity-related databases. The principles of data quality will be revised together with methods to increase this quality in our databases.

More information (usually in Spanish)...

## PRINCIPLES OF DATA QUALITY



Arthur D. Chapman<sup>1</sup>

*Although most data gathering disciplines treat error as an embarrassing issue to be expunged, the error inherent in [spatial] data deserves closer attention and public understanding ...because error provides a critical component in judging fitness for use.*  
(Chrisman 1991).



11:30 - 12:45 - Principios de calidad de

- Principios generales <<VIDEO>>
- Precisión y exactitud <<VIDEO>>
- Calidad en todo el proceso
  - Captura
  - Almacenamiento

12:45 - 13:30 - Herramientas y proced

- Uso de tesauros y otros. María E

13:30 - 15:00 - Comida.

15:00 - 15:30 - Detectar y corregir. Fra

16:00 - 16:45 - Principios sobre Calidad

- Que (taxonomía/nomenclatura)
- Dónde y Cuándo. Isabel Ortega.

16:45 - 17:15 - Pausa para café.

17:15 - 18:00 - Calidad :

- Quién. Francisco Pando. <<VIDEO>>.
- Qué (descriptivo). Francisco Pando. <<



# Tool

## Darwin Test

DARWIN TEST is a software application to validate and check *DarwinCorev2* or *Darwincore1.4* records (*DarwinCore* version 1.2 or *Darwincore* version 1.4 standard for specimen and observation data exchange).

Before publishing your biodiversity data in a public network such as GBIF it is highly recommended to test your *DarwinCorev2* or *Darwincore1.4* data using DARWIN TEST program, in order to detect possible problems. The issues analyzed include omission, typographic, convention and coherence errors. DARWIN-TEST is a *Microsoft Access®* based program. At present, the software is available only in Spanish.

Please find further information about DARWIN TEST in its website:



### Two kind of tests

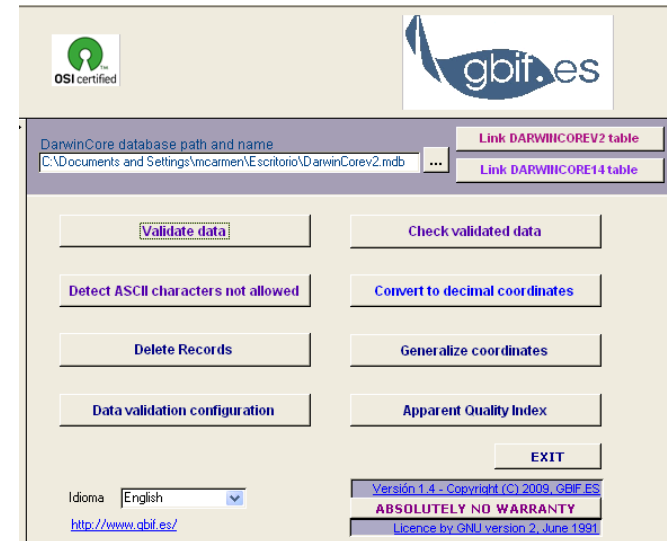
#### Technical

Field names, data types, etc

Ascii characters

#### Content

Congruence test

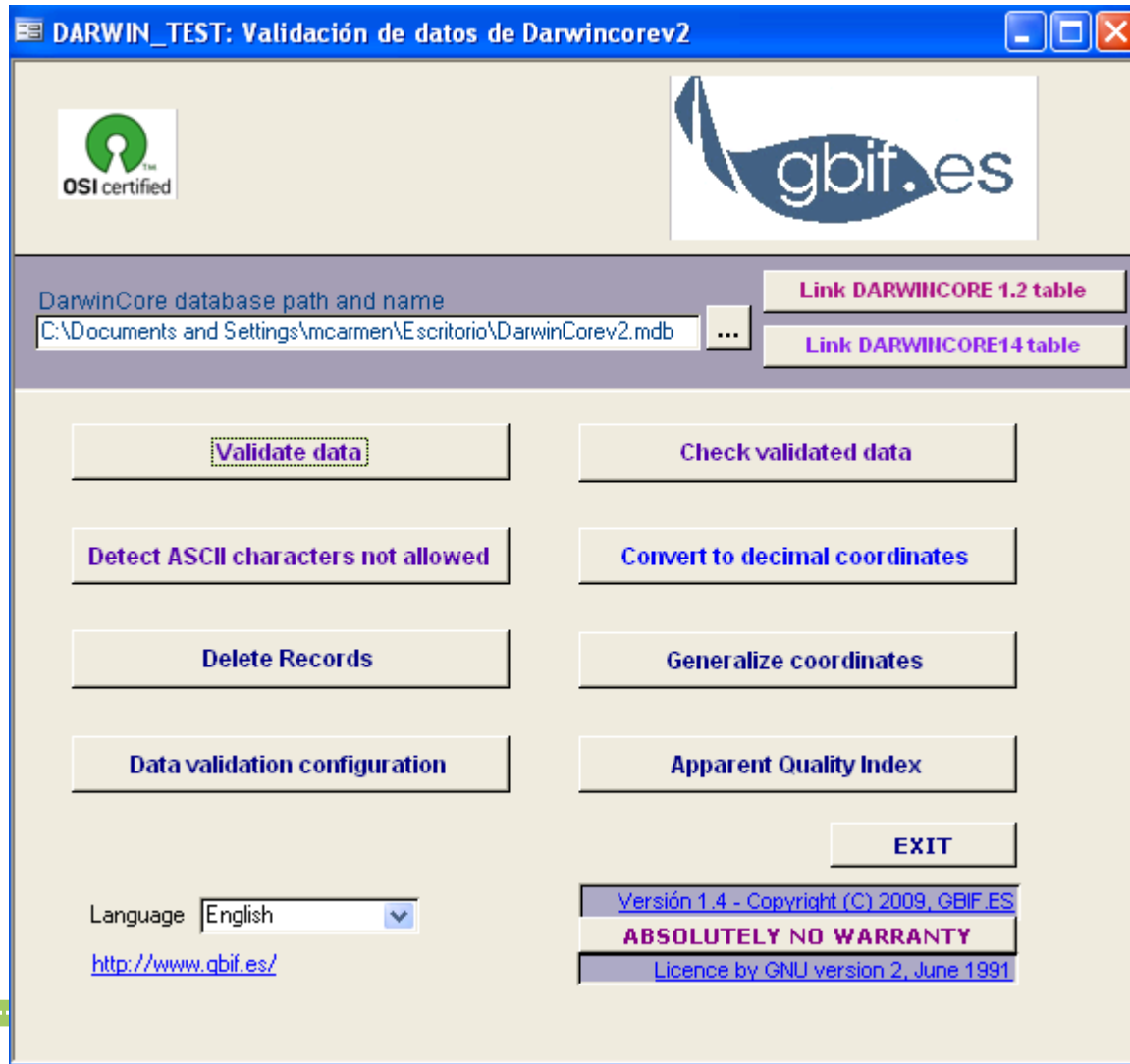




# Tool (Darwin test) deployment

- First stage
  - Tool used at the coordination unit on "ready to publish datasets"
  - Report to data providers
    - Technical part: has to be passed
    - A FYI report (on content)
- 2nd stage
  - Make it publicly available:
    - Offer to all GBIF participants as a services
    - Request to users of the hosting service to past the Darwin Test
- 3rd stage
  - Using CoL as controlled vocabulary for scientific names
  - Using the Colombia'ss AATs (Archivos de autoridad taxonómica; taxonomic reference archives) as dictionaries for scientific names
  - Making Darwin Test multilingual

# Darwin test in action



Home > Software > Darwin\_Test

- Inicio
- GBIF.ORG
- Sobre GBIF
- Actualidad
- Participación
- Consultar Datos**
- Proyectos de Biodiversidad
- Colecciones y Proyectos
- Formación y Divulgación
- Recursos
- Contactar

## Darwin\_Test

An application to test and check data in the Darwincorev2 or Darwincore1.4 f databases quality in the GBIF network.

**NEW! DARWIN\_TEST 1.4**

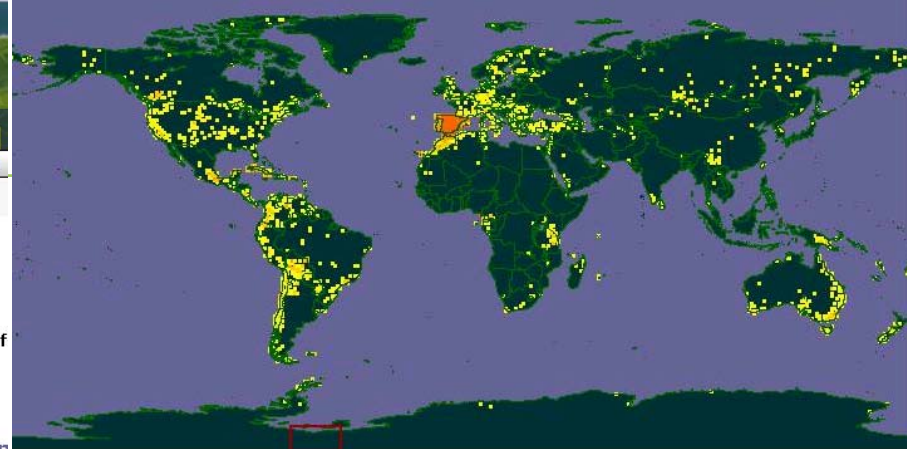
- Characteristics
- Download zone
- User Manual (In Spanish)
- How to give credit when using Darwin Test
- Credits
- Report bugs, propose enhancements

<< Octubre 2009 >>

L	M	M	J	V	S	D
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

**NEW VERSION 1.4:** Darwin\_Test 1.4 preserves all the functionalities from the old version (che better conversion between UTM's or geographical coordinates to decimal coordinates, management Darwincore records, and generalization of decimal coordinates thought user filters) and includes as a display language.

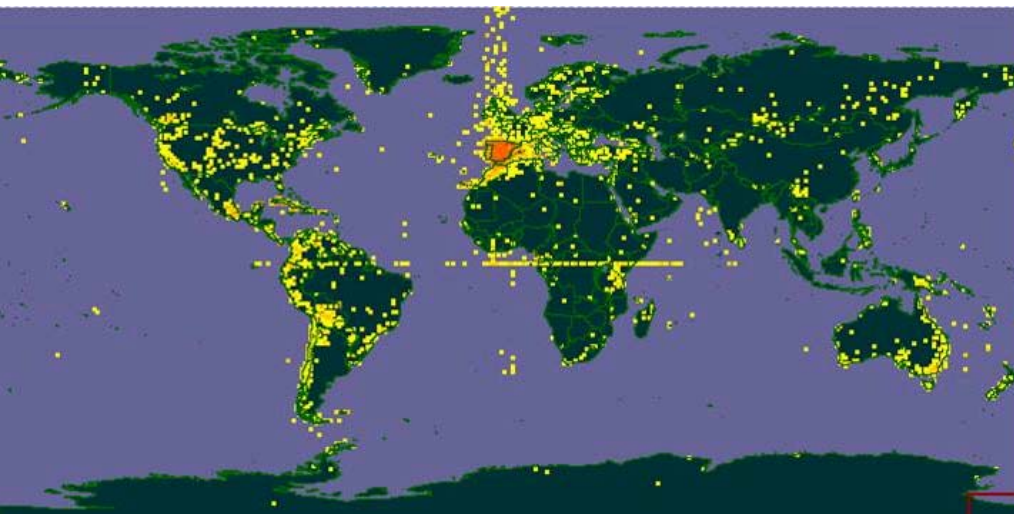
**DARWIN\_TEST** is an application to validate and check records from tables using the Darwincore (standard DarwinCore2 version 1.2 ó Darwincore version 1.4) to the exchange of information between natural history collections.



<http://sourceforge.net/projects/darwin-test>

[http://www.gbif.es/darwin\\_test/Darwin\\_Test\\_in.php](http://www.gbif.es/darwin_test/Darwin_Test_in.php)

[http://www.gbif.es/darwin\\_test/Darwin\\_test.php](http://www.gbif.es/darwin_test/Darwin_test.php)



**sourceforge** FIND AND DEVELOP OPEN SOURCE SOFTWARE

Find Software | Develop | Create Project | Community | Site Support

SourceForge.net > Find Software > Darwin\_Test

**Darwin\_Test** Alpha by ortega-makeda

Summary | Files | Support | Develop

Darwin\_Test validates the Darwincore2 format, the standard for natural history collections databases on the GBIF network.

**Download Now!**  
Darwin\_Test\_13.zip (660.0 KiB)

OR [View all files](#)

<http://darwin-test.sourceforge.net>

TAGS [edit](#)


# Part 3: Hands-on Darwin Test



DARWIN\_TEST: Validación de datos de Darwincorev2

## Darwin Test

Data validation and geographic coordinates generalization for Darwin Core datasets



DarwinCore database path and name  
K:\GBIF.ES\Formacion\2011\2011 Kenya\DarwinTest\_KENIA\Darwin ...

Link DARWINCORE 1.2 table  
Link DARWINCORE 1.4 table

Validate data

Convert to decimal coordinates

DwC data verification

Generalize coordinates

Update linked DwC table

Create a DwC table to correct the original database

Selectively delete records from linked DwC

Detect ASCII characters not allowed


Data validation configuration

Apparent Quality Index

Language

<http://www.gbif.es/>

Versión 2.0 - Copyright (C) 2010. GBIF.ES  
**ABSOLUTELY NO WARRANTY**  
Licence by EUPL version 1.0, January 2007



# At your service:



Francisco [Paco] Pando ([pando@gbif.es](mailto:pando@gbif.es))

Director

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

[pando@gbif.es](mailto:pando@gbif.es)

[www.gbif.es](http://www.gbif.es)

Phone: + 34 91 420 3017

Fax: + 34 91 420 0157



