Datos de investigación: reflexiones sobre su acceso abierto

Luis-Millán González (Universitat de València)
Fernanda Peset (Universitat Politècnica de València)

21 de octubre de 2013, 11.30 h.

¿Es posible una ciencia abierta si no se comparten los datos de investigación? Nuestra intervención reflexiona sobre este interrogante. Investigadores, agencias de financiación, bibliotecas, etc. son los agentes involucrados en este proceso. Cada uno cuenta con unos intereses y unos roles que determinan cómo se está abriendo la ciencia a pares y a la sociedad en su conjunto. Revisaremos también los avances en el acceso a los datos y las herramientas que están disponibles en estos momentos.

OPENDATASCIENCE, centro de recursos para la preservación y gestión de datos abiertos de investigación", CS02012-39632-C02-02

Agradecimientos

Justificación

el acceso a las publicaciones ya está encarrilado, pero ¿y los datos de investigación, el “nuevo” petróleo de la ciencia?

El futuro está demasiado cerca y participamos de él

Estructura

I. Las cuestiones técnicas: definiendo y detallando sus beneficios

II. La realidad. ¿Quiénes son los actores y cuál es su visión?

- la financiación : agencias y empleadores
- la producción de la ciencia: investigador
- la gestión del producto: OTRIS y bibliotecas

III. El presente y el futuro: conclusiones/tendencias/necesidades

IV. Nuestras inquietudes, ¿y las vuestras?

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA
CRUE
REBIUN
Red de Bibliotecas Universitarias
SB SocialBiblio

¿Acceso abierto a los datos de investigación?

I. Veamos qué son datos de investigación y qué es abierto

Datos de investigación cerrados y abiertos

- “datos de investigación [es] todo aquel material que ha sido **registrado durante la investigación**, reconocido por la comunidad científica y que sirve para **certificar** los resultados de la investigación que se realiza. […] debe provenir de una **fuente única** y deben ser **difíciles o imposibles de obtener de nuevo**”

Torres-Salinas; Robinson-García; Cabezas-Clavijo, 2012

Heterogéneos

Los datos incluyen: “cuadernos de laboratorio, cuadernos de campo, datos de investigación primaria (incluidos los datos en papel o en soporte informático), cuestionarios, cintas de audio, videos, desarrollo de modelos, fotografías, películas, y las comprobaciones y las respuestas de la prueba. Las colecciones datos para la investigación pueden incluir diapositivas; diseños y muestras. […] El código de software “

Universidad de Melbourne, del Informe Recolecta

Numerosos

- *A research data set constitutes a systematic, partial representation of the subject being investigated.*

Ergo, ya en la década del 2000 se valoran y comienzan a preservarse

(como las publicaciones)

Beneficios

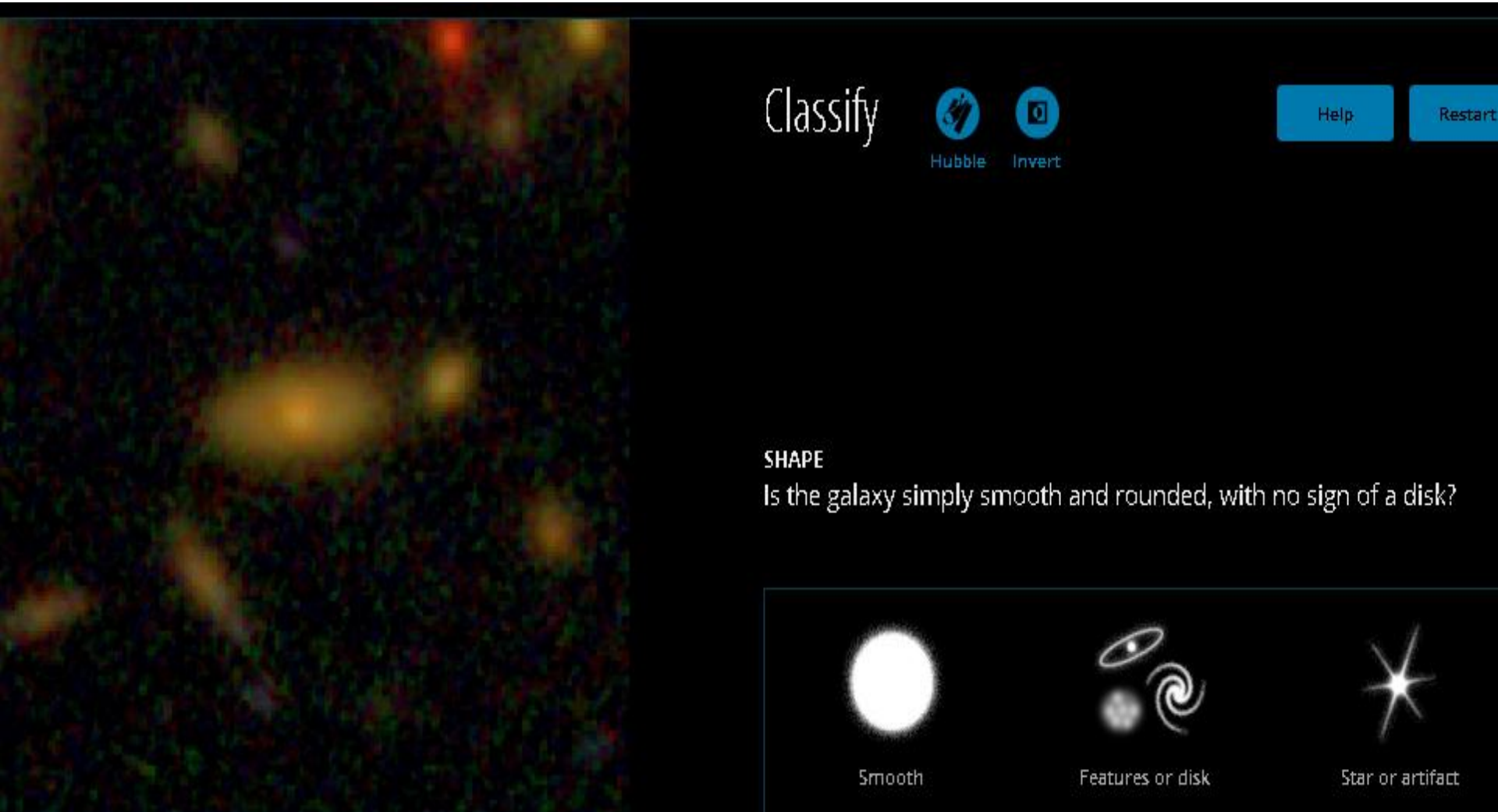
- Convengamos en que los poseen ***per se***

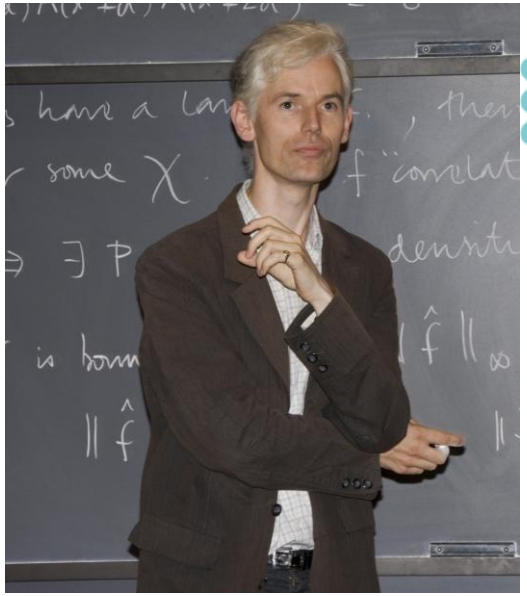
Compartir datos tiene beneficios nuevos

- responder de forma rápida y mucho más eficiente ante las emergencias (e-coli 2011)
- capacidad potencial de señalar los fraudes (datos inventados: reciente polémica del artículo de Science) o las malas prácticas (tensionar la integridad del sistema o de las personas).
- estimula nuevas y altamente creativas formas de colaboración científica
- estimula un movimiento social hacia la ciencia, quizá fundamental para el cambio de las dinámicas científicas: Galaxy Zoo, Fold-it, Ash-Tag, etc.

Boulton, 2013

Galaxy zoo


Classify
Hubble
Invert
Help
Restart
SHAPE
Is the galaxy simply smooth and rounded, with no sign of a disk?
Smooth
Features or disk
Star or artifact



Tim Gowers, matemático, planteó un problema en su blog y fue resuelto por la comunidad —crowdsourcing— de manera inauditamente veloz.

OPEN International ACCESS WEEK

Pero además…

- los **ciudadanos**, quienes pagan con sus impuestos la ciencia, demandan cada vez más las evidencias que dan lugar a las políticas que se les acaban imponiendo en todos los órdenes sociales
- única forma de afrontar **retos globales** que implican a toda la sociedad: cambio climático, crisis energética, las pandemias o residuos mundiales.

¿Qué significa comunicar los datos de forma abierta?

Lo libre, lo open… recordemos

- *Gratuito.* Sólo implica el acceso sin coste, pero no otras dimensiones de reutilización y difusión. “Free” tiene esa acepción.
- *Libre.* A menudo se mantiene en su forma española para diferenciarlo de lo simplemente gratis. En realidad este es el término que acoge otras connotaciones con la reutilización del contenido sin restricciones.
- *Abierto*. Desde su semántica inicial, de código accesible y manipulable, incluye actualmente la posibilidad técnica y legal de transformarse.

Veamos los Principios Panton

- "En el ámbito de la ciencia, por **datos abiertos** nos referimos a su disponibilidad gratuita en Internet permitiendo a cualquier usuario **descargarlos, copiarlos, analizarlos, volverlos a procesar, transferirlos a un software, o utilizarlos** para cualquier otro propósito, sin barreras económicas, legales o técnicas fuera de aquellas que son inseparables del acceso a Internet.

- En el momento de publicar los datos o colecciones de datos es fundamental que dicha publicación esté acompañada de una **declaración clara y explícita** de los deseos y expectativas de quienes los publican con respecto a la reutilización, y propósitos de uso de los elementos individuales de los datos, de la colección completa así como de subconjuntos dentro de la colección.

- son apropiadas las licencias del 'Creative Commons' (CCZero) y Licencia al Dominio Público (Public Domain Dedication & Licence PDDL-OKF). Se desaconseja ENCARECIDAMENTE el uso de licencias que limitan la reutilización comercial o la producción de obras derivados

Open data de Royal Society

- Criterios a cumplir cuando son liberados:
  - Accesibles (accessible): fáciles de encontrar y en una forma en que puedan ser usados [y preservados].
  - Evaluados/certificados (asessable): credibilidad para diferentes grupos de interés.
  - Inteligibles (intelligible): ser entendidos
  - Reutilizable (useable): en un formato para usar y con licencias adecuadas.

include the JPEG 2000, PNG and SVG standard image formats; ASCII, PDF, Open Document Format and Office Open XML format (the native format for recent versions of Microsoft Word) for text; HTML, XHTML, RSS and CSS for the web and NetCDF for some scientific data.

Resumen

- Heterogéneos y variables
- Cerrados y abiertos
- Criterios de liberación bastante costosos

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA
CRUE
REBIUN
Red de Bibliotecas Universitarias
SB SocialBiblio

¿Por qué ahora?

II. La realidad. ¿Quiénes son los actores y cuál es su visión?

!!!Potenciar la **innovación** (buzzword en H2020)!!!

- Uso intensivo de herramientas tecnológicas y de comunicación: e-ciencia
- Producción de datos en aumento y por ende de herramientas de análisis de éstos: data y text mining sobre big data y linked data…

Producir nuevo conocimiento a partir de datos previos

- Remezclar
- Mostrar patrones ocultos con las antiguas técnicas
- Inventar/intentar nuevas aproximaciones
- Visualizaciones…

A de Agencias de financiación y empleadores (públicos)

- Aprovechemos el potencial desaprovechado de los datos aislados (sin compartir)
- Los datos abiertos pueden ser el combustible de la **innovación, crecimiento y creación de trabajo** (Neelie Kroes, com. Agenda Digital)
- Plan de gestión de datos en las solicitudes

C de Científicos
Pero… ¿qué piensa un científico?

- Producción
- Explotación
- Compartido/intercambio
- Reconocimiento
- Preservación
- Más trabajo

S de Servicios de apoyo (bibliotecas, OTRI)

- Entender el trabajo de los científicos!
- Colaborar
- Prepararnos y conocer
- Asesorar
- Apoyar

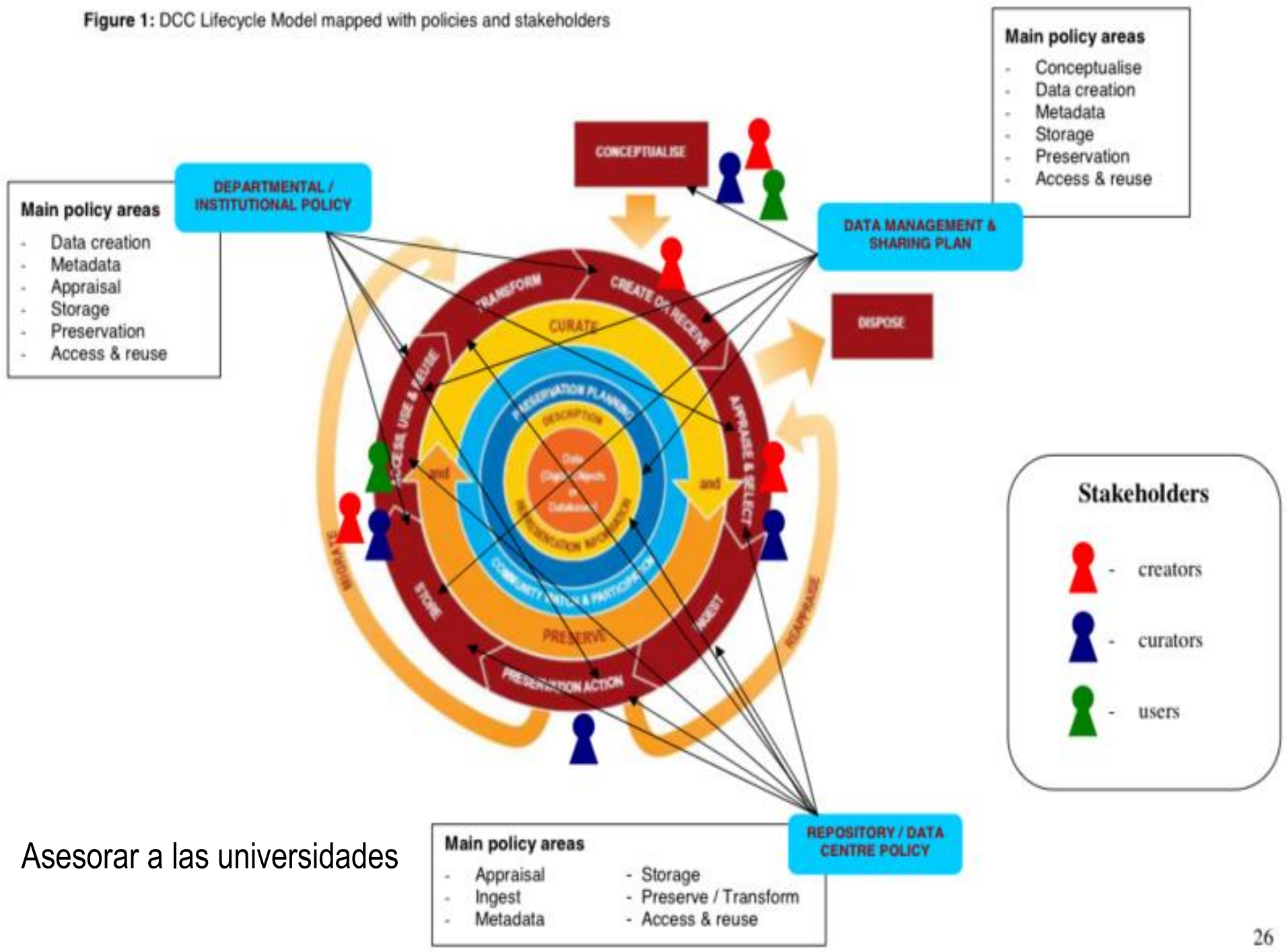
- **En los inicios:** planes de gestión de las solicitudes de financiación: costes de las actividades de gestión y herramientas online;
- **Durante el proyecto:** la documentación, los formatos y los estándares sobre reutilización de los datos; y sobre almacenamiento, gestión y análisis de los datos de acuerdo con prácticas ya probadas (+fuentes de datos)
- **Una vez terminado:** qué datos tendrán valor en el futuro; ayuda para visibilizar y hacer disponibles los datos para varios tipos de grupos de interés

International OPEN ACCESS WEEK

Aprendizajes en

- Plan de gestión de datos
- Organizar y documentar los datos
- Almacenar los datos y seguridad
- Aspectos éticos y de propiedad intelectual
- Compartir datos, preservación y licencias

Figure 1: DCC Lifecycle Model mapped with policies and stakeholders



Asesorar a las universidades

Liberar datos no es fácil

- Características de la producción condiciona los **metadatos** de preservación
- **Herramientas** heterogéneas (otras generalistas como Zenodo)
- **Licencias** y confidencialidad
- Pero sobre todo… **time-consuming**!

DRYAD

About | For researchers | For organizations | Contact us | Profil

Describe publication

Submitting data to Dryad consists of three simple steps:

- **1. Describe your publication**
- 2. Upload and describe your data files
- 3. Approve data for publication

Please describe your publication in as much detail as possible. Providing a detailed description will make it easier for others to fi
Dryad. Please describe the **publication only**. Do not enter information specific to your data files on this page.

Fields marked with an asterisk (*) are required. For more information on expected contents for a field, hold your mouse over the
question.

Publication metadata

Title*

Authors*

Last name, *e.g. Smith* | First name + initial, *e.g. Donald F.*

Add

Please include the entire first name, ar
initial if appropriate. Do not use only in
this makes the author more difficult to
later.

Journal name*

Abstract

DOI

If your publication has not been assign
please leave this blank.

Journal issue

Volume | Number

Year

If your publication has not been assign
volume or issue number, please leave

Primary contact for data associated with this article

Subject keywords

Add

Please enter general keywords associ
the data file. Keywords may be separat
commas, or added individually. For exa
adaptation
evolutionary contingency
founder effects

Taxonomic names

Add

Please enter taxonomic names associa
the publication. These may be names a
(species or higher-level clade). Names
separated by commas, or added indivi
example:
Drosophila melanogaster

(LastName, FirstName)

Producer *
Producer | Abbreviation
Affiliation
URL
(http://example-domain.edu/)
Logo URL
(http://example-domain.edu/image)

Production Date *
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Funding Agency *

Distributor *
Distributor: Harvard Dataverse Network | Abbreviation
Affiliation
URL: http://thedata.harvard.edu/dvn/
(http://example-domain.edu/)
Logo URL: http://thedata.harvard.edu/dvn/resources/images
(http://example-domain.edu/image)

Contact *
Contact | Affiliation
E-mail

Distribution Date *
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Deposit Date *
2013-07-05
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Description and Scope

Description *
Copying and pasting from a Word document can create errors when you save this page.
Description
Description Date
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Keyword *
Keyword | Vocabulary
URL
(http://example-domain.edu/)

Topic Classification *
Topic Classification | Vocabulary
URL
(http://example-domain.edu/)

Time Period Covered - Start *
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Time Period Covered - End *
(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

Geographic Coverage *

Resumen

- Costoso
- Nuevo para las bibliotecas
- Capacidades nuevas
- Equipos mixtos como técnicos de laboratorio
- Bibliotecario especializado

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA
CRUE
REBIUN
Red de Bibliotecas Universitarias
SB SocialBiblio

¿ tan incierto es el futuro?

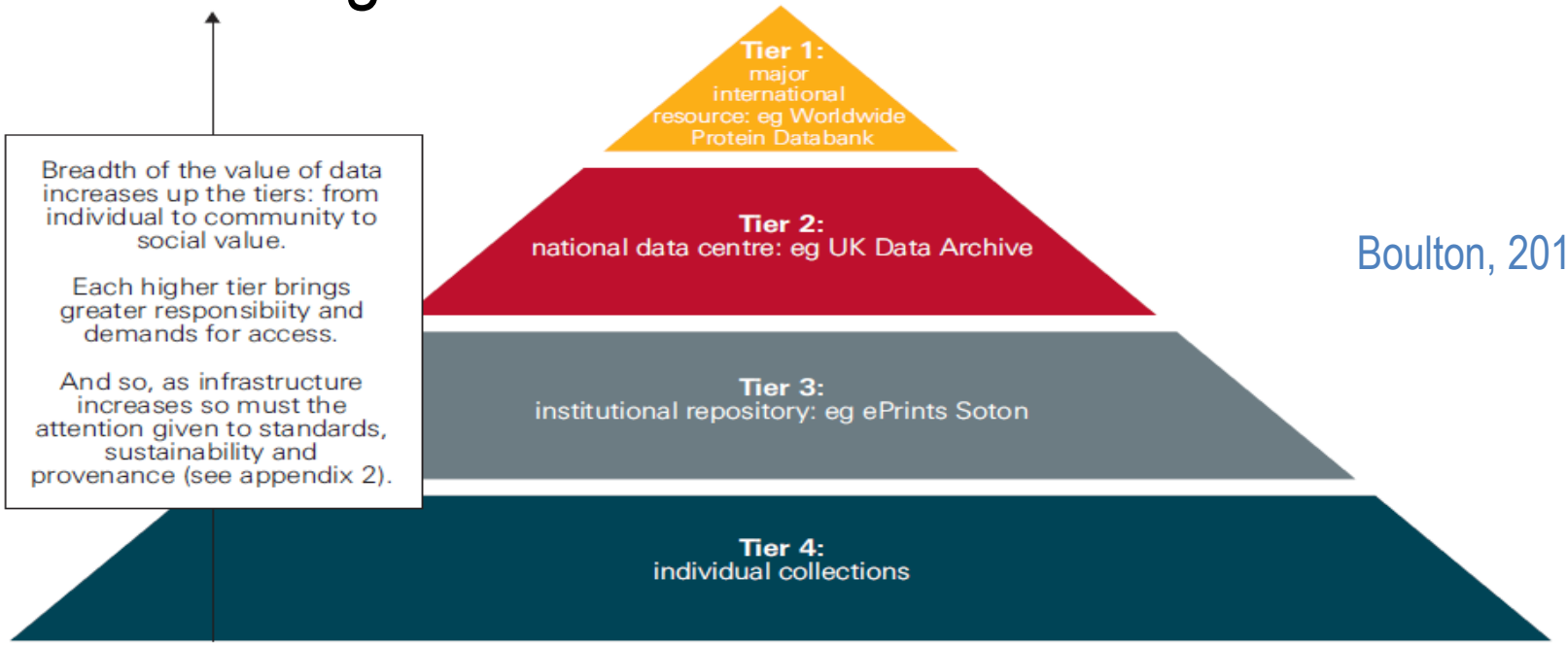
III. El presente y el futuro: conclusiones/tendencias/necesidades

Presente

- España: RECOLECTA-FECYT *Informe preliminar para la conservación y reutilización de los datos científicos en España*
  - CSIC; CEACS; UPF

Herramientas: especializadas y huérfanos

- Figshare, Dryad o Zenodo (Pedro Príncipe)
- Pisos desiguales:



Boulton, 2013

- Y de otros países, lo que podemos aprender y adaptar
  - DCC
  - RDA
  - ANDS
  - …

¿Qué podemos hacer?

- Gestionar desde el inicio /// Pensar en clave de datos
- Participación en foros públicos como RDA o DCC
- Preparar guías y difundir nuestros servicios

Plan was successfully created.

Project Details

PROJECT	SEDIC
PLAN CREATED	28 August 2013, 11:42
BUDGET	€16,000.00
START DATE	28 August 2013
END DATE	28 December 2013
LEAD ORGANISATION	UPV
OTHER ORGANISATIONS	UV

Edit Project Details

Project Sharing

This plan is not shared

Share Plan

My Plans | Fill in Plan | Export | Lock | Duplicate | Delete

Project Phases

Generic DMP Project funding

Application
[Initial funder form in DMP system.]
Not Started

→

Funded
[Initial funder form in DMP system.]
Not Started

Pensar en clave de datos desde posgrado

- Definir los datos
  - ¿Cómo los obtienes e instrumentos?
  - ¿Cuánto se actualizan?
  - ¿Cuántos generas y en qué formatos?
  - ¿Cuántas versiones almacenas?
- Controlar
  - ¿Cuánta información adicional es necesaria para entenderlos?
  - ¿Dónde los almacenas?
  - Directorios y nombres de archivos
  - Copias de seguridad ¿cómo y cuándo? ¿testeas?
- Compartir
  - ¿De quién es la propiedad?
  - ¿Quién puede usarlos y quién podría?
  - ¿Qué compartes y qué no? ¿por?
- Archivo de datos
  - ¿Qué debe ser archivado?
  - ¿Por cuánto tiempo y dónde?
  - ¿Cuándo pasan al estado “archivo”?
  - ¿Quién es el responsable de moverlos?
  - ¿Quién tendrá acceso? Condiciones
- Supervisión del plan
  - ¿Quién es responsable?
  - ¿Con qué frecuencia se actualizará?
  - …

Univ. Bath

International OPEN ACCESS WEEK

Necesidades

- Reconocimiento: ¿dónde está la E de Evaluadores)
  - Proyecto DATACITE

Grande T, Borden WC, Smith WL (2013) Data from: Limits and relationships of Paracanthopterygii: a molecular framework for evaluating past morphological hypotheses. Dryad Digital Repository. doi:10.5061/dryad.k4m8t

- Integración
  - Proyecto ODIN con DataCite
  - CRIS/Repositorios institucionales (Pablo de Castro)

¿cambio de paradigma?

- entender cómo se están manejando los datos en la institución: **auditarlas**
- construir un ejemplo y apoyarlo: **trabajar desde la práctica**
- definir la posición de la institución: **normativa y estrategia**
- asegurarse que los investigadores conozcan los servicios de apoyo: **guías**
- proporcionar un servicio de **almacenamiento** de datos fácil y robusto
- hacer **descubribles y citables** los datos a otros
- **ir a la cabeza** del movimiento creando servicios de gestión

Hodson y Jones (2013)

Saber más…

- Proyecto I+D+i: DATASEA Datos Abiertos de investigación
- Curso SEDIC: Datos de investigación, para marzo 2014
- Conferencia Javier Hernández: 22 octubre en UPV
  Horizonte 2020. Del acceso abierto a los datos abiertos

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA CRUE REBIUN Red de Bibliotecas Universitarias SB SocialBiblio

IV. Nuestras inquietudes, ¿y las vuestras?

Gracias por vuestra atención

OPEN ACCESS International WEEK