DATA SHARING

Acceso abierto a los datos de investigación: definición, políticas y actores

---

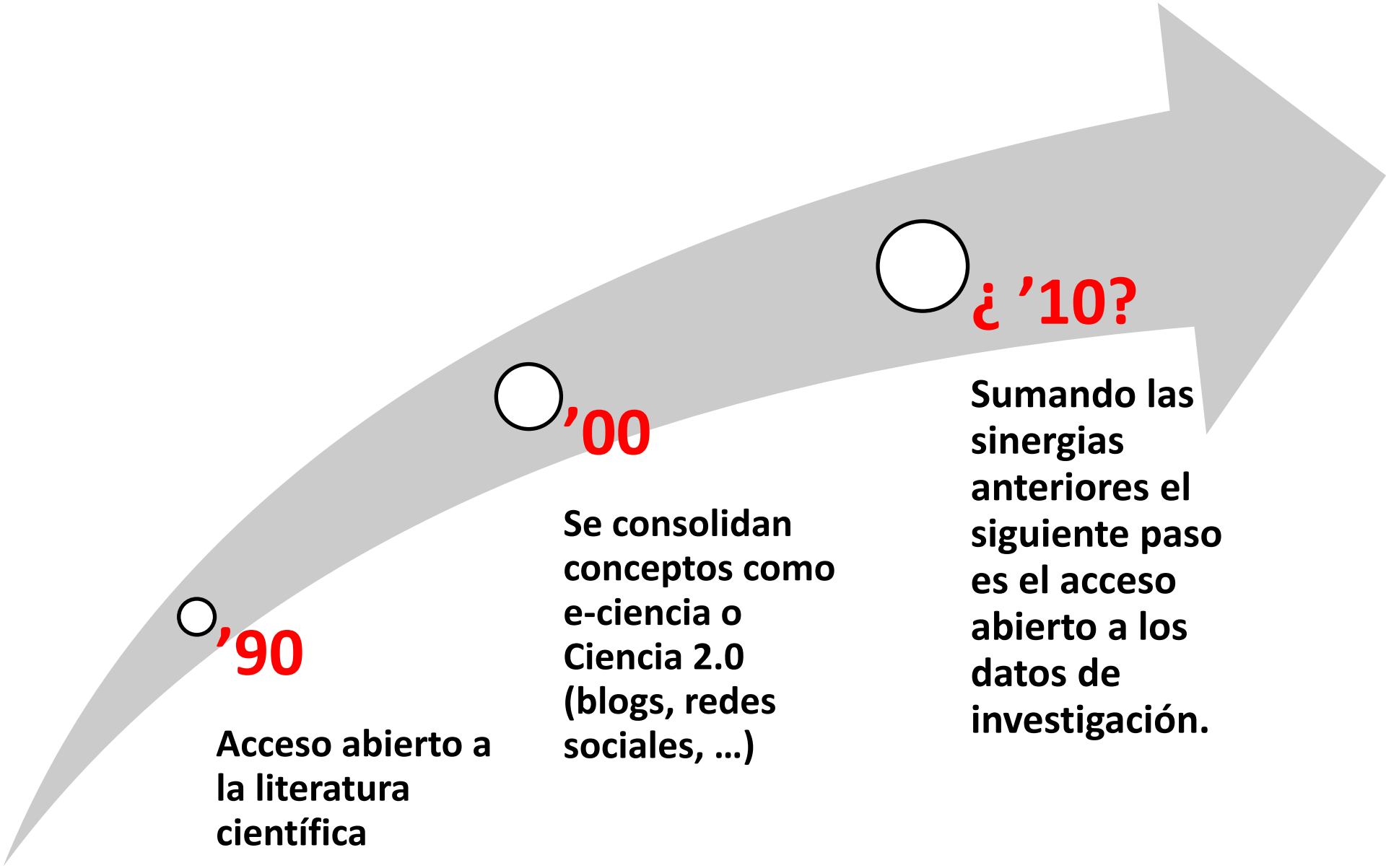
daniel torres-salinas

Instituto Politécnico de Braganca

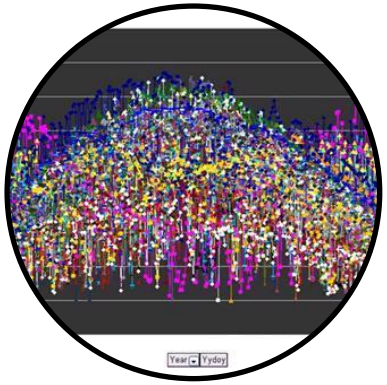
26 de Octubre 2012

INTRODUCCIÓN

DEL ACCESO ABIERTO DE LOS ARTÍCULO AL ACCESO ABIERTO DE DATOS



¿QUÉ ESTÁ FAVORECIENDO EL DATA SHARING?



Como pilar de la E-Ciencia.

Apoyo de todo tipo organismos.

Apoyo de las principales revistas.

Contexto Tecnológico

EXISTE UN INTENSO DEBATE ENTRE LOS PROPIOS CIENTÍFICOS

A raíz de éxitos como el *Human Genome Project* muchas comunidades científicas debate sobre adoptar políticas de *Data Sharing* más intensas

global burden of non-communicable diseases, but also shows the impressive progress we have made over the previous decade. We must continue to

inactivity, and obesi
published online No
8 The NCD Alliance. Jo
www.ncdalliance.org

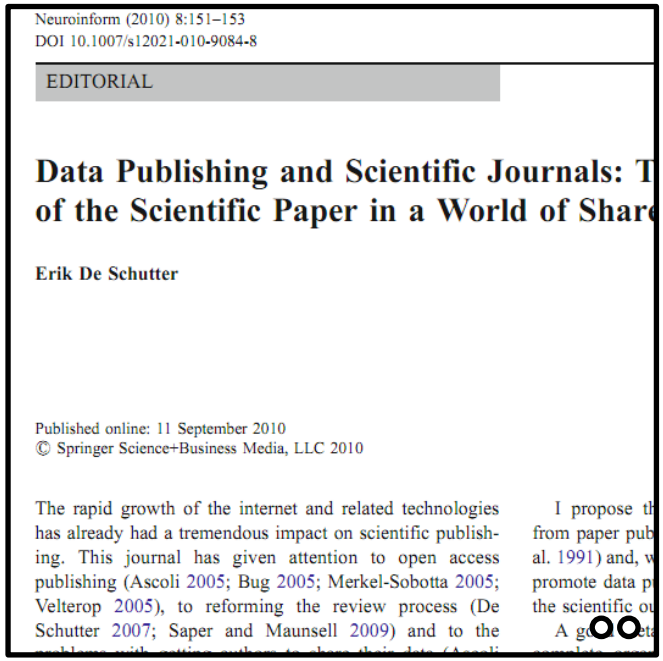
Sharing research data to improve public h

The purpose of medical research is to analyse and understand health and disease. A key and expensive element is the study of populations to explore how interactions between behaviour and environment, in the context of genetic diversity, determine causation and variation in health and disease. As funders of public health research, we need to ensure that research outputs are used to maximise knowledge and potential health benefits. In turn, the populations who participate in research, and the taxpayers who foot the bill, have the right to expect

that every last ou
the research.

Ensuring data
research commun
and enhances the
In many research
biology to the so
ingrained in ho
and genomics, t
populations has

www.thelancet.com Vol 377 February 12, 2011



Neuroinform (2010) 8:151–153
DOI 10.1007/s12021-010-9084-8

EDITORIAL

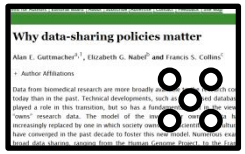
**Data Publishing and Scientific Journals: T
of the Scientific Paper in a World of Share**

Erik De Schutter

Published online: 11 September 2010
© Springer Science+Business Media, LLC 2010

The rapid growth of the internet and related technologies has already had a tremendous impact on scientific publishing. This journal has given attention to open access publishing (Ascoli 2005; Bug 2005; Merkel-Sobotta 2005; Velterop 2005), to reforming the review process (De Schutter 2007; Saper and Maunsell 2009) and to the

I propose th
from paper pub
al. 1991) and, w
promote data p
the scientific ou
A g



Why data-sharing policies matter

Alan F. Guttmacher[a,1], Elizabeth G. Nabel[b] and Francis S. Collins[c]

+ Author Affiliations

dical Image and Data S
e We There Yet?[1]
m E. Flanders, MD

tomorrow belongs to the
day.

ata sharing: reaching con

mmy Whitworth [b]

Wellcome Trust, Gibbs Building, 215 Eust
gland.

rrespondence to Jimmy Whitworth (e-mail

lletin of the World Health Organization 2
2471/BLT.10.079202

ani & AbouZahr write passionately about t

Lancet Neuroinformatics Publich Health Radiographics PNAS

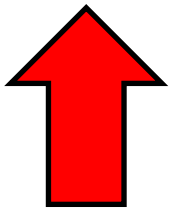
DATOS: TIPOS Y DEFINICIÓN DEL NIH.

SEGÚN FORMATO

- Texto
- Imágenes
- Etc..

SEGÚN FASE INVESTIGACIÓN

- Prelimares
- Finales



SEGÚN OBJETIVO

- Específicos
- Medio
- Generales

"Por datos finales de investigación entendemos material factual registrado, aceptado por la comunidad científica y necesarios para validar los resultados de la investigación. No son datos finales : notas de laboratorio, sets de datos parciales, análisis preliminares, borradores de trabajos, planes para investigaciones futuras, informes que, comunicaciones con colegas, u objetos físicos"

EJEMPLOS DE DATOS DE INVESTIGACIÓN

Bibliometría

Title: Anhydrous polyproline helices and globules
Author(s): Counterman AE, Clemmer DE
Source: JOURNAL OF PHYSICAL CHEMISTRY B 108 (15): 4885-4898 APR 15 2004
Document Type: Article
Language: English
Cited References: 51 Times Cited: 0 FIND RELATED RECORDS

Abstract: ion mobility/time-of-flight methods and molecular modeling calculations have been used to examine the conformations of a range of polymer lengths and charge states of polyproline peptides, [Pro(n) + zH](z+) (n = 3-56; z = 1-6). Ions formed from 1-propanol solutions {[Pro(n) + H](+) (n = 5-11) and [Pro(n) + 2H](2+) (n = 10-22)} favor extended forms of the classical polyproline I helix. In these conformers, all proline residues are in the cis configuration, and protonation at the N-termimls allows hydrogen bonds to be formed with backbone carbonyl groups of the second and third proline residues in each polymer. Protonation of this all-cis form at the N-terminus also stabilizes the helix macrodipole. Singly charged ions formed from aqueous solutions favor globular and hairpin-like conformers that contain both cis- and trans-proline residues. Higher char-e state ions (z = 3-6) formed from aqueous Solutions favor relatively extended conformations, although these are not as extended as the polyproline II structural limit. As polymer size increases, higher charge state ions become more compact. Several conformer states of varying size that appear to be favored structural types are observed; however, we have not been able to identify the type of structures based on comparison of molecular modeling data and experimental measurements.

KeyWords Plus: ION MOBILITY MEASUREMENTS; POLY-L-PROLINE; TANDEM MASS-SPECTROMETRY; GAS-PHASE; UNSOLVATED PEPTIDES; ELECTROSPRAY-IONIZATION; HIGH-RESOLUTION; CONFORMATIONS; BIOMOLECULES; DISSOCIATION

Addresses: Clemmer DE (reprint author), Indiana Univ, Dept Chem, Bloomington, IN 47405 USA Indiana Univ, Dept Chem, Bloomington, IN 47405 USA

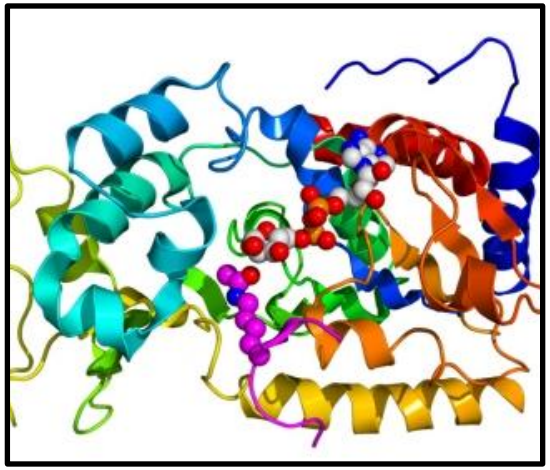
Publisher: AMER CHEMICAL SOC, 1155 16TH ST, NW, WASHINGTON, DC 20036 USA

Subject Category: CHEMISTRY, PHYSICAL

IDS Number: 810ST

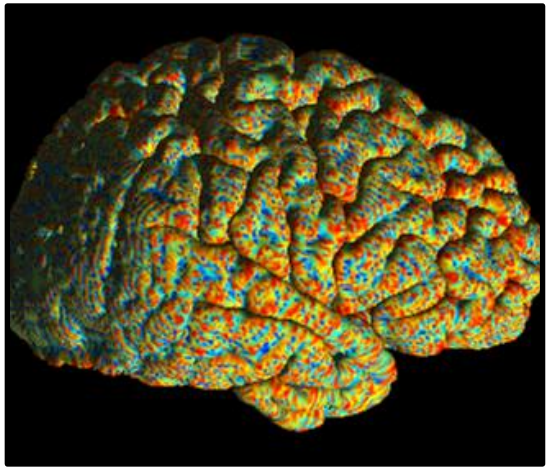
Los registros descargados en una base de datos para un estudio.

Proteómica



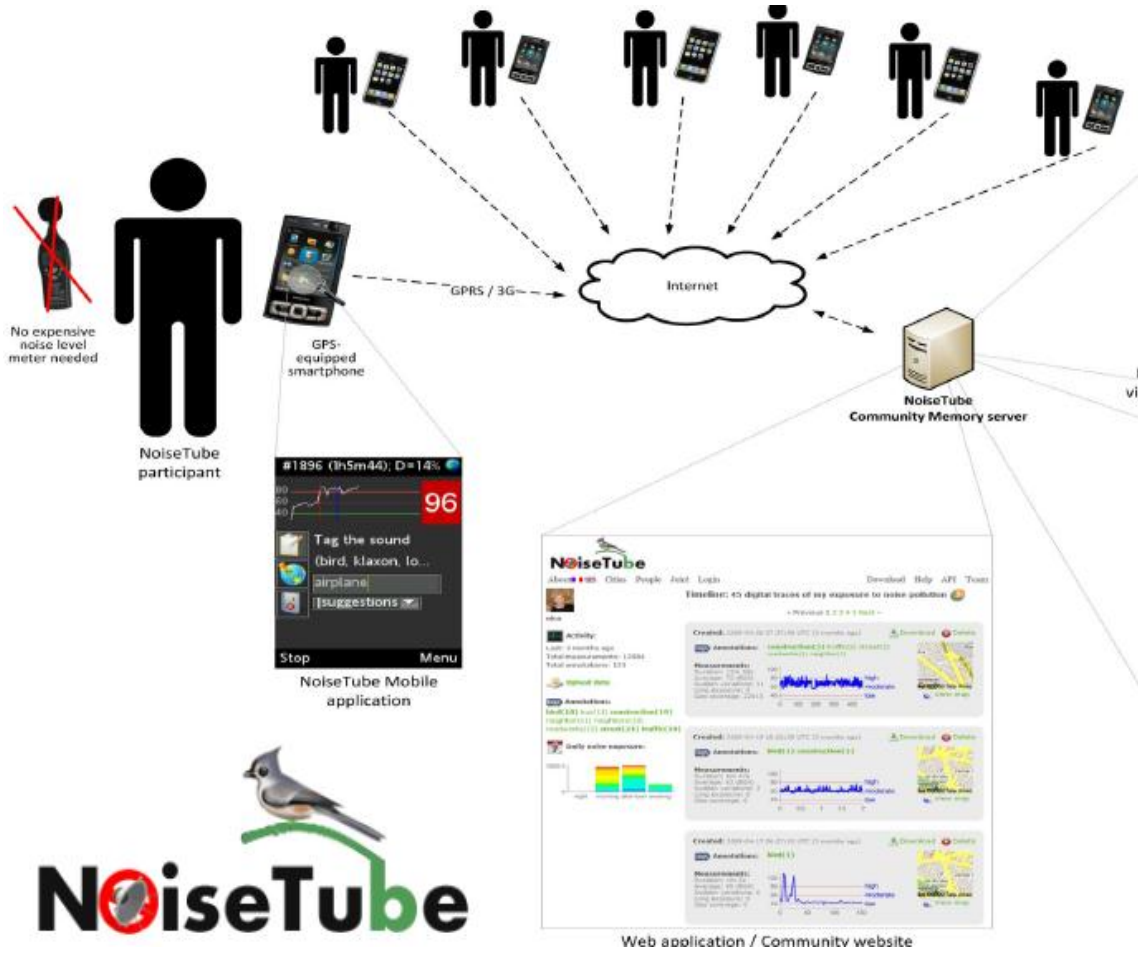
Serían todas aquellas proteínas y péptidos que se han conseguido identificar

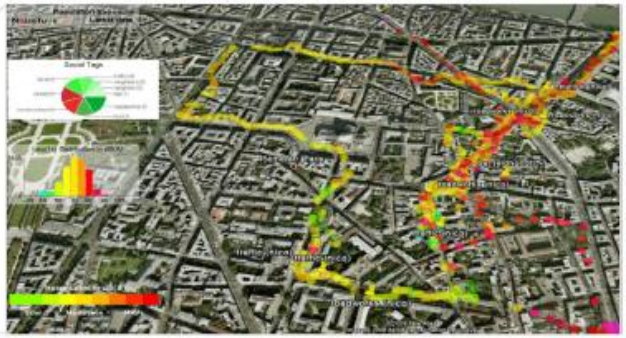
Neurociencias



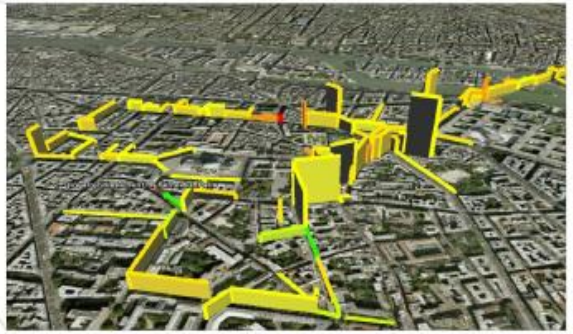
En neurociencias las neuroimágenes

EJEMPLOS DE DATOS DE INVESTIGACIÓN + SOCIAL





Innovative visualisations



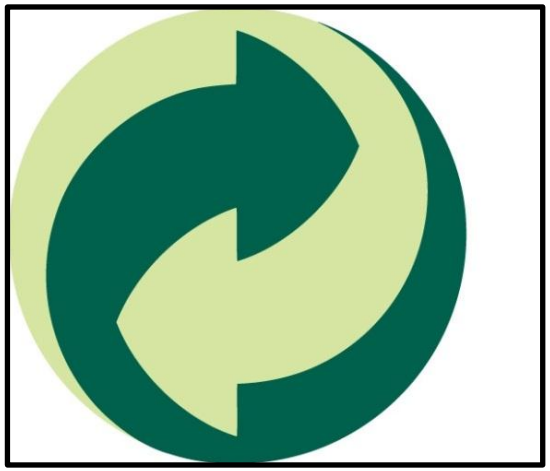
NoiseTube

Sony CSL

Sony Computer Science Laboratory Paris

¿PERO CUÁLES SON LAS VENTAJAS DE COMPARTIRLOS?

Reaprovechamiento



- Mejor aprovechamiento dinero publico.
- Permite acabar con proyectos duplicados.
- Permite nuevos estudios.

Transparencia

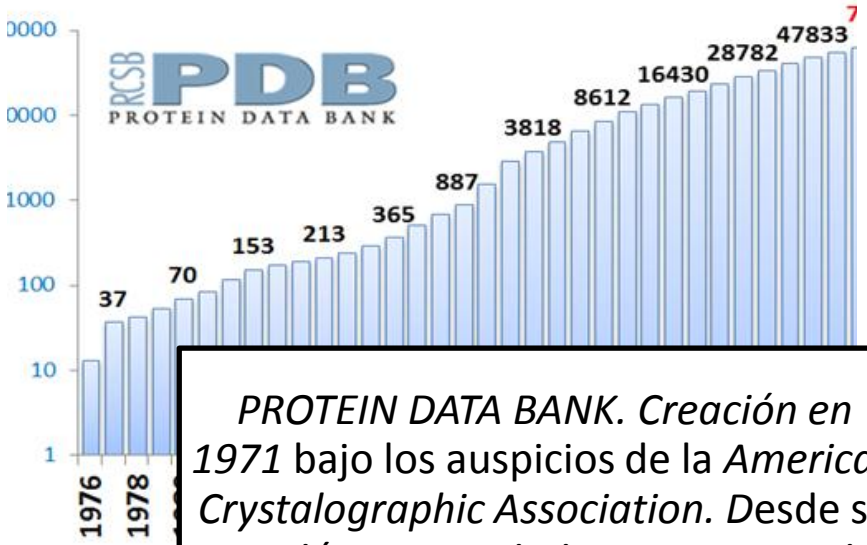
- Permitiría acabar con el fraude científico.
- Replicar trabajos fácilmente

Impacto

- Oportunidad de dar mayor visibilidad a nuestro trabajo.
- Mayor número de citas
- Fomenta colaboración

NO ES UNA CUESTIÓN NUEVA. ALGUNOS HITOS

GALTON. Pienso que nadie debiera publicar resultados biométricos sin depositar una copia de sus datos bien redactada y presentada en algún lugar donde todo aquel que lo deseara pudiera verificar su trabajo



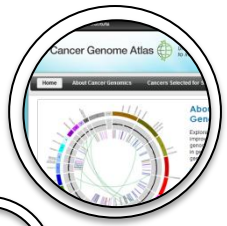
PROTEIN DATA BANK. Creación en 1971 bajo los auspicios de la *American Crystalographic Association. D*esde su creación, un crecimiento exponencial pasando de apenas 500 estructuras a finales de los ochenta a más de 75.000 en 2010

CADA VEZ EXISTEN UN MAYOR NÚMERO DE REPOSITORIOS

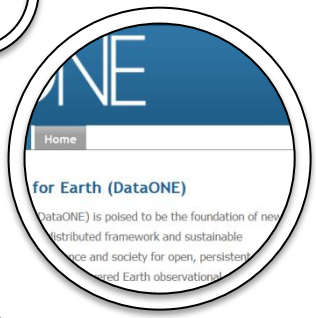
45% de los trabajos que utilizan gene expressions hacen públicos su datos

La creación de repositorios disciplinares es más la excepción que la regla

Neurodatabase

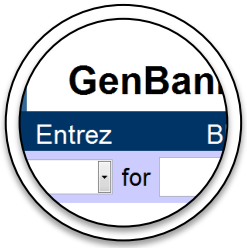


Cancer Genome Atlas



Data Observation Network for Earth

Protein Data Bank

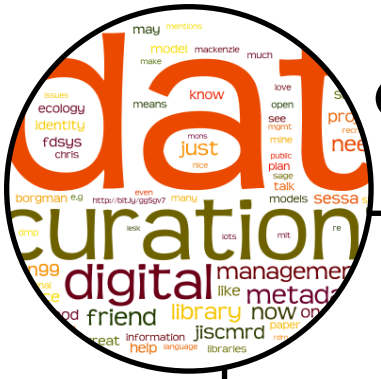


GenBank

CONSOLIDADOS NUEVOS ROLES PROFESIONALES

En el ámbito de las bibliotecas también es un tema en debate y ya se están empezando a asumir nuevos roles en relación a los datos

Data Curation Centers



Data Curation

¿CÓMO SE COMPARTEN LOS DATOS?

COMPARTIR DATOS: LOS CANALES INFORMALES

CANALES INFORMALES

PEER TO PEER (e-mail)

DESCENTRALIZADOS (WEBS..)

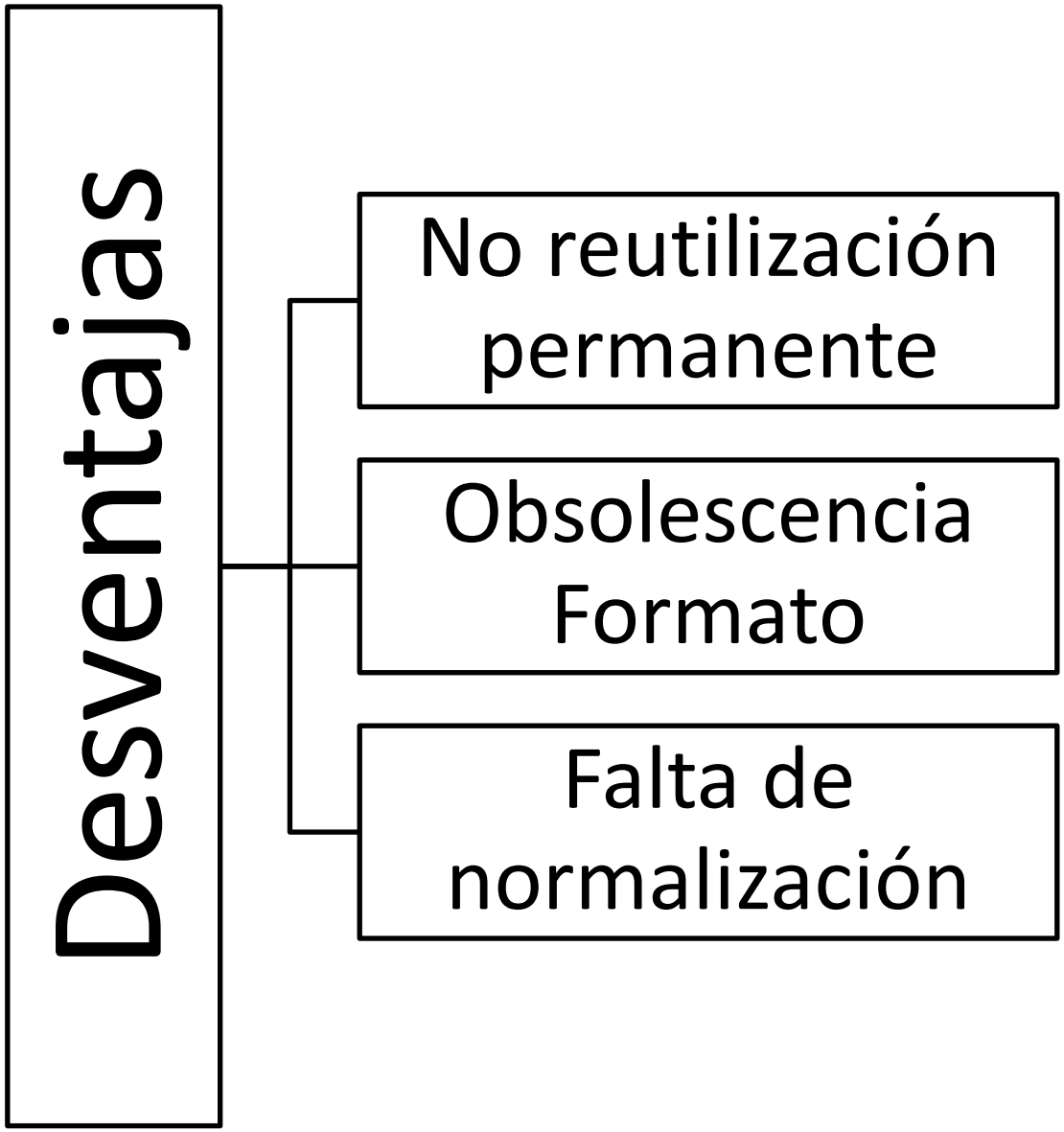
COMPARTIR DATOS: LOS CANALES INFORMALES

- [1380] R. Alcalá, M.J. Gacto, F. Herrera, **A Fast and Scalable Multi-Objective Genetic Fuzzy System for Linguistic Fuzzy Modeling in High-Dimensional Regression Problems**. *IEEE Transactions on Fuzzy Systems doi: 10.1109/TFUZZ.2011.2131657 19:4 (2011) 666-681.* ***COMPLEMENTARY MATERIAL to the paper*** *here: dataset partitions, results, figures, etc..*

- [1371] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, **An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-One and One-vs-All Schemes**. *Pattern Recognition 44:8 (2011) 1761-1776, doi: 10.1016/j.patcog.2011.01.017.* ***COMPLEMENTARY MATERIAL to the paper*** *here: dataset partitions, results, figures, etc..*

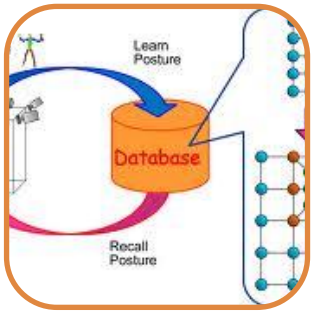
1. Paper abstract
2. Complementary Material 1: 5fcv partitions for the 17 data sets; Excel file with the results obtained by the different regression methods used in the paper; and average Pareto fronts obtained in all the studied datasets.
3. Complementary Material 2: Some examples on the influence of the Lateral Displacements and the Rule Cropping Strategy.
4. Complementary Material 3: Graphical representation of some examples on the extracted knowledge bases. It includes a zip file with the most accurate solution obtained by the proposed method in the first fold at each dataset.
5. Complementary Material 4: Tables of results and tables with Wilcoxon's Signed Rank test for two different running conditions of the comparison methods, GR-MF, GA-WM and GLD-WM (number of labels in {2,...,7} or {3,...,9}).
6. Complementary Material 5: Wilcoxon's Signed-Ranks test and its application in the regression framework.

COMPARTIR DATOS: LOS CANALES INFORMALES



COMPARTIR DATOS: LOS CANALES INFORMALES

CANALES FORMALES



REPOSITORIOS

REVISTAS CIENTÍFICAS

COMPARTIR DATOS: LOS CANALES FORMALES

REPOSITORIOS

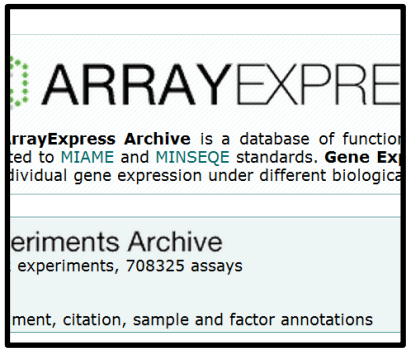
base de datos en línea donde los datos son depositados y descritos conforme a un estándar quedando listos para su posterior recuperación

existe una gran variedad de bancos de datos y las soluciones adoptadas en cada ámbito son muy diversas.

Asimismo también se diferencian en una mayor complejidad en su uso, tanto en el depósito como en la recuperación

COMPARTIR DATOS: LOS CANALES FORMALES

ARRAY ESPRESS



Genómica

863.732 experimentos y ensayos

DRYAD

Biociencias

902 paquetes de datos y 2157 ficheros de datos

CLINICAL TRIAL

Medicina

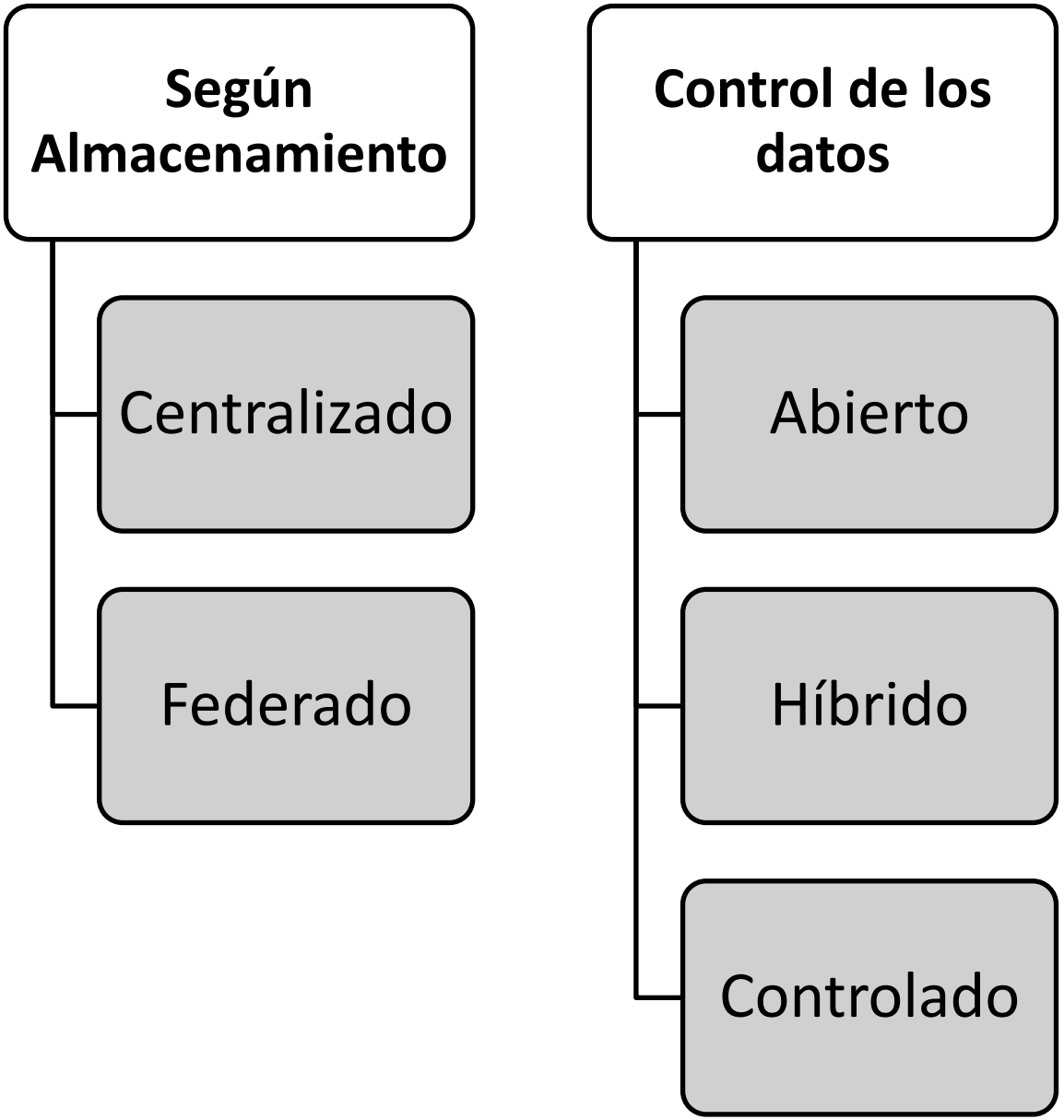
113.224 ensayos

Species 2000

Biología

1.368.009 especies de 100 bases de datos

TIPOS DE REPOSITORIOS



LA DIFICULTAD DE CONSUTA EN LOS REPOSITORIOS

Description of the queriable fields (249) :		
Criteria on basic data		
pm	proper motion (mas)	< / <= / > / >=
pmqual	proper motions quality (A:best, E:worst)	= / != / in
plx	parallaxes (mas)	= / != / < / <= /
plxqual	parrallax quality (A:best, E:worst)	= / != / in
rvtype	radial velocity type as the value was entered in the database ('v' = radial velocity, 'z' = redshift, 'c' = cz velocity)	= / !=
radvel	radial velocity (km/s)	= / != / < / <= /
redshift	redshift	= / != / < / <= /
cz	'cz' velocity	= / != / < / <= /

Un registro de DRYAD

Dryad File Identifier	doi:10.5061/dryad.1324/1 **58 views 26 downloads**
Description	Preparing input file for A. gymnandrum and its close species
Keywords	Alpine plants, coalescent tests, glacial refugia, Pleistocene divergence, Qinghai-Tibetan Plateau, Hybridization,
Date Deposited	2010-03-11T02:18:58Z
Scientific Names	Aconitum gymnandrum
Contained in Data Package	Data from: History and evolution of alpine plants endemic to the Qinghai-Tibetan Plateau: Aconitum gymnandrum (Ranunculaceae).

Show Full Metadata

Ficheros en el ítem

Ficheros	Tamaño	Formato	Vista
ITS-1-20 intragroup.nxs	26.10Kb	Unknown	Vista/<wbr/>Abrir

POR ESO EXISTEN UNA GRAN CANTIDAD DE ESTÁNDARES

Research community, funding agencies, and journals participate in the development of reporting standards for the bioscience domain to ensure that shared experiments are reported with enough information to be comprehensible and (in principle) reproducible, compared or integrated. Similar trends in both the regulatory arena and commercial science.

biosharing

STANDARD	FULL NAME	TYPE	DOMAIN(S) COVERED
SAO	Subcellular Anatomy Ontology	terminology artifact	anatomy
FuGEFlow	FuGEFlow	exchange format	experimental description (flow cytometry)
Gating-ML	Gating-ML	exchange format	data transformation methods (flow cytometry)
FuGE-ML	Functional Genomics Experiment Markup Language	exchange format	experimental description (functional genomics)
GCDML	Genomic Contextual Data Markup Language	exchange format	experimental description (genomics)

El rol de las agencias de financiación y las revistas científicas

COMPARTIR DATOS: LOS CANALES FORMALES

El valor de los bancos de datos como la vía óptima para compartir datos y sobre todo el reconocimiento que éstos han tenido en diversas comunidades científicas altamente especializadas se puede atribuir, gracias al establecimiento de políticas y al fomento de su uso, por parte de dos agentes principales.

ORGANISMOS FINANCIADORES

REVISTAS CIENTÍFICAS

POLÍTICAS INSTITUCIONALES DE DATA SHARING. EJEMPLOS

1994

Economic and Social Research

ESRC Data Policies and Standards

2000

National Science Foundation

NSF Data Sharing Policy and Data Management Plan Requirements

2011

Wellcome Trust

Sharing research data to improve public health: joint statement of purpose

POLÍTICAS INSTITUCIONALES DE DATA SHARING. NIH.

La razón que justifica compartir los datos desde el punto de vista institucional es que los datos que provienen de proyectos también son resultados de la investigación y por tanto, al igual que los artículos, deben hacerse públicos.

" To facilitate data sharing, investigators submitting a research application requesting $500,000 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible"

(NIH 2003)

POLÍTICAS SUPRANACIONALES DE DATA SHARING. OECD y UE

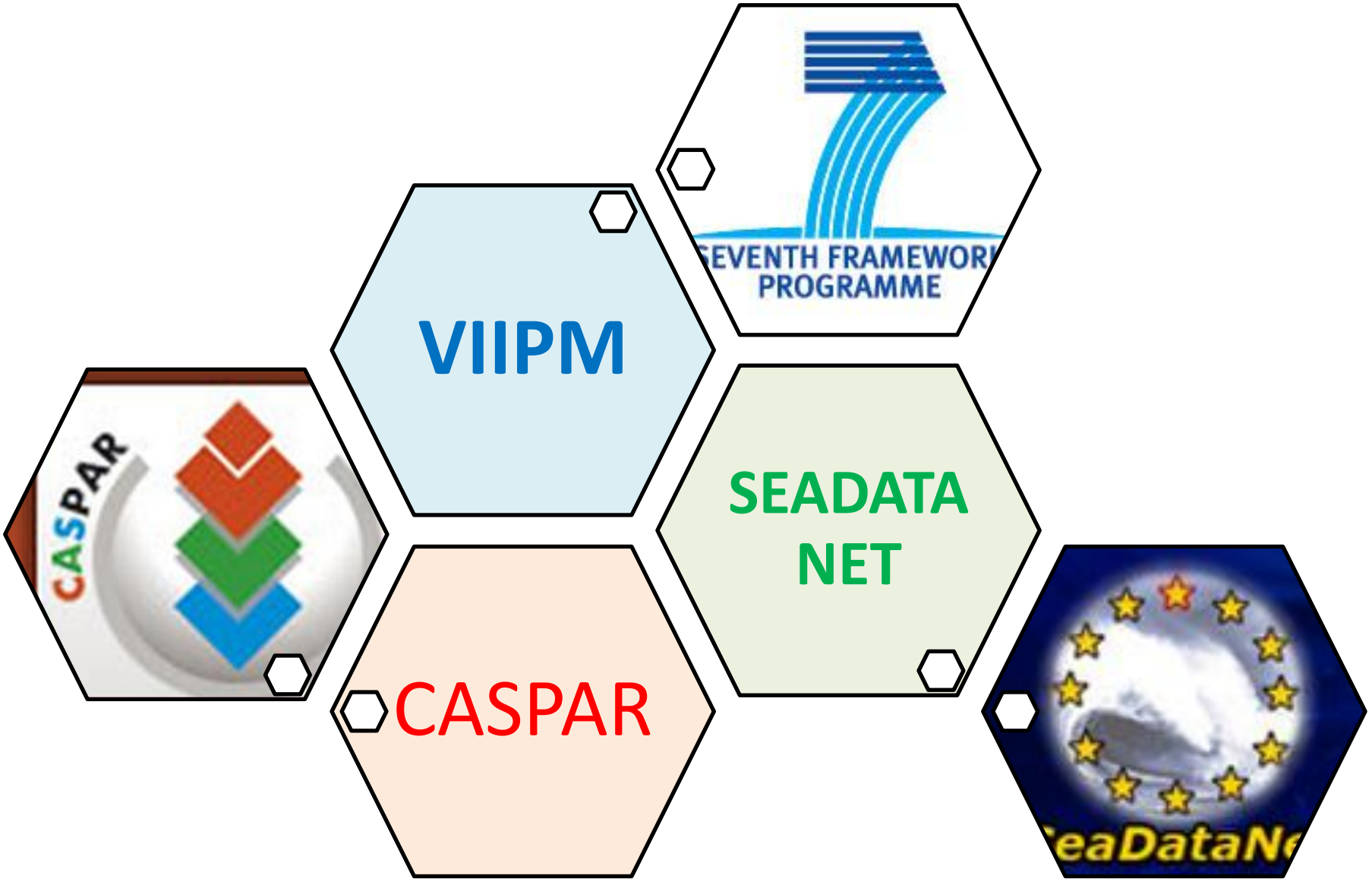
OECD

Principles and Guidelines for Access to Research Data from Public Funding

EU

On Scientific Information in the Digital Age: Access, Dissemination and Preservation

COMPARTIR DATOS: LOS CANALES FORMALES



COMPARTIR DATOS: LOS CANALES FORMALES

Las revistas están fomentando compartir en sus instrucciones. Las razones son favorecer la replicación, ayudar a los revisores y apoyar las políticas institucionales

Los requisitos que exigen generalmente es el respeto por la descripción de los estándares que describen los datos y emplear un repositorio “reconocido”

Los autores que publican en las revistas suelen estar obligados a compartir lo datos todos aquellos lectores que se los soliciten

POLÍTICAS DE DATOS DE LAS REVISTAS CIENTÍTICAS

Science apoya los esfuerzos de las bases de datos para compartir datos para el uso de la comunidad científica. Por ello, los sets de datos deberán ser depositados en un repositorio, y deberá indicarse en el artículo el número de registro o la dirección para su localización.

Los materiales, datos y protocolos asociados estén disponibles para ser consultados por quien lo desee. Los sets de datos deberán estar accesibles de manera gratuita desde la fecha de publicación y deberán suministrarse a los editores y revisores…

PLoS ONE promociona una investigación abierta y pretende que todos los trabajos que publica puedan servir como punto de partida para futuros científicos. Por ello, exigimos la aceptación de los estándares existentes parar el depósito público de datos.

NATURE: REPOSITORIOS RECOMENDADOS

Sharing data sets

A condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without preconditions.

Other datasets

In addition to the above-mentioned mandatory requirements for data submission to community-endorsed public databases, Nature journals strongly recommend deposition of other types of data sets into appropriate public repositories that are at an earlier stage of development. Examples of such repositories that facilitate sharing large data sets, some of which can offer the option of anonymous referee access to data before publication, include:

For **proteomics** data: PRIDE , PeptideAtlas, Tranche

For **protein interaction** data: IMEx consortium of databases including DIP, IntAct and MINT

For **cryoelectron micrographs**: EMDB (unified data resource for cryo-EM)

For **chemical compound** screening and assay data: PubChem

Other databases recommended by Nature journals include IntAct and the Global Proteome Machine Organization.

Earth sciences databases recommended by Nature journals include Pangaea, the publishing network for geoscientific and environmental data; PetDB, for geochemical data of rocks on the ocean floor; and GEOROC, geochemistry of rocks from the oceans and continents.

See also: World Data Center system; National Climatic Data Center.

Astronomy and Astrophysics: NucAstroData; Plasma gate; Smithsonian/NASA astrophysics data system; SIMBAD astronomical database; UK solar system data centre.

Physics: NIST physical reference data; Hepdata reaction data.

Biology: NBII; ITIS (taxonomy); NCBI taxonomy; Species 2000; National Center for Ecological Analysis and Synthesis; Dryad.

ALGUNAS REVISTAS FACILITAN COLGAR LOS DATOS DE SU WEB

RESEARCH ARTICLE OPEN ACCESS

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Article | Metrics | Related Content | Comments: 5

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

To **add a note**, highlight some text. Hide notes
Make a general comment

SUPPORTING INFORMATION Top

Text S1.

Cohort Publication Bibliography

(0.05 MB DOC)

Text S2.

Locations of Publicly Available Data for the Cohort

(0.05 MB DOC)

Text S3.

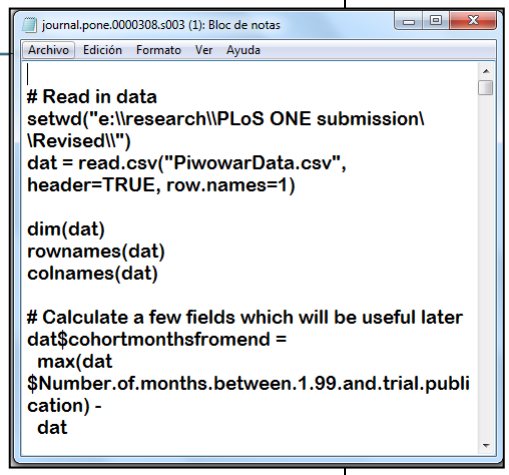
Statistical Analysis R-code

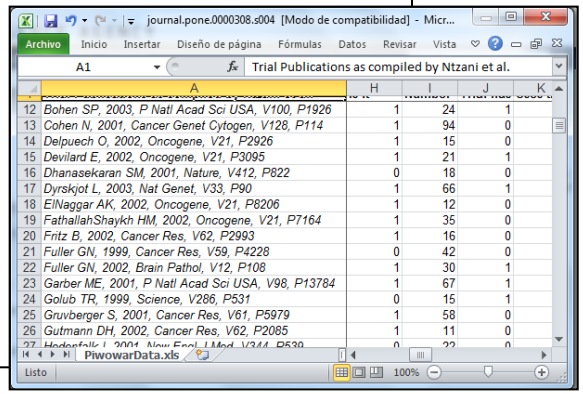
(0.01 MB TXT)

Data S1.

Raw Citation Counts and Covariates

(0.04 MB XLS)





ASCESSION NUMBER

Up-regulation of Expression of the *Ubiquitin Carboxyl-Terminal Hydrolase L1* Gene in Human Airway Epithelium of Cigarette Smokers

Brendan J. Carolan, Adriana Heguy, Ben-Gary Harvey, et al.

Cancer Res 2006;66:10729-10740. Published online November 15, 2006.

Web deposition of data. All data has been deposited in the Gene Expression Omnibus site,[3] which is curated by the National Center for Bioinformatics. Accession numbers include (*a*) HuGeneFL accession number GSE5056; (*b*) small airways HG-U133A, accession number GSE3320, already cited in ref. 31; (*c*) large Airways HG-U133A accession number GSE5057; (*d*) large airways HG-U133 Plus 2.0 accession number GSE5059; and (*e*) small airways HG-U133 Plus 2.0 accession number GSE[illegible]58 (31).

ASCESSION NUMBER

NCBI

DATASET BROWSER (CURATED)

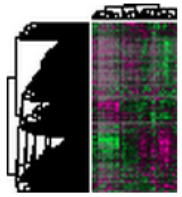
Gene Expression

Search for GSE5056[ACCN] Search Clear Show All Advanced Search

DataSet Record GDS2489: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Large airway epithelium response to cigarette smoking (HuGeneFL)		
Summary:	Analysis of large airway epithelial cells of phenotypically normal smokers. Lung cancer tumors exhibit neuroendocrine properties, and chronic smokers have increased numbers of neuroendocrine cells. Results provide insight into the effect of cigarette smoking on neuroendocrine cells.		
Organism:	*Homo sapiens*		
Platform:	GPL80: [Hu6800] Affymetrix Human Full Length HuGeneFL Array		
Citation:	Carolan BJ, Heguy A, Harvey BG, Leopold PL et al. Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers. *Cancer Res* 2006 Nov 15;66(22):10729-40. PMID: 17108109		
Reference Series:	GSE5056	**Sample count:**	44
Value type:	count	**Series published:**	2006/11/21

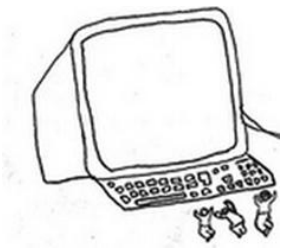
Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Consideraciones finales: el papel de los profesionales de la información



¿QUÉ PUEDEN HACER LOS BIBLIOTECARIOS?

- Proveer de demandas de datos
- Promoción del Data Sharing
- Desarrollo normas y directrices
- Intermediación con los repositorios

CAMPUS DATA MANAGMENT

Data Asset Framework

Four steps to effective data management

DATA SHARING

Acceso abierto a los datos de investigación: definición, políticas y actores

---

daniel torres-salinas

Mail: torressalinas@gmail.com

Twitter: @torressalinas

Instituto Politécnico de Braganca

26 de Octubre 2012