

EVENTOS

XV WORKSHOP REBIUN DE PROYECTOS
DIGITALES. DATOS Y BIBLIOTECAS

2016



Diseño de un repositorio de datos de investigación agrarios y alimentarios



crue

Universidades
Españolas

Red de Bibliotecas
REBIUN



crue

Universidades
Españolas

Red de Bibliotecas
REBIUN



Instituto Nacional de Investigación
y Tecnología Agraria y Alimentaria

Diseño de un repositorio de datos de investigación agrarios y alimentarios

Jorge García, Antonio Jesús Padial

Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria

XV WORKSHOP REBIUN. CASTELLÓN, 29-30 SEPTIEMBRE 2016



¿Por qué un repositorio de datos de investigación?

Entre otros muchos motivos:

- Como herramienta para evitar la pérdida de datos.
- Para dar respuesta a las obligaciones que imponen las convocatorias públicas de investigación y las revistas científicas.
- Para aumentar la visibilidad de los centros y los departamentos.
- Como plataforma de gestión interna de datos de investigación.
- Para facilitar la reutilización de los datos.



https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

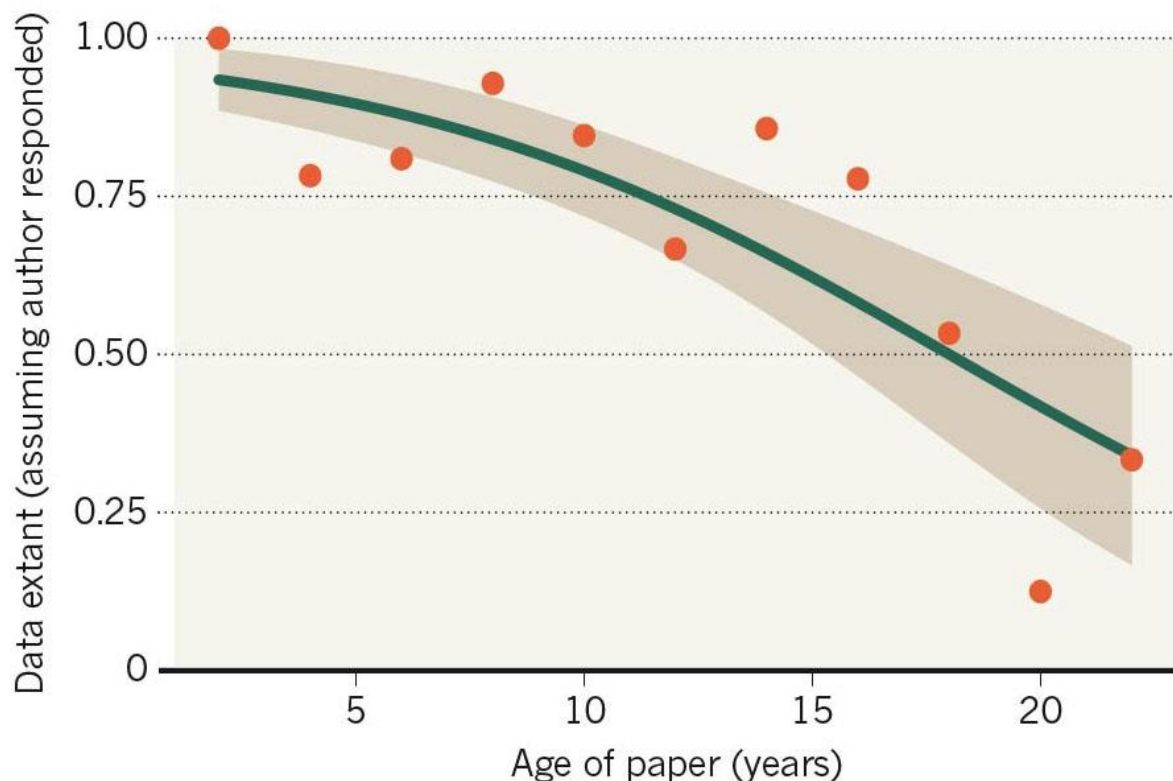
Para evitar pérdida de datos.



The availability of research data declines rapidly with article age. *Vines et al*

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



Estudio de 2013 sobre 516 artículos publicados con antigüedad de 2 a 22 años.

Mediante una consulta a los autores sobre los datos de los artículos.

Los resultados del estudio eran bastante lineales con pérdidas anuales de datos del 7% al 17%.

<http://dx.doi.org/10.1016/j.cub.2013.11.014>



Horizon 2020

https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

PROJECTS MUST HAVE



Provides information on:



the data the research will generate

Data management costs are fully eligible for funding



how to ensure its curation, preservation and sustainability

No repository imposed: deposit data where you want

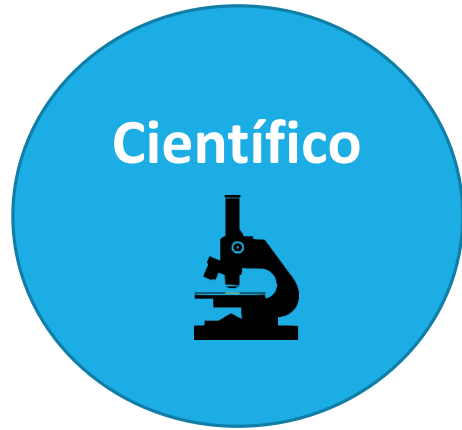
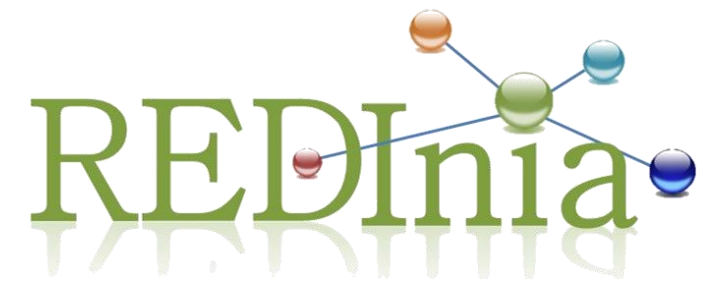


what parts of that data will be open (and how)



From 2017, **research data is open by default**, with possibilities to opt out

Equipos que trabajan en el proyecto



Investigadores
y personal de
apoyo a la
investigación.



Informática,
Biblioteca y
Biometría.





Nos inspiran, nos ayudan y han compartido con nosotros su tiempo y su conocimiento



datos.gob.es
reutiliza la información pública



Organización de las Naciones Unidas para la Alimentación y la Agricultura



¿Cómo debe ser nuestro repositorio?
¿Qué vamos a almacenar?
¿Cómo son nuestros datos?

Estimación inicial de tipos y volúmenes de datos.

Censo de Dataset

Estimación de volúmenes de datos

Colecciones de acceso público actualmente disponibles

¿A qué se dedica el INIA?

Organismo público de investigación dependiente de la AGE dedicado a la investigación agroalimentaria y forestal.

La misión del INIA, es el progreso sostenible del sector agroalimentario español mediante la investigación en: producción agrícola, ganadera y forestal, mitigación de los efectos del cambio climático, contribución a la seguridad alimentaria mundial y a la lucha contra el hambre en línea con los Objetivos del Milenio.

A continuación podemos observar sus centros y las áreas de investigación que se desarrollan.

3-Centros de investigación

CISA Centro Investigación en Sanidad Animal



Epidemiología,
Vigilancia enfermedades emergentes
Inmunología y Vacunas
Laboratorio seguridad Biológica niveles: 3 y 3+

CRF Centro de Recursos Fitogenéticos



Investigación aplicada a la conservación y
Utilización de Los Recursos Fitogenéticos para la
Agricultura y la Alimentación
Inventario Nacional de Recursos Fitogenéticos

CIFOR Centro De Investigación Forestal



Cultivos y Plantaciones Forestales
Dinámica y Funcionamiento de ecosistemas forestales
Protección Forestal
Productos Forestales



7 Departamentos

BIOTECNOLOGÍA

MEDIO AMBIENTE

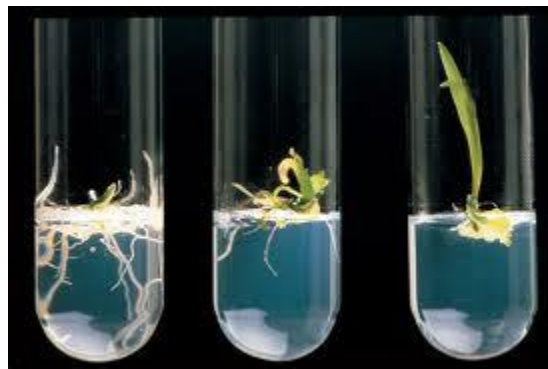
MEJORA GENÉTICA ANIMAL

PROTECCIÓN VEGETAL

REPRODUCCIÓN ANIMAL

CENTRO CALIDAD DE LOS ALIMENTOS (SORIA)

DIRECCIÓN TÉCNICA DE EVALUACIÓN DE VARIEDADES Y PRODUCTOS FITOSANITARIOS (DTEVPF)

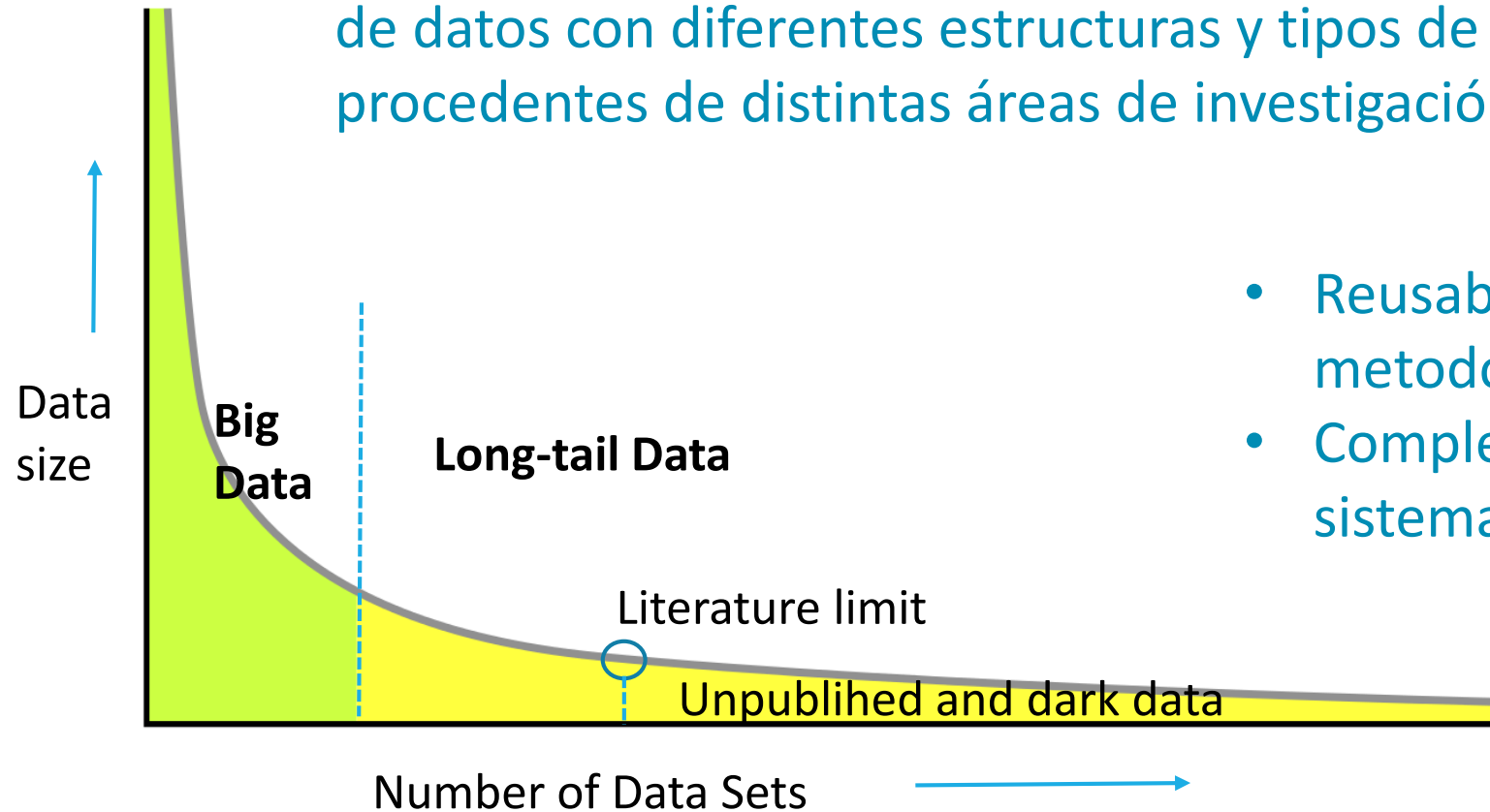


Multitud de áreas de investigación = diferentes tipos de datasets.



Datos de investigación del INIA = Datos de cola larga

INIA->Multitud de pequeñas y medianas colecciones de datos con diferentes estructuras y tipos de datos, procedentes de distintas áreas de investigación.



- Reusabilidad interdisciplinaria (de datos y metodología).
- Complejidad en los análisis y los sistemas de gestión.

Luis Martínez-Urbe, Biblioteca Fundación Juan March (Rebiun 2015)

Explore and download the Natural History Museum's research and collections data.

4.1M records

49 datasets

20 contributors

Search the Natural History Museum Specimen Collection

3,385,787 of the Museum's 80 million specimens are now available online.



1,355,303 Zoology

646,010 Botany

611,007 Entomology

392,148 Mineralogy

381,319 Palaeontology

Featured datasets



Index lots



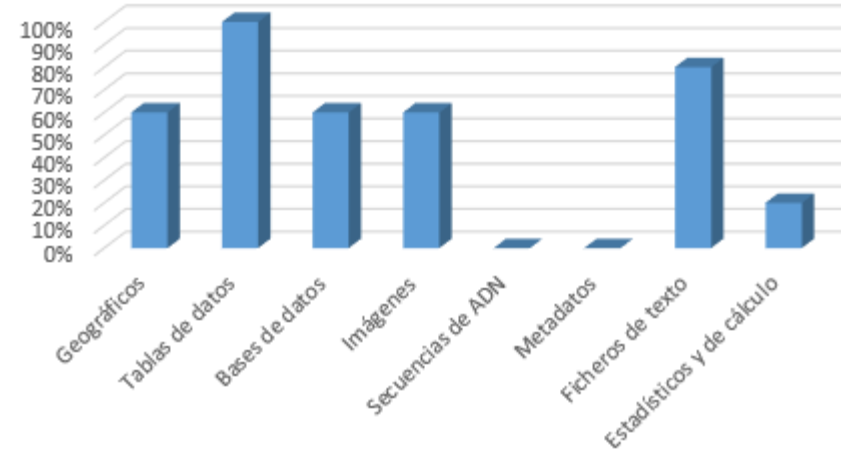
BioAcoustica



PREDICTS

data.nhm.ac.uk/
Un magnífico repositorio, compuesto por pocas colecciones con gran cantidad de elementos

Censo de datasets



CENSO DE DATASETS DEL INIA

Investigador

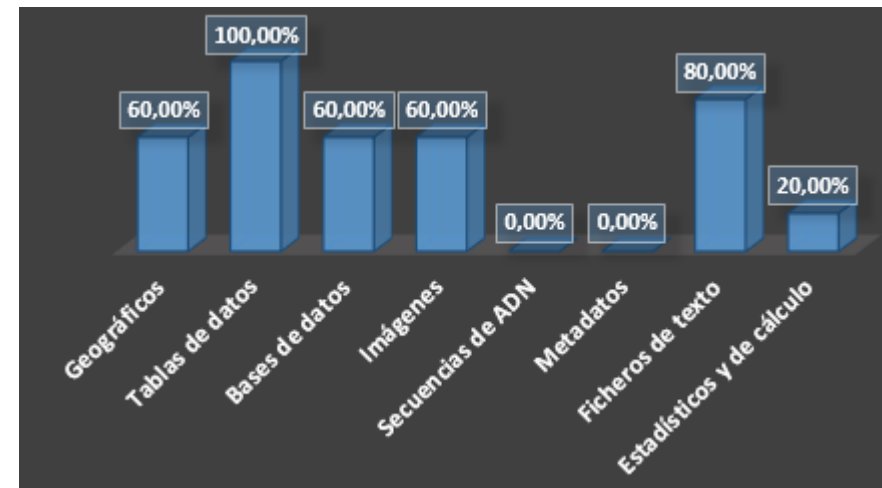
Centro/Departamento

Área/Sección

Línea de Investigación

5 cuestionarios, 3 centros, 18 datasets habituales

Tipos de Ficheros	Extensión						Tamaño de los ficheros en Mbytes	Número de ficheros por trimestre
	Shapefile	GeoCSV	DXF	Gml	arcGIS	Otros		
Geográficos	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Otros	--	
Tablas de datos	Excel	CSV	XML	Json		Otros	--	
Bases de datos	Access	Sql Server	MySql	Oracle		Otros	--	
Imágenes	Jpg	PNG	TIFF	RGB		Otros	--	
Secuencias de ADN	embl	genbank	Fasta			Otros	--	
Metadatos	XML	schema	SGML	RDF	Json	Otros	--	
Ficheros de texto	txt	CSV	Word	Pdf	HTML	Otros	--	
Estadísticos y de cálculo	.R	Matlab				Otros	--	
Otros							--	



Título descriptivo de los Datasets (P.e.: Red Nacional de Ensayos Genéticos Forestales (GENFOREST))

Volumen estimado basado en las compras de discos duros en los últimos años.

Compras de discos duros con cargo a proyecto -> estimación de 1TB/Mes*

(*los datos genéticos pueden sobrepasar estas previsiones)

**Estimación
cuantitativa**



**Recursos de
almacenamiento
adquiridos en INIA en los
últimos años**



**Estimación de recursos de
almacenamiento necesario
para los próximos años**

Catálogos de datos en las webs del INIA

Actualmente están disponibles unas 17 colecciones, de diferentes tipos:

- Herbarios digitales (soporte digital de la colecciones) 3500 elementos
- Bases de datos 1004 elementos
- Inventario Nacional de recursos Fitogenéticos, 77286 elementos
- Colecciones del centro de recursos Fitogenéticos 48000 elementos
- Mapas de distribuciones de especies 60 mapas
- Catálogo de insectos

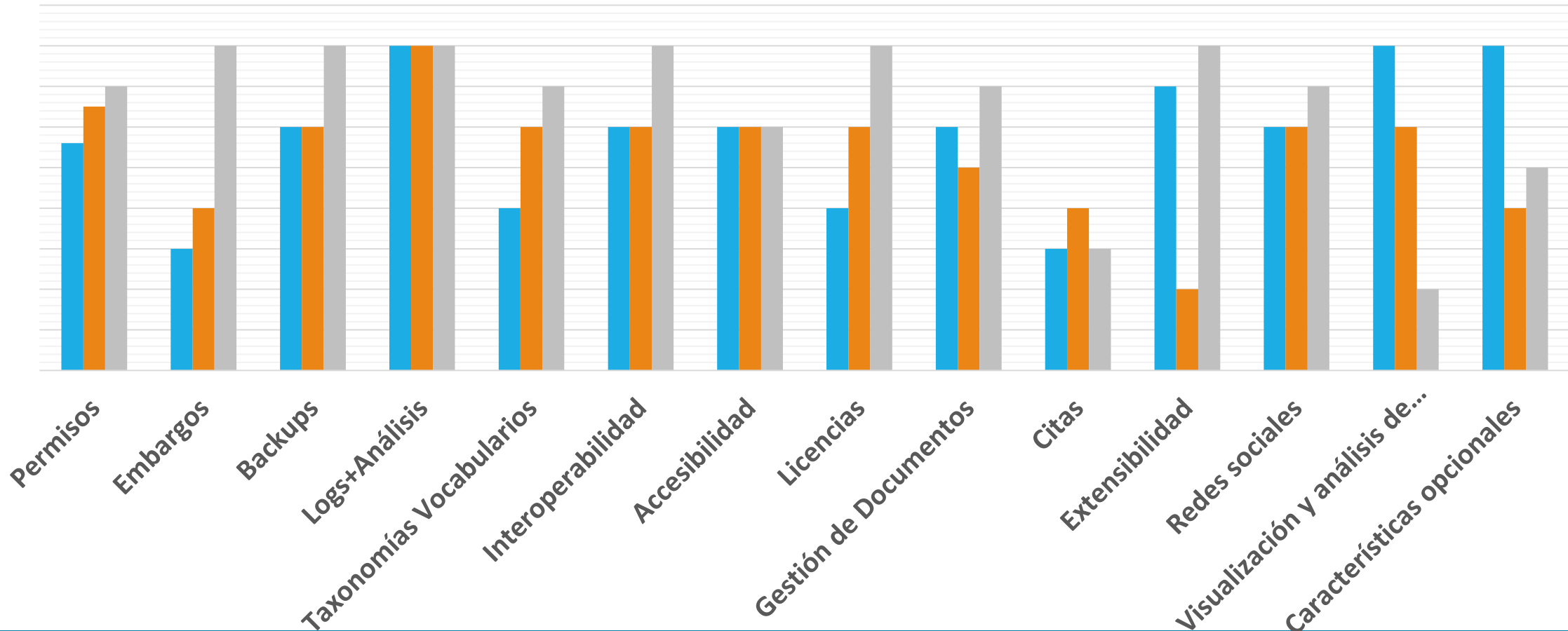
Otras colecciones (no son específicamente datos de investigación):

- Colecciones de fotografías históricas catalogadas 19.000 elementos.
- Hojas técnicas, archivos históricos, etc...

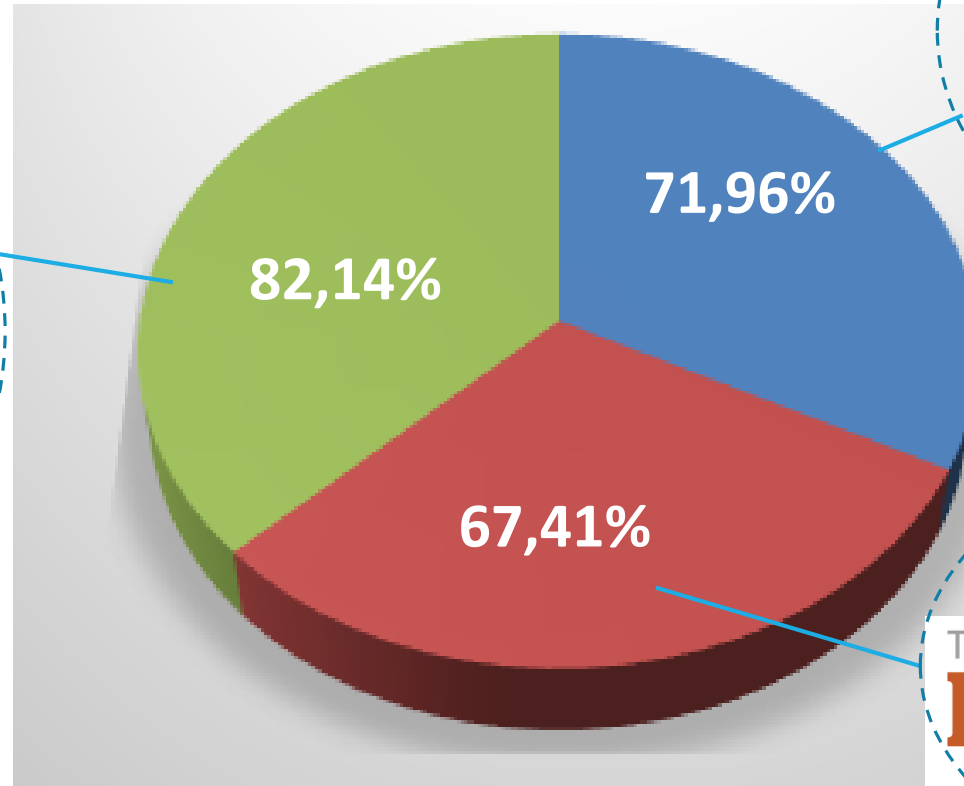
Evaluación de las plataformas



Resultados test Requerimientos



Resultados del test



eprints repository software

Apache

MySQL

Perl

Apache Solr

python

ckan

PostgreSQL

The Dataverse Project

GlassFish

PostgreSQL

Elección de para desarrollar



Elección de Ckan frente a ePrints y Dataverse:

1. Ckan es una plataforma de software libre y cuenta con una comunidad de desarrollo muy activa.
2. Es una plataforma específica para la gestión de datos de investigación (frente a eprints que está orientada a la gestión documental de resultados de investigación)
3. Interoperable y ampliamente utilizada en investigación y gestión de datos abiertos.

Inicio Sesión Registro

Conjuntos de datos Organizaciones Grupos Acerca de Búsqueda

REDInia

Organizaciones

¿Qué son las Organizaciones?

Las organizaciones en CKAN son usadas para crear, gestionar y publicar colecciones de conjuntos de datos. Los usuarios pueden tener diferentes perfiles en una organización, dependiente de su nivel de autorización para crear, editar y publicar

Buscar organizaciones

4 organizaciones encontradas Ordenar por: Nombre Ascendente

Biblioteca
El INIA cuenta con una Biblioteca científica especializada en Aricultura,...
1 Conjunto de Datos

CIFOR
El Centro de Investigación Forestal (CIFOR) realiza una parte importante de...
2 Conjuntos de Datos

CRF_INIA
El CRF tiene como objetivo principal contribuir a evitar la pérdida de...
2 Conjuntos de Datos

Repositorio de Datos
El objetivo de este grupo es

Piloto Ckan
INIA

Fases del proyecto

Servicio a los
Investigadores del
INIA

Fase 1

Servicio abierto a
la comunidad, datos
abiertos + datos
encontrables

Fase 2

Recolector de datos
agroalimentario y
forestales,
interoperable con
repositorios
supranacionales

Fase 3

Repositorio
Interno de
Datos de
Investigación

Portal de Datos
Agroalimentarios

Comunicación
con Fuentes de
Datos Externas

Nivel de
ejecución
actual



Metadata INIA



Metadata

Actualmente se está trabajando en el modelo de metadatos INIA, basándonos en diferentes repositorios e iniciativas y en las particularidades de nuestra organización.

DCAT Data Catalog Vocabulary



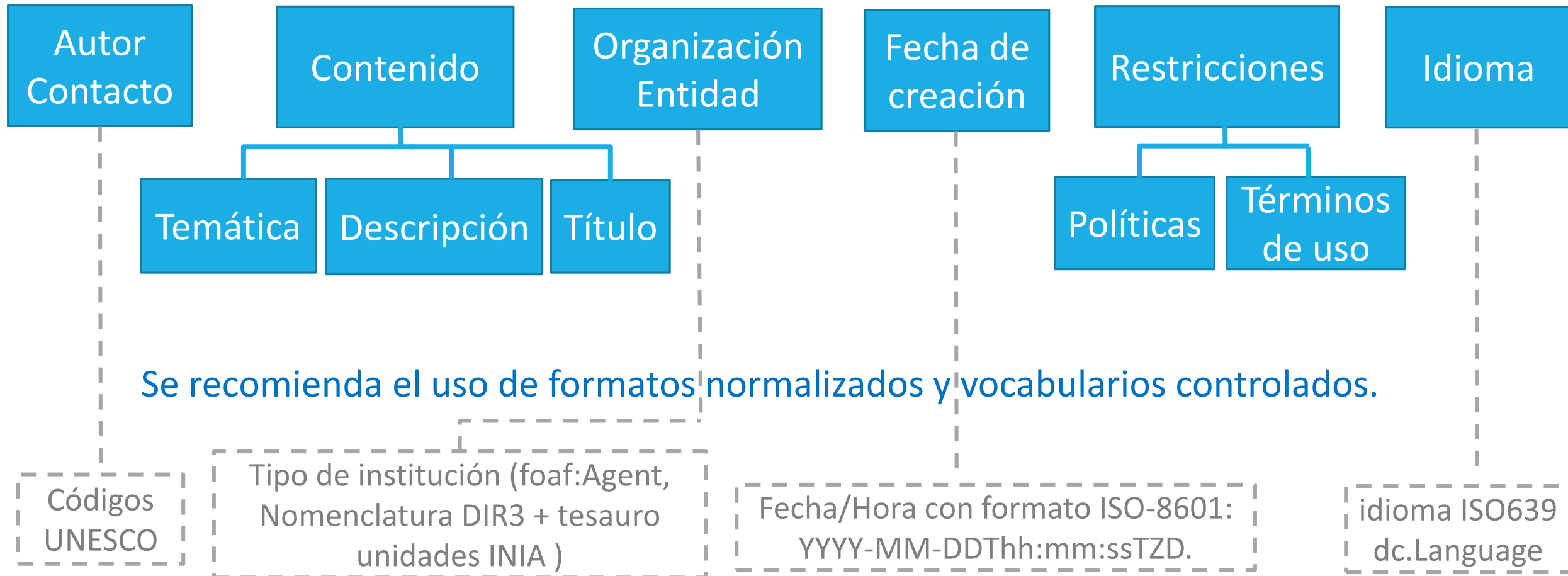
Esquema Nacional de Interoperabilidad



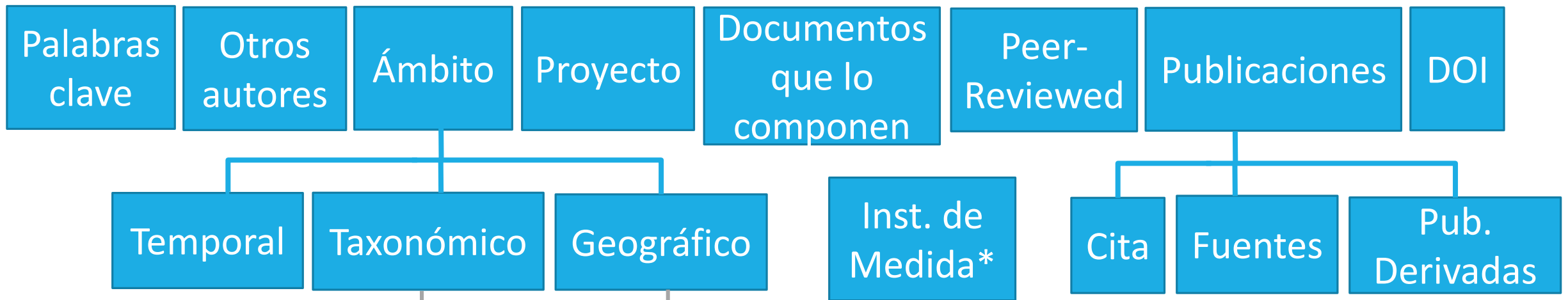
Documento de trabajo Grupo de Interés (IG) en METADATA de la Research Data Alliance (RDA) congreso 2016 Tokyo.



Campos requeridos (10)



Campos recomendados



Se recomienda el uso de formatos normalizados y vocabularios controlados.

GBIF, Darwin Core

ISO 19115:2003- Geographic Information Metadata

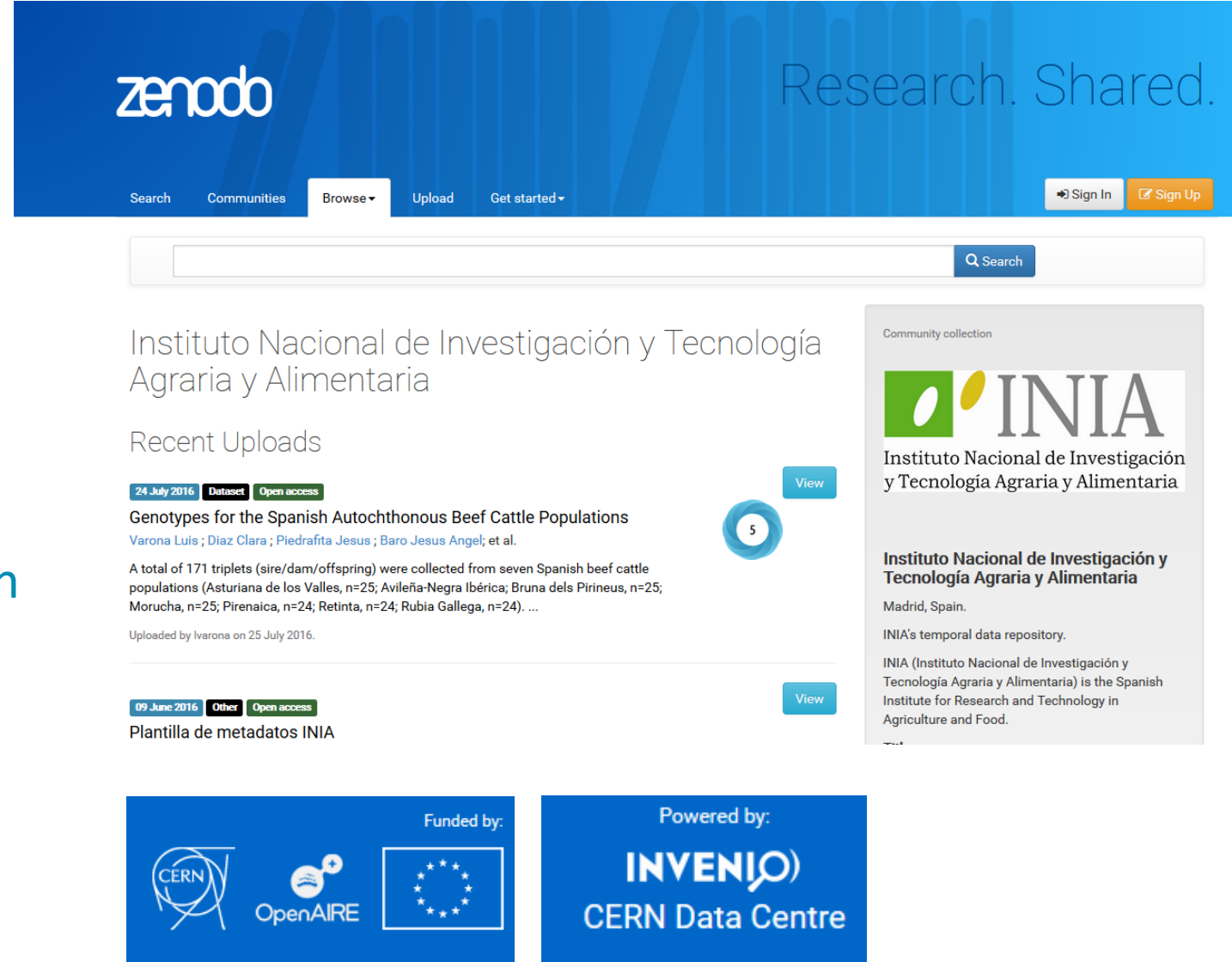
Utilización de Zenodo como repositorio provisional.

Se ha optado Zenodo para dar respuesta a las necesidades puntuales (hasta que esté operativa nuestra plataforma)

Zenodo cumple con muchas de las características requeridas

Proporcionando, facilidades para el alojamiento actual y su posterior migración

Es un repositorio gestionado por organismos públicos con solvencia técnica



The screenshot displays the Zenodo website interface. At the top, the Zenodo logo is on the left, and the tagline "Research. Shared." is on the right. Below the header, there are navigation links for "Search", "Communities", "Browse", "Upload", and "Get started", along with "Sign In" and "Sign Up" buttons. A search bar is positioned below the navigation. The main content area features a "Community collection" for "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (INIA). A "Recent Uploads" section lists a dataset titled "Genotypes for the Spanish Autochthonous Beef Cattle Populations" uploaded on 24 July 2016. The dataset description includes details about 171 triplets collected from seven Spanish beef cattle populations. Below this, another upload from 09 June 2016 is listed as "Plantilla de metadatos INIA". On the right side, a sidebar provides more information about INIA, including its location in Madrid, Spain, and its role as a temporal data repository for the Spanish Institute for Research and Technology in Agriculture and Food. At the bottom of the screenshot, there are logos for funding and power: "Funded by: CERN, OpenAIRE, and the European Union" and "Powered by: INVENIO CERN Data Centre".

Otras tareas en las que se está trabajando

- Plan de Gestión de Datos.
- Definición de políticas y licencias recomendadas
- Vocabularios controlados comunes (conjunto básico)
- Desarrollo de ontologías que sirvan de standard en ámbitos específicos
- Requisitos operativos y de usabilidad de la plataforma
- Difusión y soporte a los usuarios:
 - fomentar el uso de repositorios de datos
 - soporte para:
 - la elaboración de metadatos
 - el uso de formatos adecuados para su reutilización
 - la descripción interna de los Datasets

Gracias por la atención.



Jorge García Pérez
Informática INIA
jorge.garcia@inia.es

Antonio Jesús Sánchez Padial
Biometría INIA
antonio.sanchez@inia.es