crue Universidades Españolas Red de Bibliotecas REBIUN

INIA 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

REDInia

¿Por qué un repositorio de datos de investigación?

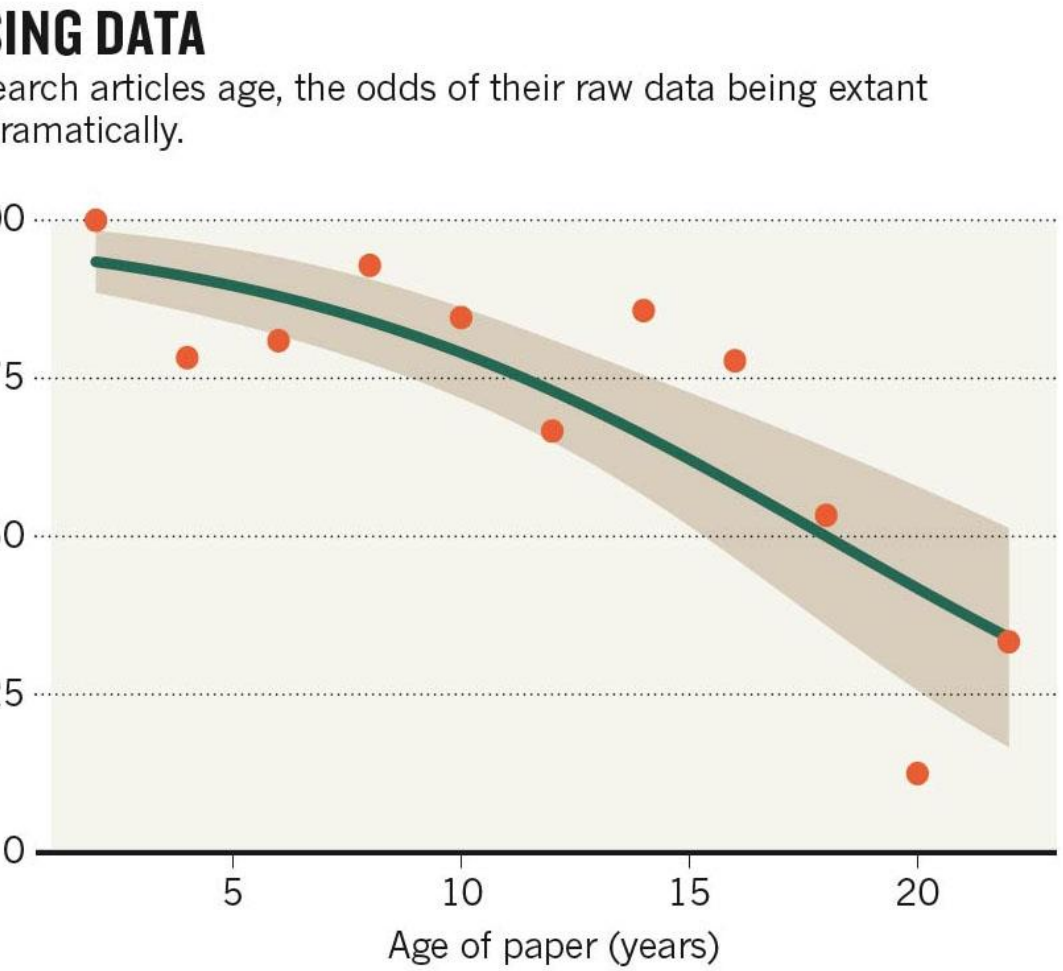
Entre otros muchos motivos:

- 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*



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

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

PROJECTS MUST HAVE

DATA MANAGEMENT PLAN (DMP)

Provides information on:

- the data the research will generate
- how to ensure its curation, preservation and sustainability
- what parts of that data will be open (and how)

Data management costs are fully eligible for funding

No repository imposed: deposit data where you want

Equipos que trabajan en el proyecto

REDInia

Científico

Investigadores y personal de apoyo a la investigación.

Dirección

Grupo de trabajo

Técnico

Informática, Biblioteca y Biometría.

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

Instituto de Salud Carlos III

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

red.es

cita
CENTRO DE INVESTIGACIÓN Y TECNOLOGÍA AGROALIMENTARIA DE ARAGÓN

FUNDACIÓN JUAN MARCH

datos.gob.es
reutiliza la información pública

GBIF
Global Biodiversity Information Facility

FECYT
FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA

PAGODA

Universidad de Alcalá

GODAN
Global Open Data for Agriculture and Nutrition

iata
Instituto de Agroquímica y Tecnología de Alimentos

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MINISTERIO DE MEDIO AMBIENTE
CONFEDERACIÓN HIDROGRÁFICA DEL JÚCAR

EUDAT

Organización de las Naciones Unidas para la Alimentación y la Agricultura

Red temática española de Open Data y Smart Cities

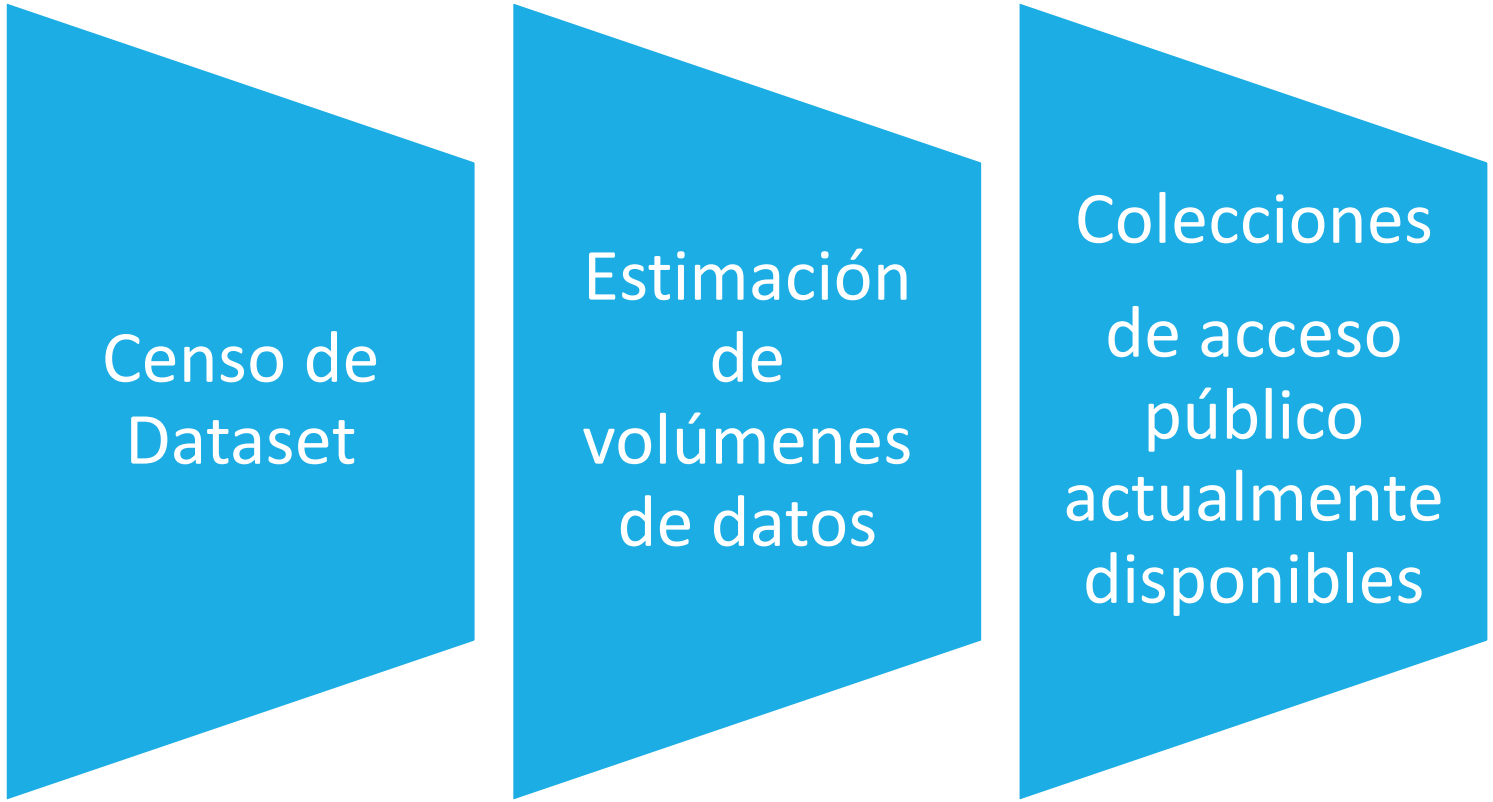
RDA
RESEARCH DATA ALLIANCE

MareData
Red Española sobre Datos de Investigación en Abierto

OPEN DATA CANARIAS

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

Estimación inicial de tipos y volúmenes de datos.



¿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

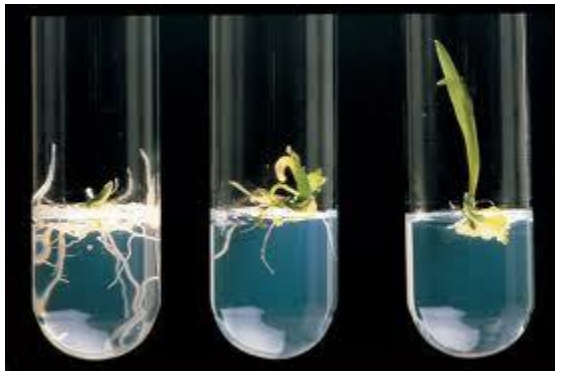
MEJORA GENÉTICA ANIMAL

PROTECCIÓN VEGETAL

REPRODUCCIÓN ANIMAL

CENTRO CALIDAD DE LOS ALIMENTOS (SORIA)

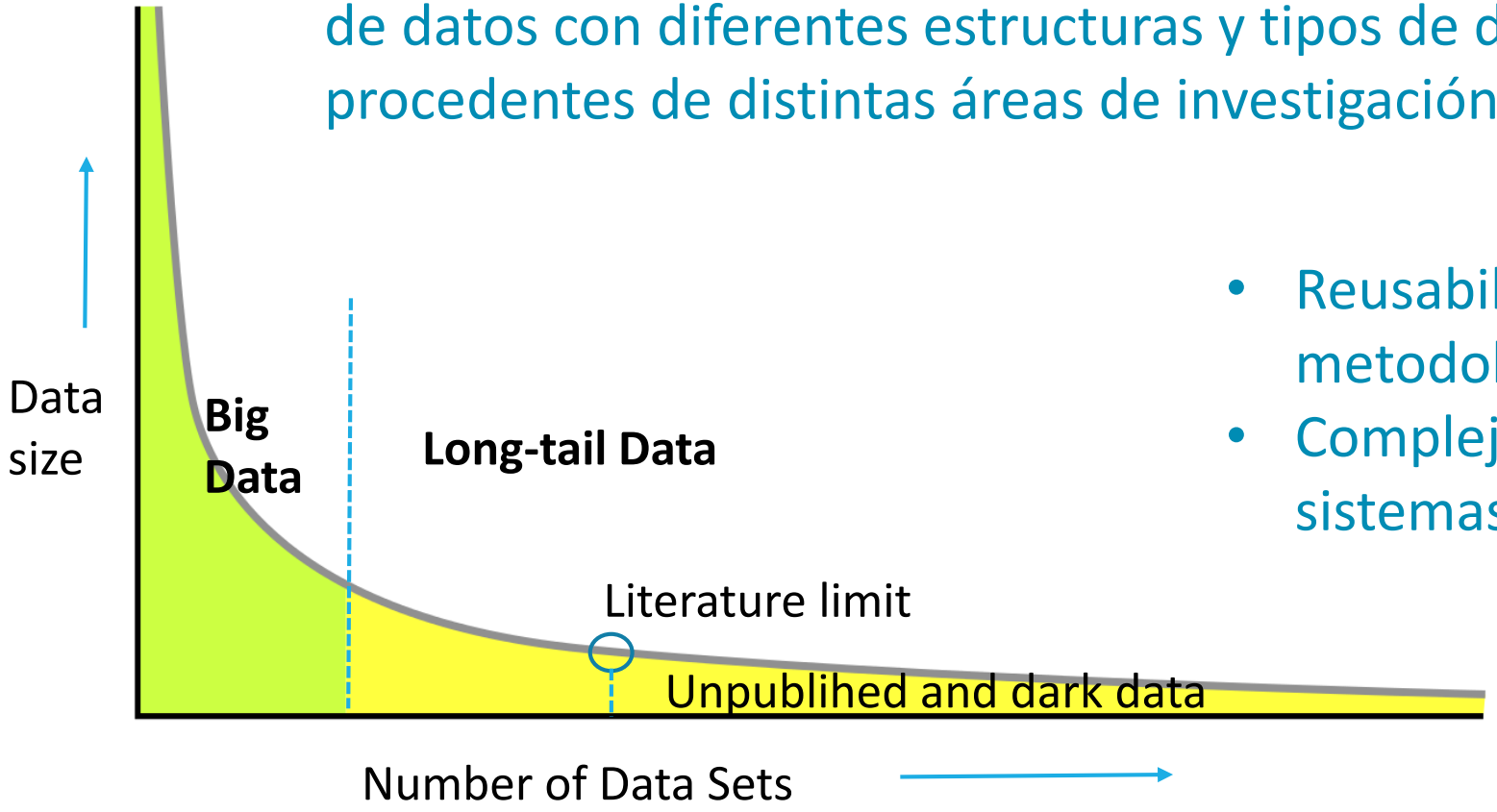
DIRECCIÓN TÉCNICA DE EVALUACIÓN DE VARIEDADES Y PRODUCTOS FITOSANITARIOS (DTEVPF)



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.



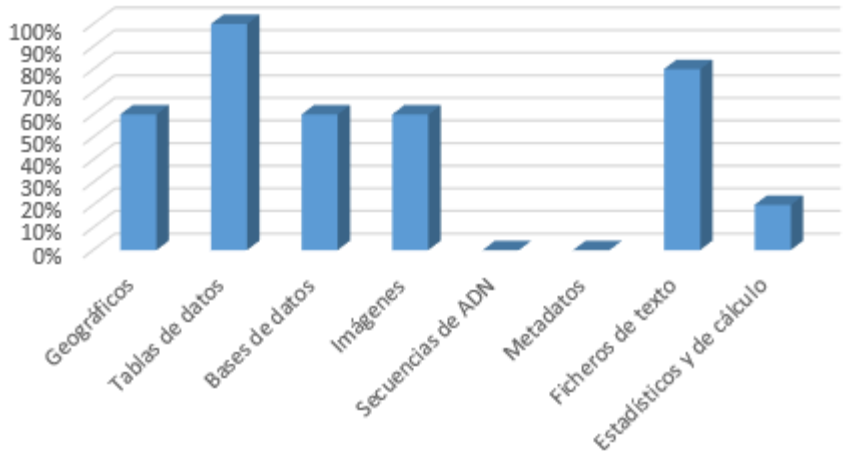
- Reusabilidad interdisciplinar (de datos y metodología).
- Complejidad en los análisis y los sistemas de gestión.

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

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

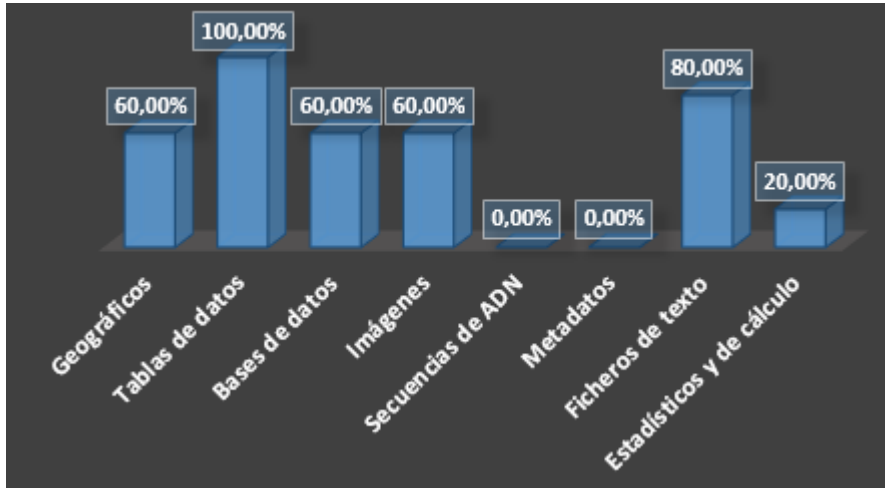
Área/Sección

Línea de Investigación

Tipos de Ficheros	Extensión						Tamaño de los ficheros en Mbytes	Número de ficheros por trimestre
Geográficos	Shapefile ☐	GeoCSV ☐	DXF ☐	Gml ☐	arcGIS ☐	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))

5 cuestionarios, 3 centros, 18 datasets habituales



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

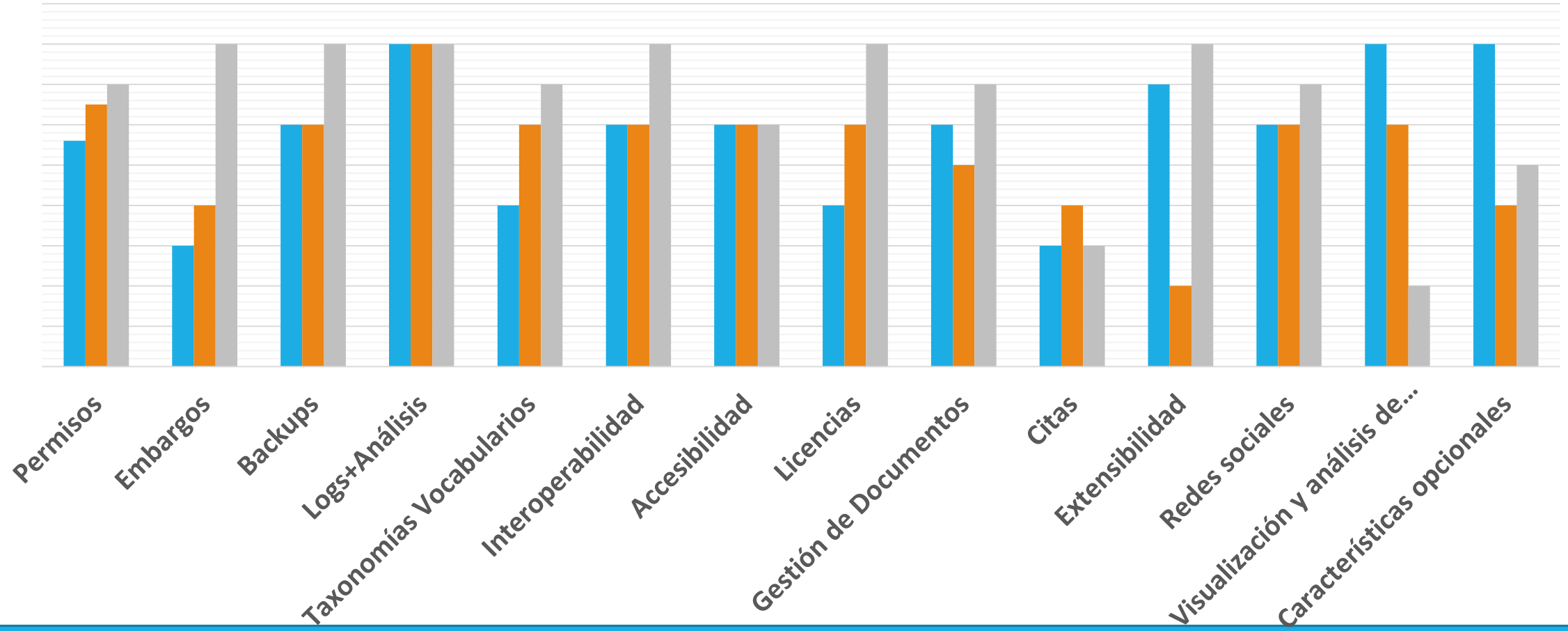
Otras colecciones (no son específicamente datos de investigación):

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

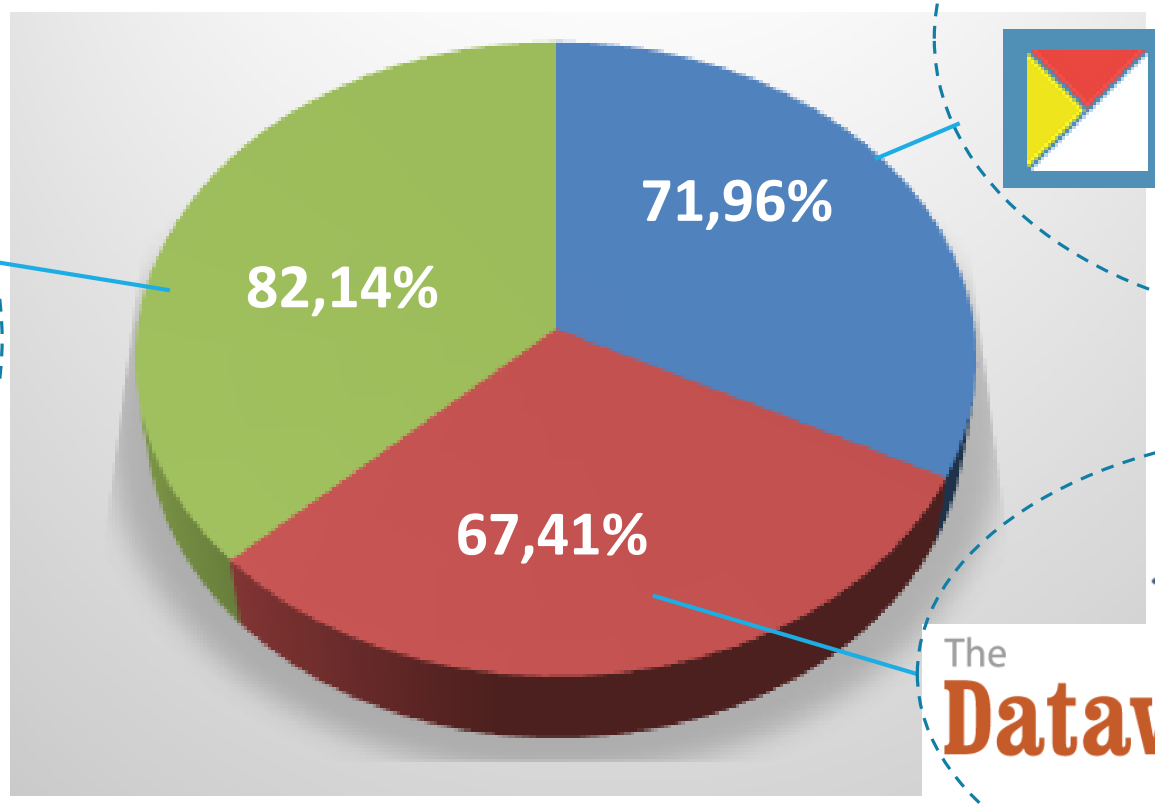
Evaluación de las plataformas

ckan · The Dataverse Project · eprints repository software

Resultados test Requerimientos



Resultados del test



Apache Solr
python
PostgreSQL
ckan

eprints repository software
Apache
MySQL
Perl

GlassFish
The Dataverse Project
PostgreSQL

Elección de ckan para desarrollar REDInia

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.

Piloto Ckan INIA

Fases del proyecto

Servicio a los Investigadores del INIA

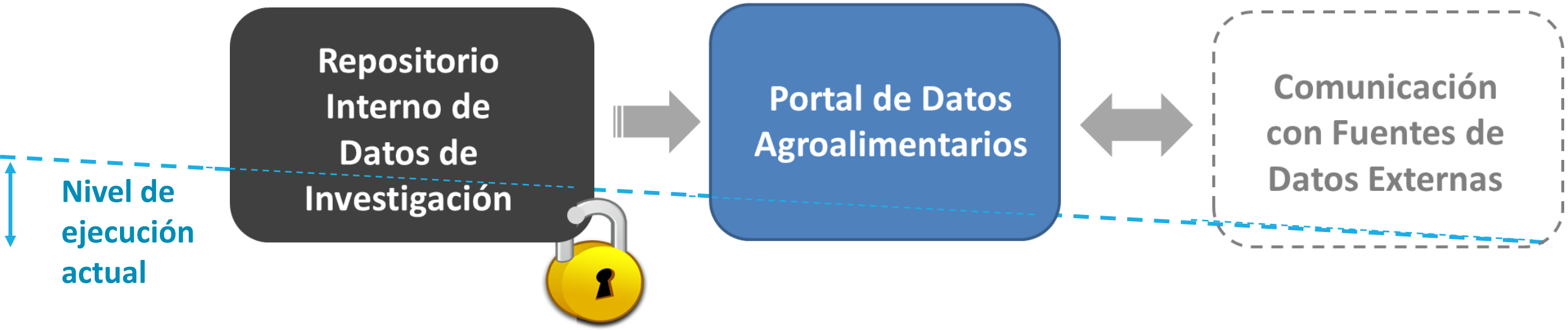
Servicio abierto a la comunidad, datos abiertos + datos encontrables

Recolector de datos agroalimentario y forestales, interoperable con repositorios supranacionales

Fase 1

Fase 2

Fase 3



Metadata INIA

REDinia 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

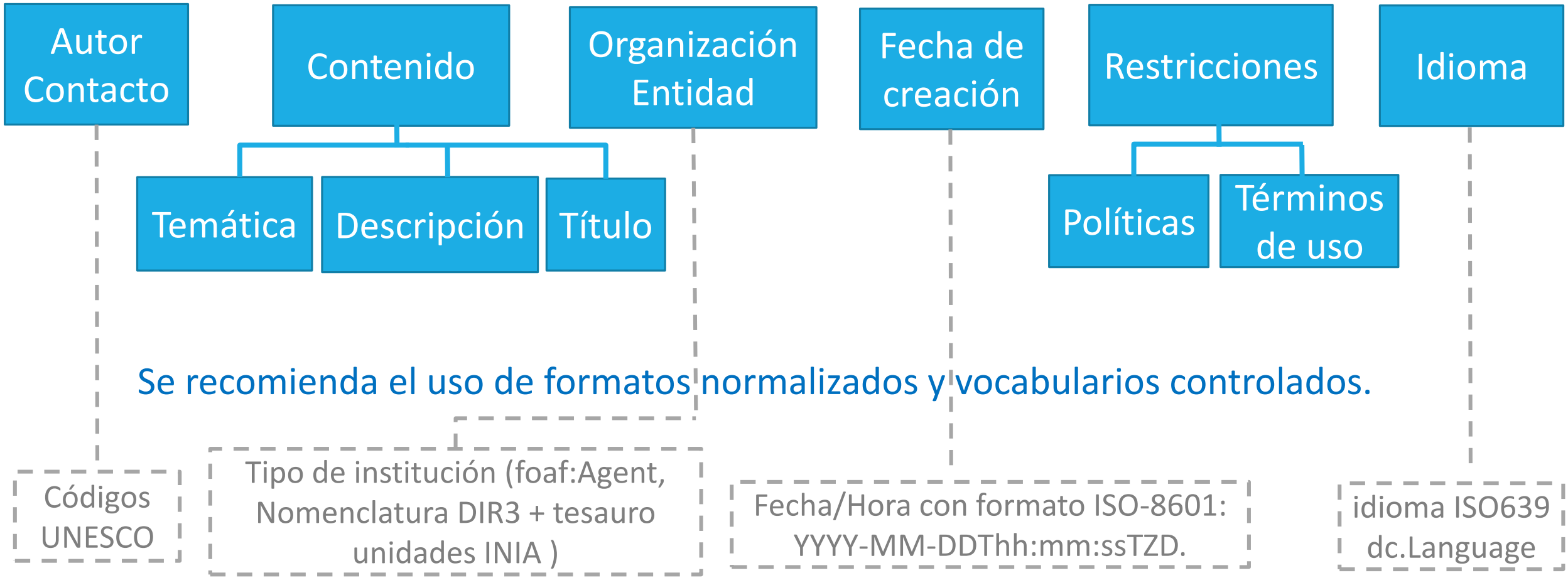
Esquema Nacional de Interoperabilidad

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

Metadatos INIA -I

REDInia Metadata

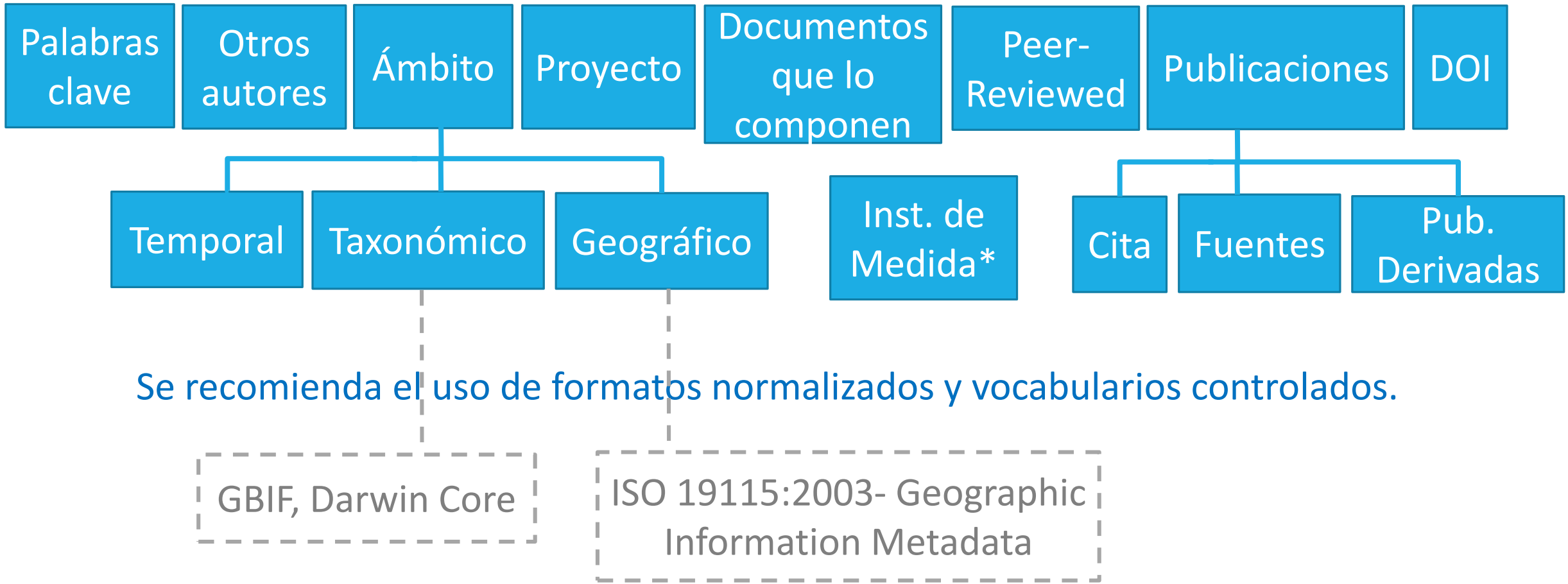
Campos requeridos (10)



Metadatos INIA -II

REDInia Metadata

Campos recomendados



Se recomienda el uso de formatos normalizados y vocabularios controlados.

GBIF, Darwin Core

ISO 19115:2003- Geographic Information Metadata

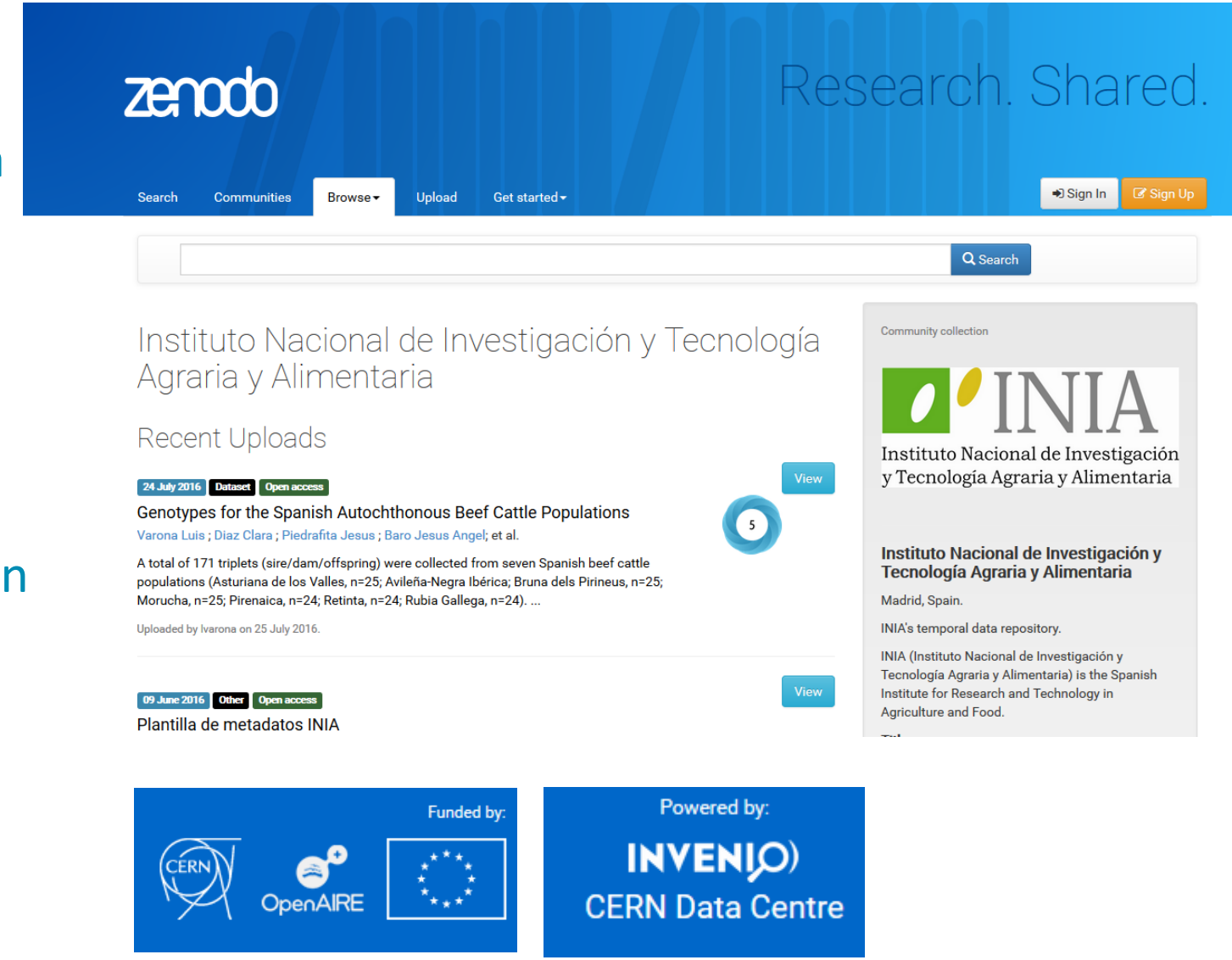
Utilización de Zenodo como repositorio provisional.

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



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