

Publicación de datos FAIR

Mikel Egaña Aranguren

mikel-egana-aranguren.github.io

mikel.egana@ehu.eus



Publicación de datos FAIR

<https://github.com/mikel-egana-aranguren/UM-Bioinformatics-MSc-FAIR-data>



Publicación de datos FAIR

```
git clone https://github.com/mikel-egana-aranguren/UM-Bioinformatics-MSc-FAIR-data.git
```

Índice

1. Introducción
2. Principios FAIR
3. Publicar datos FAIR
4. Recursos sobre FAIR
5. Linked Data
6. Ejemplo práctico
7. Proyecto a realizar

Introducción

Principios FAIR: una mejor publicación de datos (Científicos)

Para humanos y **máquinas**

No es un estándar

No promueven una tecnología concreta

Introducción

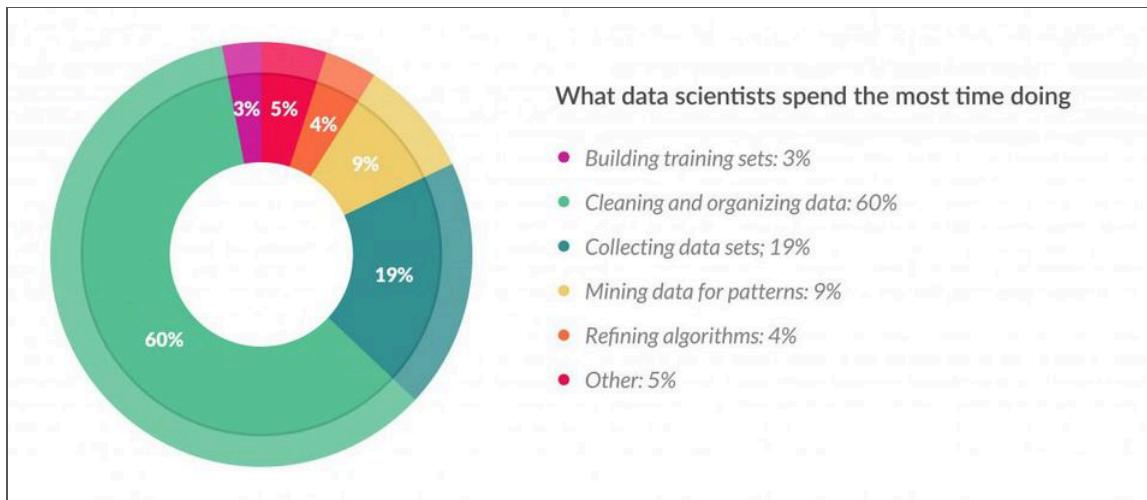
Son principios **guía**

No se cumplen de manera binaria (aprobado o no)

Un sistema siempre puede ser "más FAIR"

Introducción

80% del tiempo buscando, filtrando, masajeando e integrando datos



["Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" \[FORBES, 2020-11-19\]](#)

Introducción

La reproducibilidad es **crucial** en ciencia:

- Reproducir: ejecutar un experimento/estudio con los mismos datos/materiales
- Replicar: ejecutar un experimento/estudio con nuevos datos/materiales

Introducción

Crisis de la reproducibilidad debido a:

- Datos no publicados
- Datos publicados de manera inadecuada

Introducción

Principios FAIR para una mejor publicación de **(meta)datos**

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)

Introducción

Cada vez más agencias gubernamentales exigen cumplir los principios FAIR a la hora de publicar los resultados científicos para recibir financión (Data Management Plans): Horizon Europe, OpenAIRE, etc.

Open Data Europa para medir la calidad de los metadatos

Iniciativas del Espacio Europeo de Datos como GAIA-X

Introducción

Ley Orgánica del Sistema Universitario (LOSU): artículo 12 (Fomento de la Ciencia Abierta y Ciencia Ciudadana)

Estrategia Nacional de Ciencia Abierta 2023-2027 (ENCA) como objetivo estratégico

Grandes empresas como Novartis, Bayer, BASF, SIEMENS ENERGY etc. usan principios FAIR para publicación interna de datos

Principios FAIR

Findable

Accesible

Interoperable

Reusable

Findable

F1. (Meta)Data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (R1)

F3. Metadata clearly and explicitly include the identifier of the data it describes

F4. (Meta)Data are registered or indexed in a searchable resource

Accessible

A1. (Meta)Data are retrievable by their identifier using a standardized communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

- I1. (Meta)Data use a formal, accessible, shared, and broadly applicable language for Knowledge Representation
- I2. (Meta)Data use vocabularies that follow FAIR principles
- I3. (Meta)Data include qualified references to other (Meta)Data

Reusable

R1. (Meta)Data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)Data are released with a clear and accessible data usage license

R1.2. (Meta)Data are associated with detailed provenance

R1.3. (Meta)Data meet domain-relevant community standards

Principios y ejemplos

Las tecnologías para implementar los principios FAIR no son los principios FAIR

Findable

Data should be identified using globally unique, resolvable, and persistent identifiers, and should include machine-actionable contextual information that can be indexed to support human and machine discovery of that data

F1. (Meta)data are assigned a globally unique and persistent identifier

Globally unique

Dominios como um.es (€)

Registros

Algoritmos ([UUID](#))

etc.

Persistent

Infraestructura propia (€€€): por ejemplo [W3C URI Persistence Policy](#)

Registros: [identifiers.org](#), [DOI](#), [Orcid](#), [Zenodo](#), etc.

HTTP URIs

URI: Uniform Resource Identifier ([RFC 3986](#))

Identifica un recurso (URL: Localiza un documento)

HTTP: podemos usar HTTP para acceder (**Resolver**) a esa URI (dominio)

Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, et al. (2017) PLOS Biology 15(6): e2001414. <https://doi.org/10.1371/journal.pbio.2001414>

[papers/journal.pbio.2001414.pdf](#)

F1. Ejemplos

- <https://orcid.org/0000-0001-8888-635X>
- **doi:**10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f -
<https://doi.org/10.4121/UUID:5146DD06-98E4-426C-9AE5-DC8FA65C549F>
- <http://www.uniprot.org/uniprot/P98161>
- <http://omim.org/entry/173900>

F1. Nuestro ejemplo hipotético

Nuestro laboratorio de la UM ha descubierto un gen nuevo, PKD1, implicado en enfermedades renales de los humanos

La UM tiene un repositorio persistente de datos

URI del dataset: <https://um.es/dataset/UMGenesDataset>

URI de un gen: <https://um.es/data/LDD773322>

F2. Data are described with rich metadata (defined by R1 below)

Añadir metadatos lo más detallados posible

Metadatos de contenido: a qué especies pertenecen los genes, la temática de los datos, etc.

Metadatos técnicos: cuándo se generaron los datos, como, por quién, etc.

F2. Data are described with rich metadata (defined by R1 below)

Se usan ontologías (I1)

Repositorios de ontologías: [Linked Open Vocabularies](#), [OBO Foundry](#),
[BioPortal](#), [BioSchemas](#), etc.

F3. Metadata clearly and explicitly include the identifier of the data it describes

"<https://um.es/dataset/UMGenesDataset> was generated on 2020-12-10T13:00:07"

"<https://um.es/dataset/UMGenesDataset> is about genes"

"<https://um.es/dataset/UMGenesDataset> relates to humans"

etc.

F4. (Meta)data are registered or indexed in a searchable resource

Repositorios generales: [Zenodo](#), [DataDryad](#), [Dataverse](#) ([Harvard Dataverse](#)), etc.

Repositorios temáticos: [UniProt](#), [GenBank](#), etc.

Indexadores como Google

Indexación

Google indexa de manera "básica" ...

... pero cada vez menos, gracias a [Schema](#) (Ontología muy ligera para describir datos en la web) y [JSON-LD: Bioschemas](#)

Hay que intentar publicar buenos metadatos para una indexación adecuada (por Google o cualquier agente que *entienda* las ontologías que usamos)

Accessible

Identified data should be accessible, optimally by both humans and **machines**, using a clearly-defined protocol and, if necessary, with clearly-defined rules for authorization/authentication

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol

Por ejemplo [HTTPS](#)

A1.1 The protocol is open, free, and universally implementable

Por ejemplo [HTTPS](#)

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

Hacer explícitas las condiciones físicas de acceso, para humanos y **máquinas**

Datos protegidos por propiedad intelectual o privacidad (Ej. datos clínicos): no se publican los datos, sólo (algunos) metadatos y sus condiciones de acceso

A2. Metadata are accessible, even when the data are no longer available

Conservar datos es muy caro

Conservar metadatos es mucho más barato

Si los datos ya no existen, deberíamos ser explícitos sobre ello, por ejemplo para evitar búsquedas innecesarias

Interoperable

Data becomes interoperable when it is machine-actionable, using shared vocabularies and/or ontologies, inside of a syntactically and semantically machine-accessible format

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for Knowledge Representation

Las máquinas también tienen que *entender* los (meta)datos

Por ejemplo OWL ([Web Ontology Language](#))

I2. (Meta)data use vocabularies that follow FAIR principles

Las ontologías usadas para describir los datos también se tienen que publicar siguiendo los principios FAIR

I3. (Meta)data include qualified references to other (meta)data

Los (meta)datos son solo útiles cuando los integramos con otros datos

Enlaces explícitos a otros datos: "part-of", "catalyses", etc.

Las máquinas entienden el significado de esa relación

Reusable

Reusable data will first be compliant with the F, A, and I principles, but further, will be sufficiently well-described with, for example, contextual information, so it can be accurately linked or integrated, like-with-like, with other data sources. Moreover, there should be sufficiently rich provenance information so reused data can be properly cited

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

F2 es para descubrir datos, R1 es para decidir si los datos son útiles

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

Describe the scope of your data: for what purpose was it generated/collected?

Mention any particularities or limitations about the data that other users should be aware of

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

Specify the date of generation/collection of the data, the lab conditions, who prepared the data, the parameter settings, the name and version of the software used

Is it raw or processed data?

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

Ensure that all variable names are explained or self-explanatory (i.e., defined in the research field's controlled vocabulary)

Clearly specify and document the version of the archived and/or reused data

R1.1. (Meta)data are released with a clear and accessible data usage license

I es sobre interoperabilidad técnica; R1.1 es sobre interoperabilidad legal

Los datos deben tener una licencia clara y explícita para humanos y máquinas

Por ejemplo, [Creative Commons RDF](#)

R1.2. (Meta)data are associated with detailed provenance

¿Cómo, quién, cuándo, por qué generó los datos?

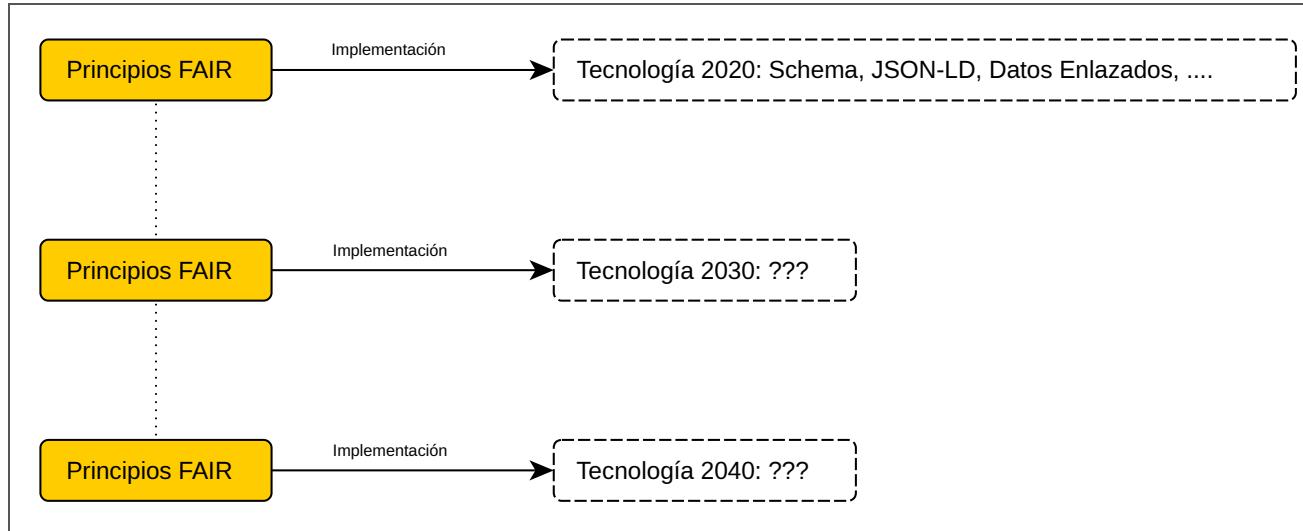
[PROV-O: The PROV Ontology](#)

R1.3. (Meta)data meet domain-relevant community standards

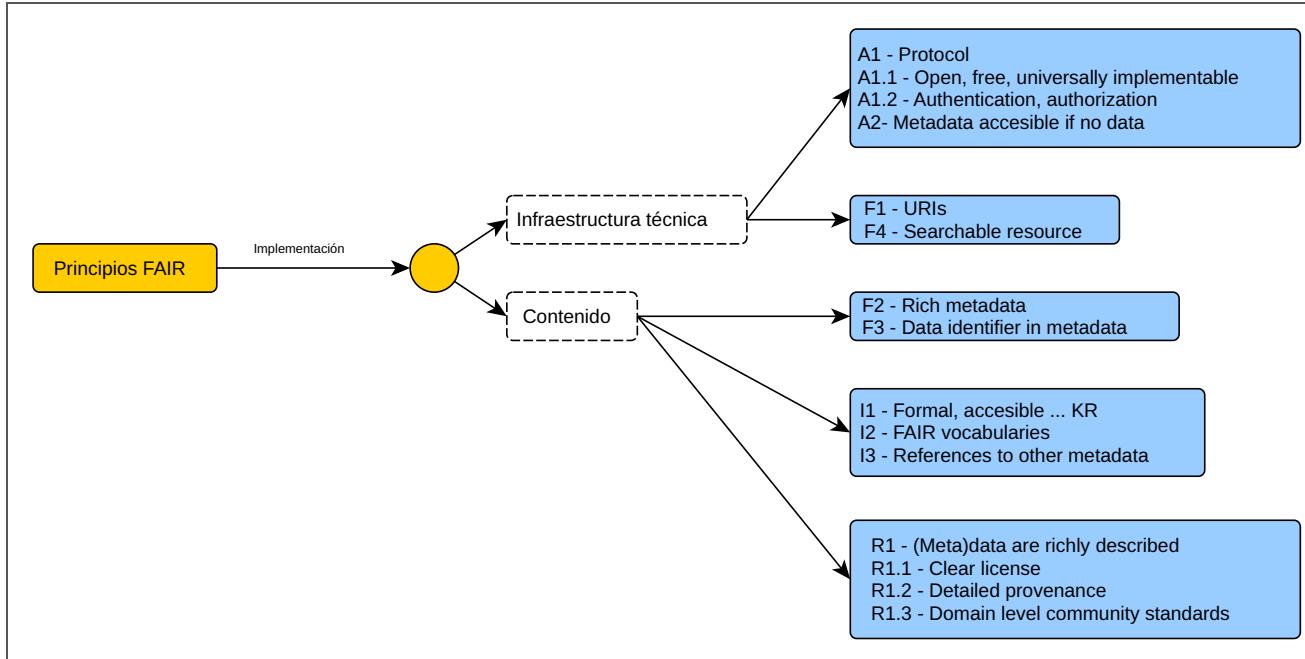
Respetar las buenas prácticas, estándares, vocabularios etc. de la comunidad científica que trabaja con esos datos

Por ejemplo [FAIR Sharing Standards](#)

Principios FAIR vs implementación



Principios FAIR: implementación



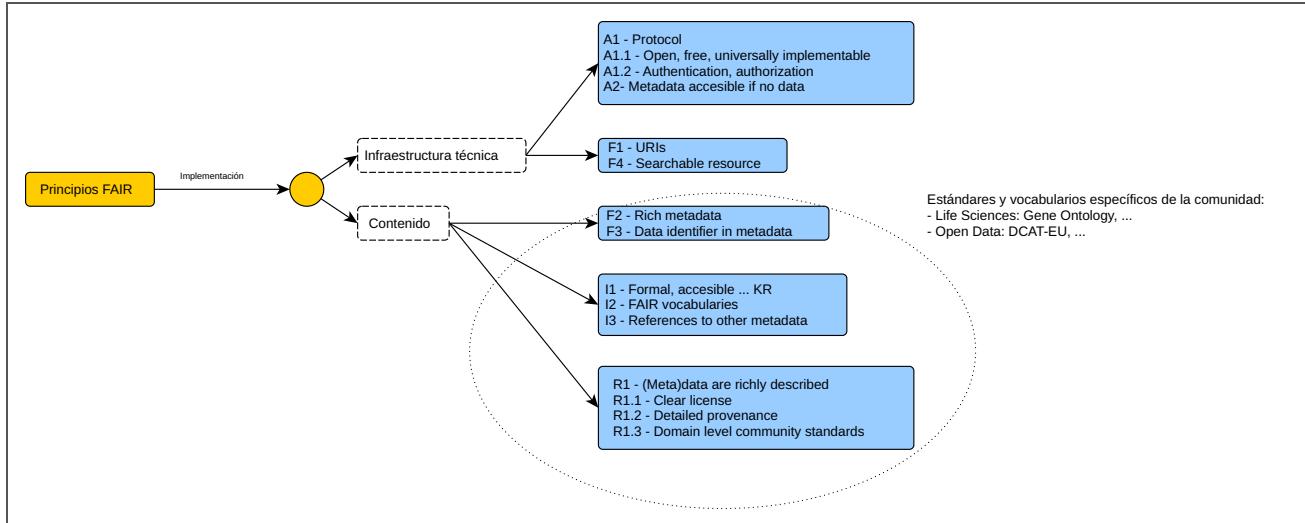
Madurez FAIR

A design framework and exemplar metrics for FAIRness

<https://github.com/FAIRMetrics/Metrics>

FAIR Evaluation Services

Madurez FAIR



Proyectos, Empresas, y centros

[Personal Health Train Network \(1812.00991.pdf\)](#)

[FAIR Data Systems](#)

[The Hyve](#)

[Eccenca GmbH](#)

[Dutch Tech Centre for Life Sciences \(DTL\)](#)

Recursos sobre FAIR en internet

[The FAIR cookbook](#)

[GO FAIR foundation](#)

[FAIR Sharing](#)

[FAIR-DOM](#)

Publicar datos FAIR

Hay muchas maneras de publicar datos FAIR, por ejemplo:

- Interoperability and FAIRness through a novel combination of Web technologies (Linked Data Platform, RML, Triple Pattern Fragments)
- FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication (FDP)
- Publishing FAIR Data: An Exemplar Methodology Utilizing PHI-Base (Linked Data)

Linked Data

Linked Data ofrece una solución *técnica* para principios FAIR

Pero no suficiente: hay que producir contenido FAIR (Metadatos, Ontologías, URIs, etc.)

Principios Linked Data

1. Usar URIs (Uniform Resource Identifier) para identificar entidades
2. Usar URIs que son accesibles mediante el protocolo HTTP(S), para que usuarios o agentes automáticos puedan acceder a las entidades
3. Cuando se acceda a la entidad, proveer datos sobre la entidad en formatos estándar y abiertos, como RDF (Resource Description Framework)
4. Añadir en los datos que publicamos en RDF enlaces a las URIs de otras entidades, de modo que un usuario o agente pueda navegar por la red de datos y descubrir más datos que también siguen los principios Linked Data

Linked Data: "Base de datos universal"

Utilizar maquinaria Web (URIs HTTP), para identificar y localizar entidades

Utilizar un modelo de datos común, tripleta RDF, para integrar datos en los que aparecen esa entidades

base de datos universal

Ventajas Linked Data

Descubrimiento e integración de datos

Programación de agentes que consuman los datos

Actualización de datos mediante enlaces

Consultas complejas

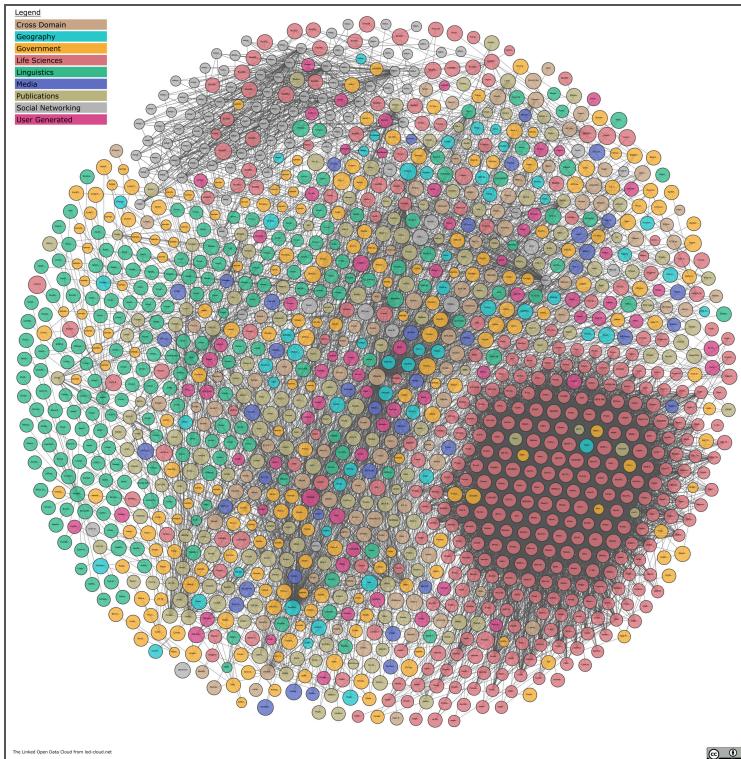
Ventajas Linked Data

Con Linked Data cualquiera puede publicar datos y enlazarlos a otros datos

El conjunto de datos abiertos publicados mediante Linked Data forma la «nube Linked Open Data»

Cada vez más instituciones públicas de todo el mundo usan Linked Data para publicar sus datos

Linked Open Data cloud



Linked Data: negociación de contenido

```
curl -L -H "Accept: text/html" "http://dbpedia.org/resource/Berlin"
```

```
| curl -L -H "Accept: application/rdf+xml" "http://dbpedia.org/resource/Berlin"
```

```
<owl:sameAs rdf:resource="http://eu.dbpedia.org/resource/Berlin" />
<owl:sameAs rdf:resource="http://linkededgeodata.org/triplify/node24e0109189" />
<owl:sameAs rdf:resource="http://sws.geonames.org/2950159" />
```



```
curl -L -H "Accept: text/html" "http://sws.geonames.org/2950159/"
```

A screenshot of a web browser displaying a map of a forested area. A search bar is overlaid on the map, containing the URL "http://eulerNames.org/". The browser's address bar shows the URL "eulerNames.org/".

```
curl -L -H "Accept: application/rdf+xml" "http://sws.geonames.org/2950159/"
```

```
<@country> /> <@parentCountry>/code  
<@population>3426354</@population>  
<@gs84_pos>lat:52.52437/+ngs84_pos:lat  
<@gs84_pos>long:13.41053/+ngs84_pos:long  
<@gs84_pos>lat:74/+ngs84_pos:lat  
<@gs84_pos>long:15.3/+ngs84_pos:long  
<parentCountry> rdfs:resource="http://www.geonames.org/6547539"/>  
<parentCountry> rdf:type="http://www.w3.org/2002/07/owl#Class"/>  
<parentCountry> rdfs:label="Poland"/>  
<parentCountry> rdfs:comment="The Republic of Poland is a country in Central Europe. It is bordered by the Baltic Sea to the north, Belarus to the east, and Ukraine to the southeast. Its capital and largest city is Warsaw. Poland's population is approximately 38 million, making it the 35th most populous country in the world. The country is divided into 16 voivodeships and includes the city of Warsaw as a separate administrative unit. The official language is Polish, and the currency is the złoty. Poland is a member of the European Union and the North Atlantic Treaty Organization."/>  
<parentCountry> rdfs:subClassOf="http://www.geonames.org/6547539"/>  
<parentCountry> rdfs:subClassOf="http://www.geonames.org/6547539#nearBy"/>  
<parentCountry> rdfs:subClassOf="http://www.geonames.org/6547539#nearBy"/>
```

Linked Data: negociación de contenido

```
curl -L -H "Accept: text/html" "http://dbpedia.org/resource/Berlin"
```

```
curl -L -H "Accept: application/rdf+xml" "http://dbpedia.org/resource/Berlin"
```

```
curl -L -H "Accept: text/html" "http://sws.geonames.org/2950159/"
```

```
curl -L -H "Accept: application/rdf+xml" "http://sws.geonames.org/2950159/"
```

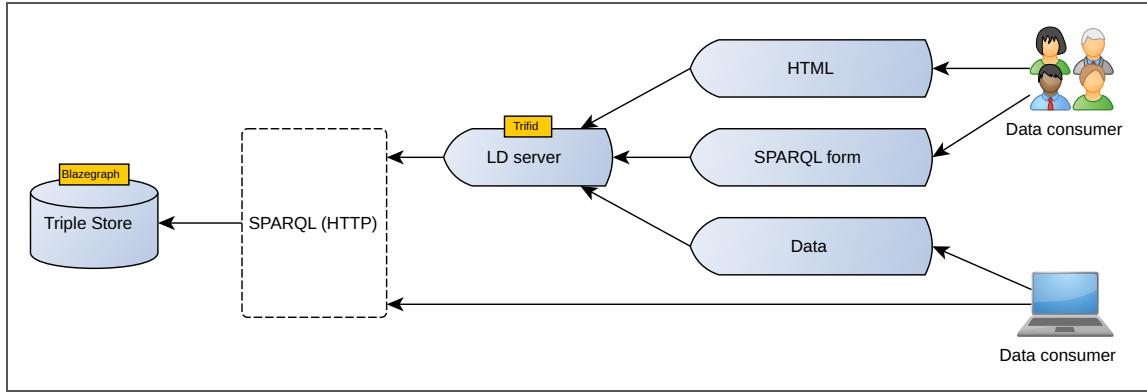
URIs/URLs en Linked Data

URI identifica a entidad; URLs localizan diferentes representaciones (RDF, HTML, ...) de la entidad

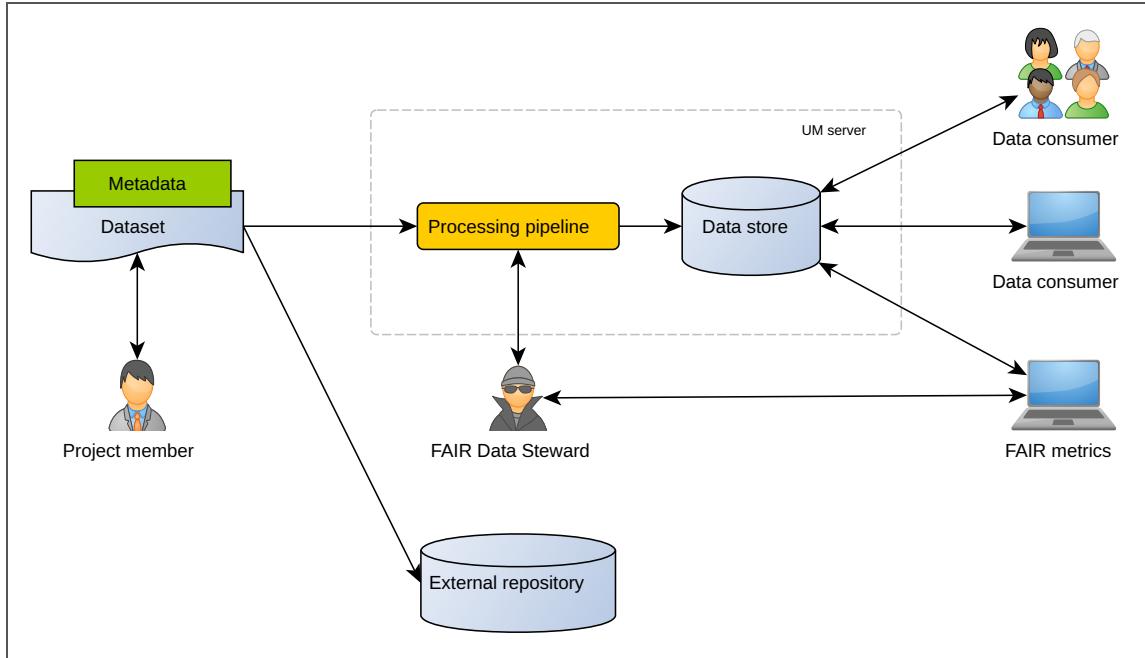
Descripción de la entidad (RDF, HTML, ...) ≠ entidad

HTTP URI dereferenciable: cuando se busca una URI, debería devolver una descripción adecuada del objeto que identifica esa URI

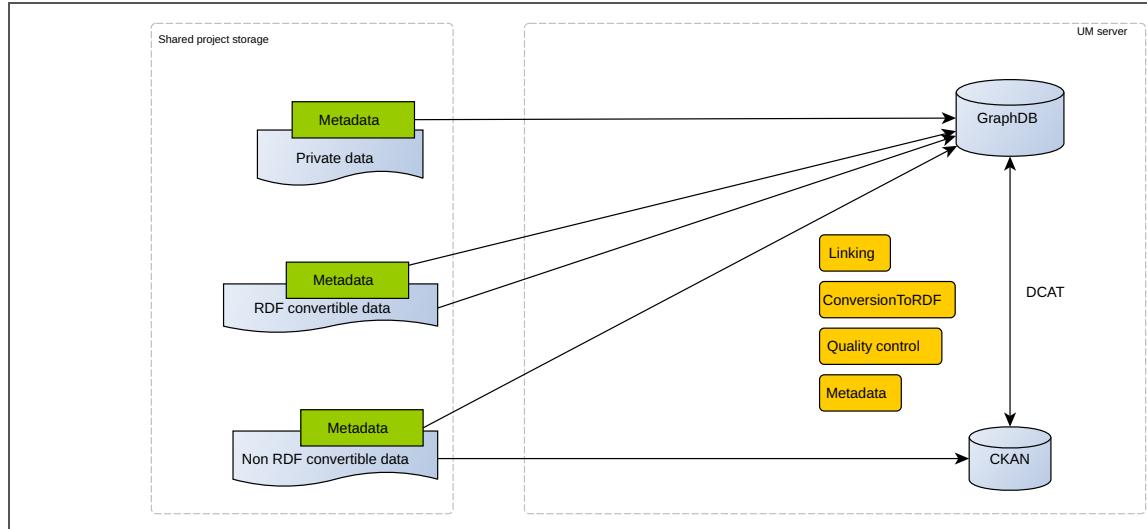
Publicar datos en Linked Data



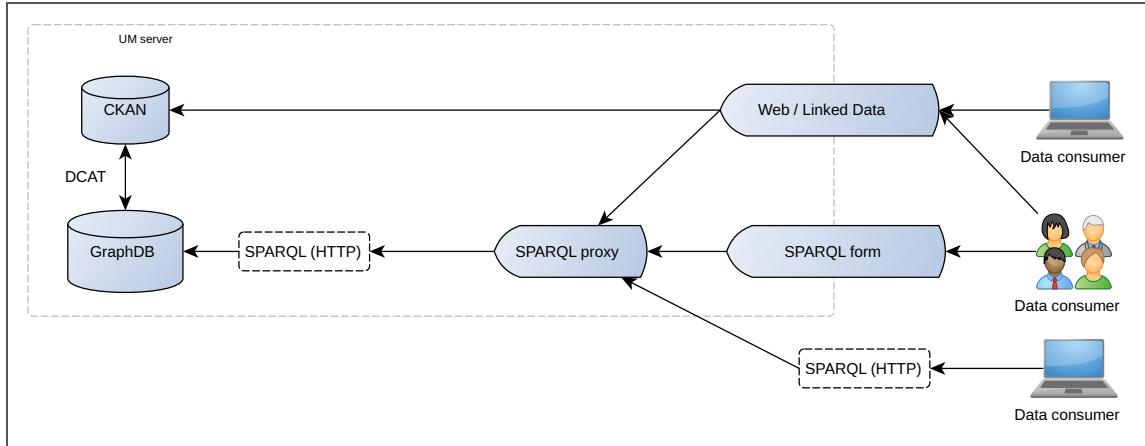
Publicar datos en Linked Data: SUPPORT4LHS



Publicar datos en Linked Data: SUPPORT4LHS



Publicar datos en Linked Data: SUPPORT4LHS



Ejemplo práctico

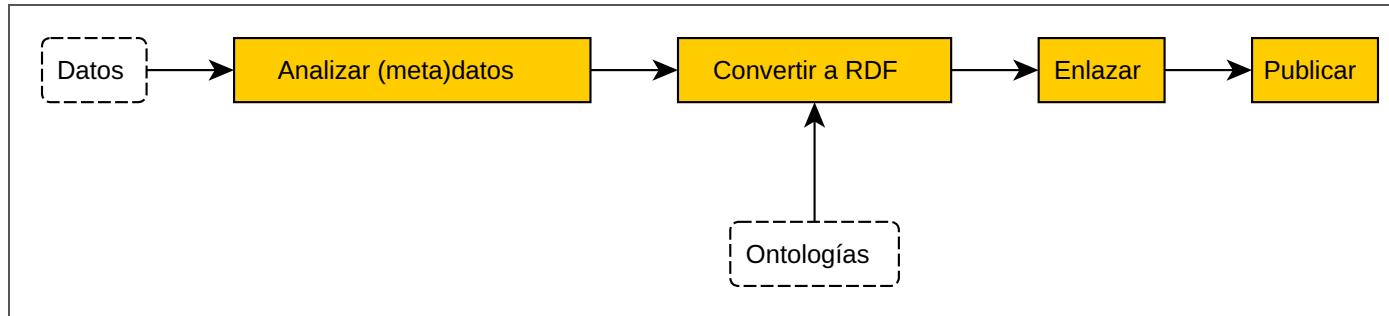
"FAIRificar" un dataset de ejemplo

Proceso vertical: intentar cubrir todos los pasos técnicos, sin entrar en detalles de contenidos

Ejemplos muy simples, nada realistas

A Generic Workflow for the Data FAIRification Process

566-Annikajacobsen-18.pdf



Datos de origen

[GenesUM.csv](#) (LinkedDataServer/data/)

Datos en RDF

[GenesUM.nt](#) (LinkedDataServer/data/)

[GenesUM.ng](#) (LinkedDataServer/data/)

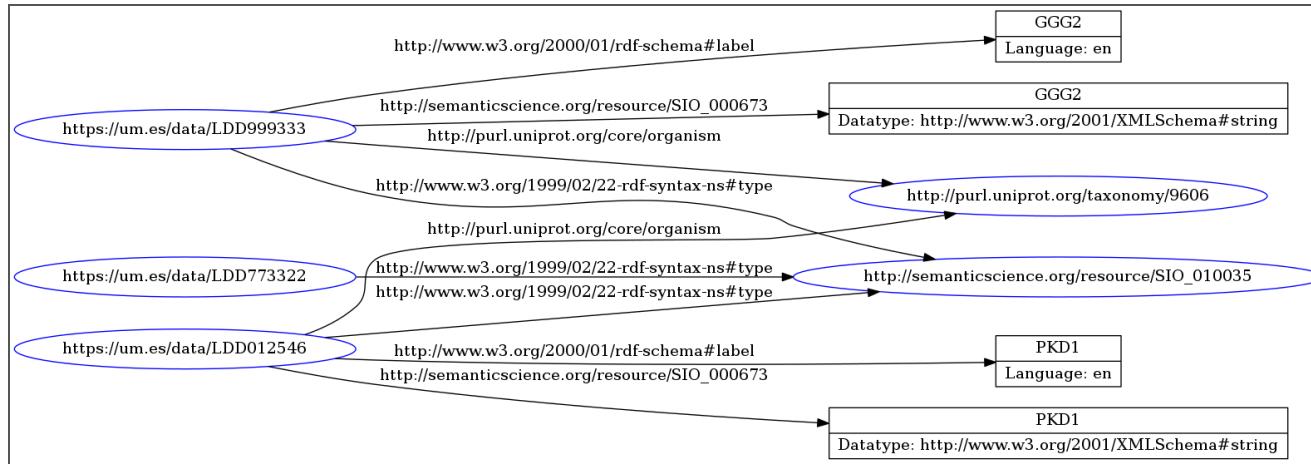
Conversión a RDF

[CSV2RDF.py](#)

Se basa en [RDFLib](#)

Otras herramientas posibles: [YARRRML](#) ([Google Enterprise Knowledge Graph Entity Reconciliation Service](#)), [TARQL](#), [OntoRefine](#), [Open Refine](#), [Eccenca CMEM](#), [Apache Any23](#), etc.

Conversión a RDF



Anotar con ontologías

Semantic Science Integrated Ontology (SIO) ([SIO_010035](#))

Uniprot Core Ontology ([9606](#))

Enlazar

A otras URIs

Manualmente, o con herramientas como [SILK](#)

Metadatos

Asignar una URI a nuestro dataset (F1):

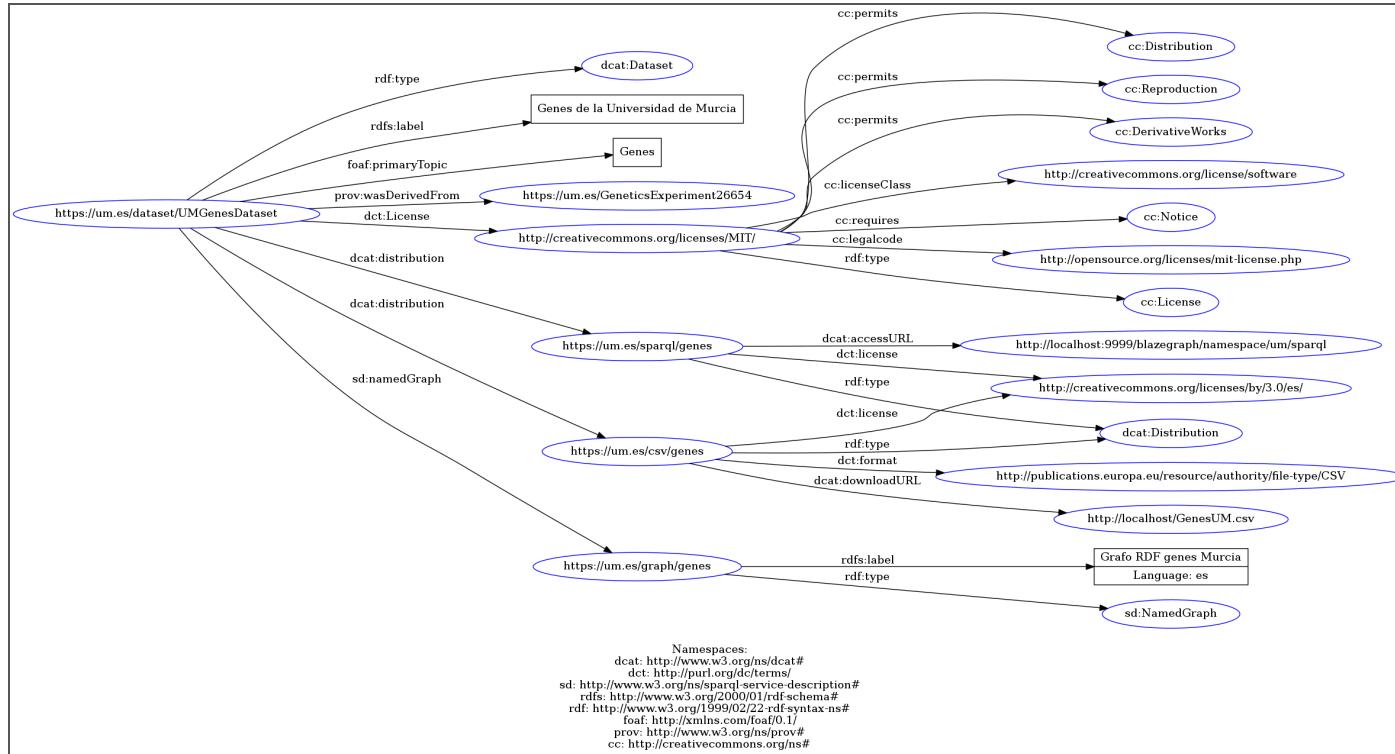
<https://um.es/dataset/UMGenesDataset>

Usar diferentes vocabularios como [DCAT](#), [VOID](#), [PROV](#), [FOAF](#), etc. para añadir metadatos

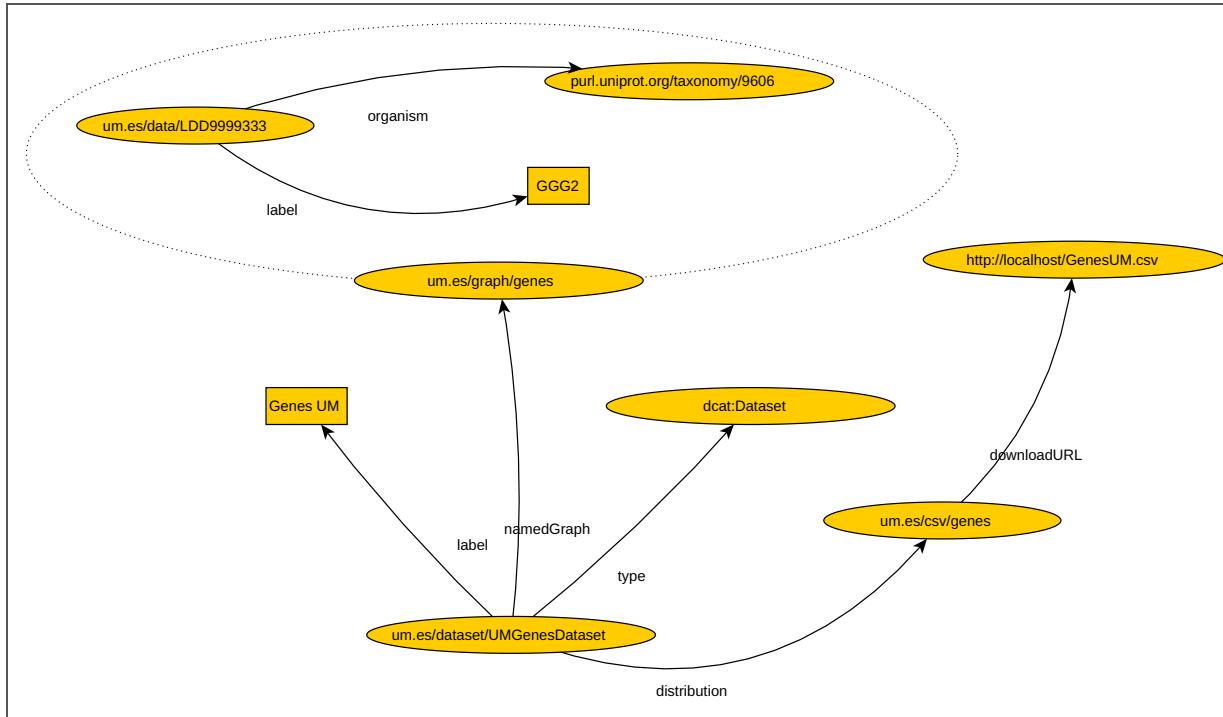
Named Graph: conjunto de triples que se identifica con una URI ([RDF W3C](#))

[MetadataGenes.ttl](#) (LinkedDataServer/data/)

Metadata



Datos/Metadatos



Datos/Metadatos

Datos: CSV (GenesUM.csv), RDF (GenesUM.nq)

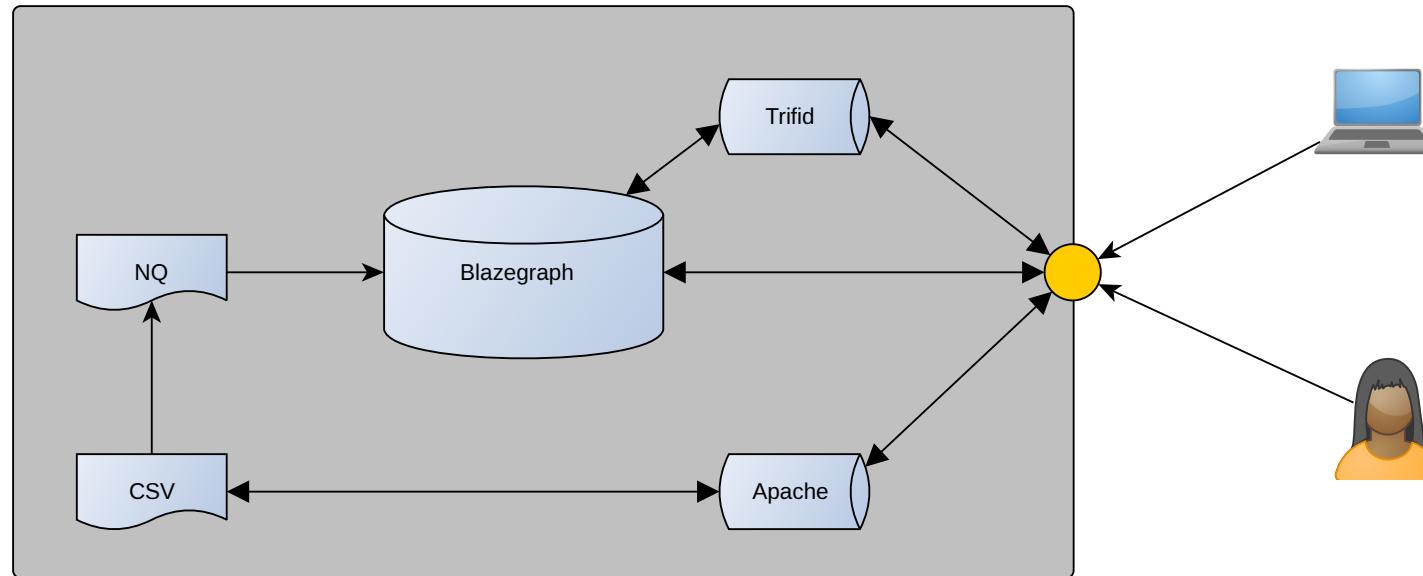
Metadatos: RDF (MetadataGenes.ttl)

Publicación

Datos: CSV (Apache), RDF (Blazegraph/Trifid)

Metadatos: RDF (Blazegraph/Trifid)

Publicación



Blazegraph, Trifid

LinkedDataServer/TrifidBlazegraph: \$ docker-compose up -d

```
version: "3"
services:
  linked_data_server:
    image: ghcr.io/zazuko/trifid
    ports:
      - "8080:8080"
    environment:
      SPARQL_ENDPOINT_URL: "http://sparql_endpoint:9999/blazegraph/namespace/um/sparql"
      #DATASET_BASE_URL: "https://um.es/data/"
      DATASET_BASE_URL: "http://fair/data/"

  sparql_endpoint:
    image: blazegraph
    ports:
      - "9999:9999"
```

Blazegraph

The screenshot shows the Blazegraph web interface with a red ribbon banner at the top right. The main navigation bar includes 'WELCOME', 'QUERY', 'UPDATE', 'EXPLORE', and 'NAMESPACES'. The 'NAMESPACES' tab is active, displaying the 'Namespaces' section. Below this, there's a toolbar with buttons for 'kb', 'Use', 'Delete', 'Properties', 'Rebuild Full Text Index', 'Clone', and 'Service Description'. A link to 'Download VOID description of all namespaces' is also present. The central area is titled 'Create namespace' and contains instructions: 'There are a number of features to enable. There's full documentation [here](#). You must select "Use" after A quick reference is below:'. It lists three options: 'PropertyGraph: Select triples.', 'RDF + SPARQL with named graphs: Select quads mode.', and 'Support for [Reification Done Right \(RDR\)](#): Select rdr mode.' At the bottom, there are input fields for 'Name' (containing 'um'), 'Mode' (set to 'quads'), and 'Inference' (unchecked). There's also a checkbox for 'Enable geospatial' which is also unchecked. A large 'Create namespace' button is at the bottom.

Namespaces

kb Use Delete Properties Rebuild Full Text Index Clone Service Description

[Download VOID description of all namespaces](#)

Create namespace

There are a number of features to enable. There's full documentation [here](#). You must select "Use" after
A quick reference is below:

- o PropertyGraph: Select triples.
- o RDF + SPARQL with named graphs: Select quads mode.
- o Support for [Reification Done Right \(RDR\)](#): Select rdr mode.

Name: Mode: Inference: (Inference disabled -

Enable geospatial:

Trifid

Configurar Trifid para que haga consultas contra SPARQL endpoint de Blazegraph (Servicio sparql_endpoint):

http://sparql_endpoint:9999/blazegraph/namespace/um/sparql

[SPARQL Service Description](#)

[SPARQL Protocol](#)

Trifid

URIs:

- <https://um.es/data/>
- <http://IP:8080>

Apache

/var/www/html/

Resumen

¿Hemos conseguido implementar todos los principios FAIR?

¿En qué medida?

Proyecto a realizar

Reproducir en el servidor Google Cloud este proceso (Con documentación extra)

Incluir en Entregable de Explotación semántica de datos (Publicación de datos)