

# Transformación y enriquecimiento de datos

Adquisición y preparación de datos



Universitat d'Alacant  
Universidad de Alicante

# Contenido

- Web Semántica y Linked Open Data
  - Wikidata
  - Consultas con SPARQL
- Ontologías y vocabularios/taxonomías
  - Schema.org
  - Otros ejemplos
- Transformación de datos
  - Tecnologías
  - Ejemplos
- Reutilización
  - Calidad de datos
  - Visualización
  - Retrieval-Augmented Generation

# Web semántica y Linked Open Data



Slide1

[Source of slides](#)

## The Semantic Web

DAML

2000 August 15

Tim Berners-Lee, Dan Connolly, Lynn Andrea Stein, Ralph Swick

<http://www.w3.org/2000/Talks/0815-daml-sweb-tbl>

This text corresponds approximately to the intent of the [slides](#).

## Semantic Web

The essential property of the World Wide Web is its universality. The power of a hypertext link is that "anything must be able to link to anything". This requires that anything must be able to be put on the web. The web technology, then, must not discriminate between cultures, between media, between the scribbled draft and the polished performance, between commercial and academic information, and so on. Information varies along many axes. One of these is the difference between information produced primarily for human consumption, and that produced primarily for machines. At one end of the scale we have the 5 second TV commercial and poetry. At the other end we have the database. To date, the web has developed most rapidly as a medium of documents for people rather than data which can be processed automatically. In this article we look toward the *Semantic Web*, in which data with well defined meaning is exchanged, and computers and people work side by side in cooperation.

## Weblike

We are looking for a web in which data has processable qualities typical of databases and mathematical formulae, but which is still weblike. Weblike things are, like the Internet, decentralized. They typically involve a lot of fun at every level, and provide benefits on a macroscopic scale which are hard or impossible to predict in advance. To actually build such systems typically requires some compromises, just as the Web initially had to throw away the ideal of total consistency, thus ushering the infamous *Error 404 not found* message but allowing unchecked exponential uncontrolled growth. The semantic web must be weblike in that, to be universal -- to ask any information which may be useful to be easily available -- it must be minimally constraining. Like good government, it must require only that which is essential for everything to work. If you think of the web today as turning all the documents in the world into one big book, then think of the Semantic Web as turning all the data into one big database, or one big mathematical formula. The semantic web is not a separate web - it is a new form of information adding one more dimension to the diversity of the one web we have today.

## First level: RDF

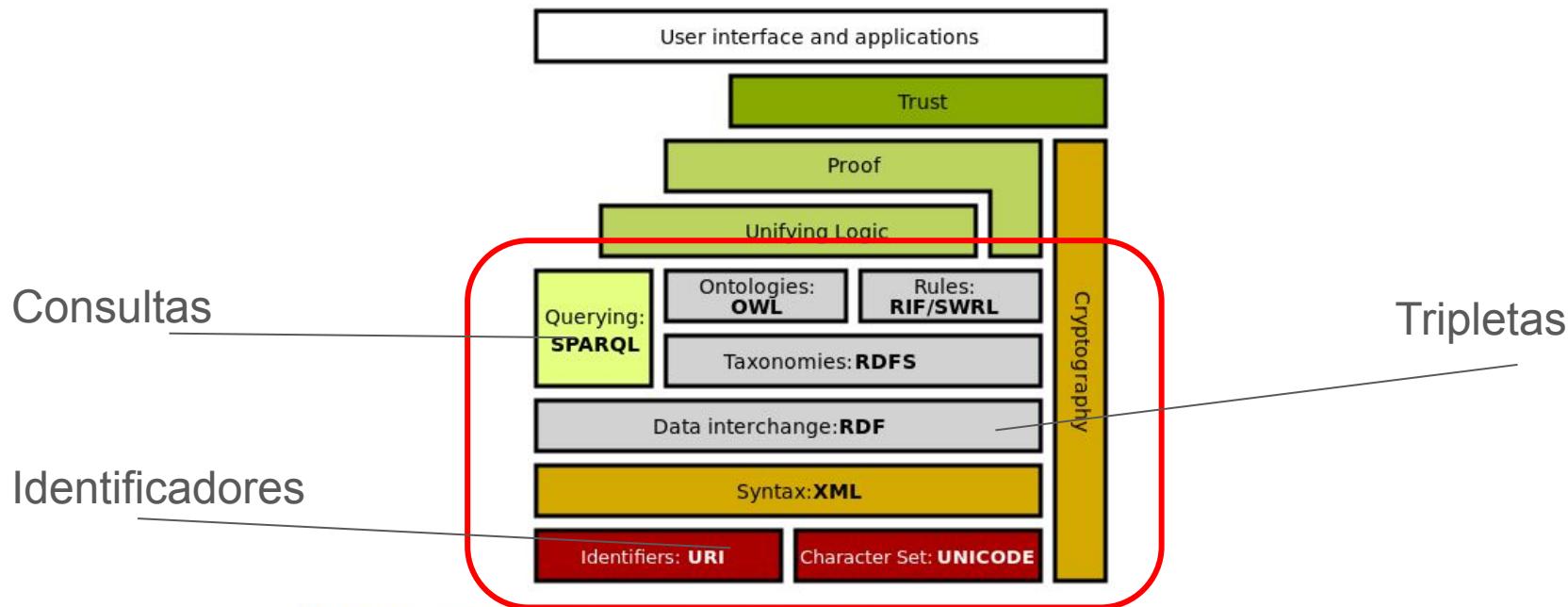
Two important technologies for the semantic web are already in place. The XML language for expressing information gives the community a common way of creating structured documents, though it doesn't have a way of saying in any sense what they mean. However, a starting point for talking about meaning comes from developments in the metadata world of information about information such as library cards and catalogs of web resources. This was the initial impetus behind the development of the Resource Description Framework (RDF) which provides a simple model of conveying the meaning of information as sets of triples rather like the subject, verb and object in a sentence.

Extensión de la Web tradicional para dotarla de significado y pueda ser entendida por máquinas y humanos

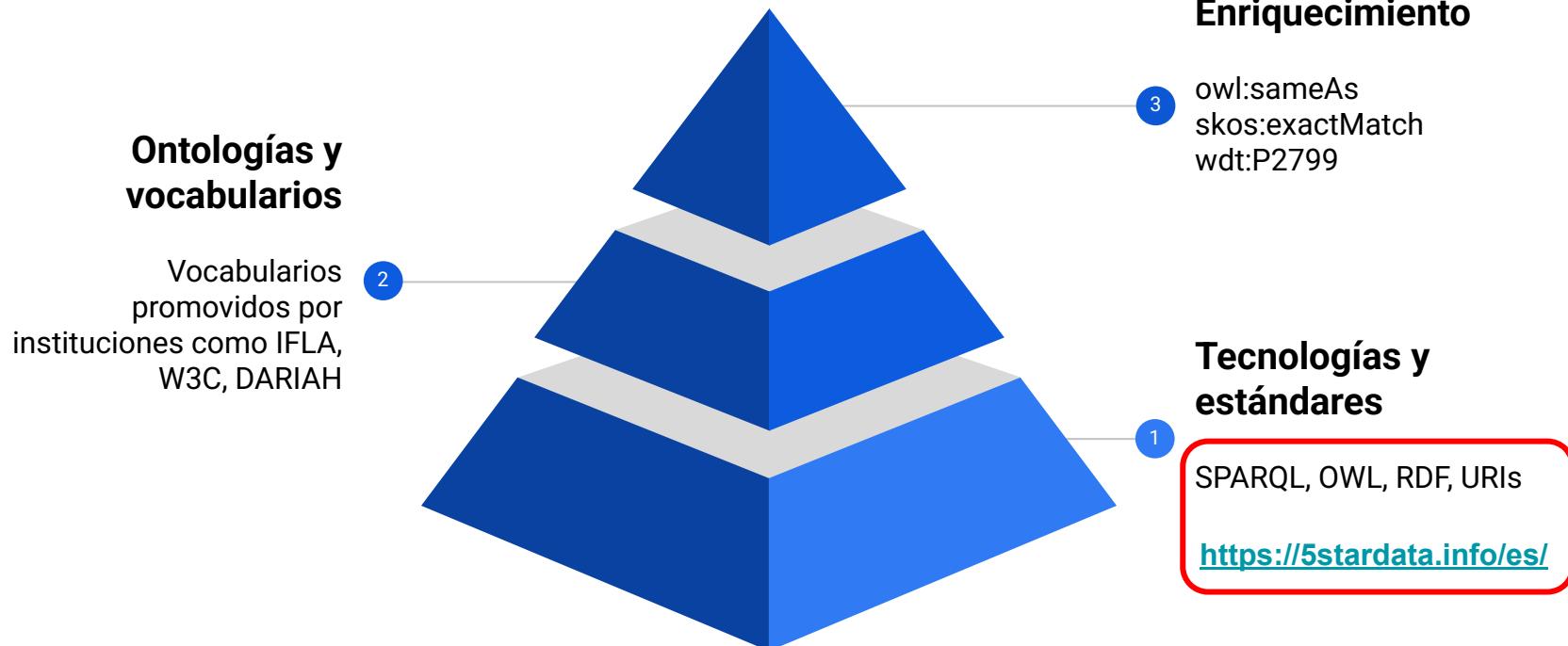
<https://www.w3.org/2000/Talks/0906-xmlweb-tbl/text.htm>

# Web semántica y Linked Open Data

## Tecnologías y estándares



# Web semántica y Linked Open Data



# Web semántica y Linked Open Data

Cuando trabajamos con RDF, la información se almacena en forma de triplets, en forma de Sujeto-Predicado-Objeto.



Sujeto

es\_autor\_de



Objeto

Predicado

# Web semántica y Linked Open Data

Se usan **URIs** para identificar a los recursos y a las propiedades.

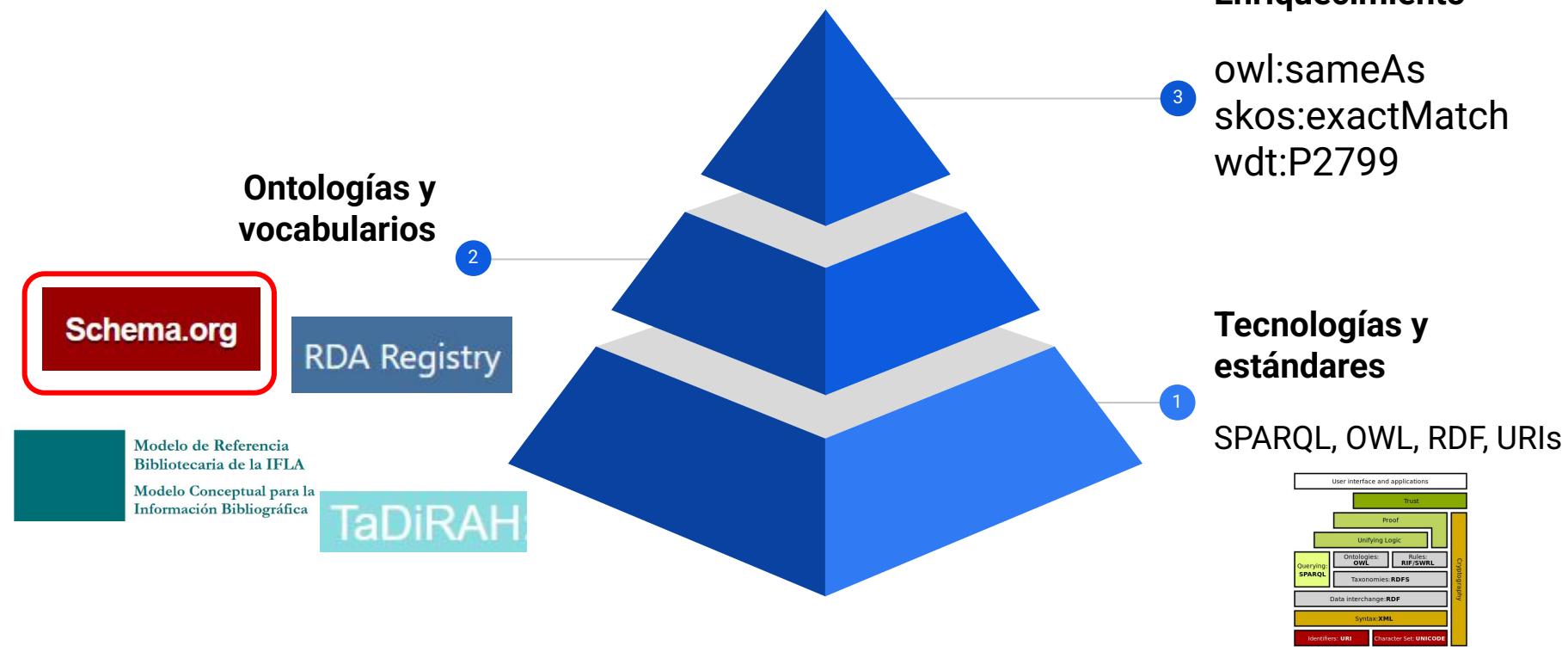


<https://data.cervantesvirtual.com/person/40>



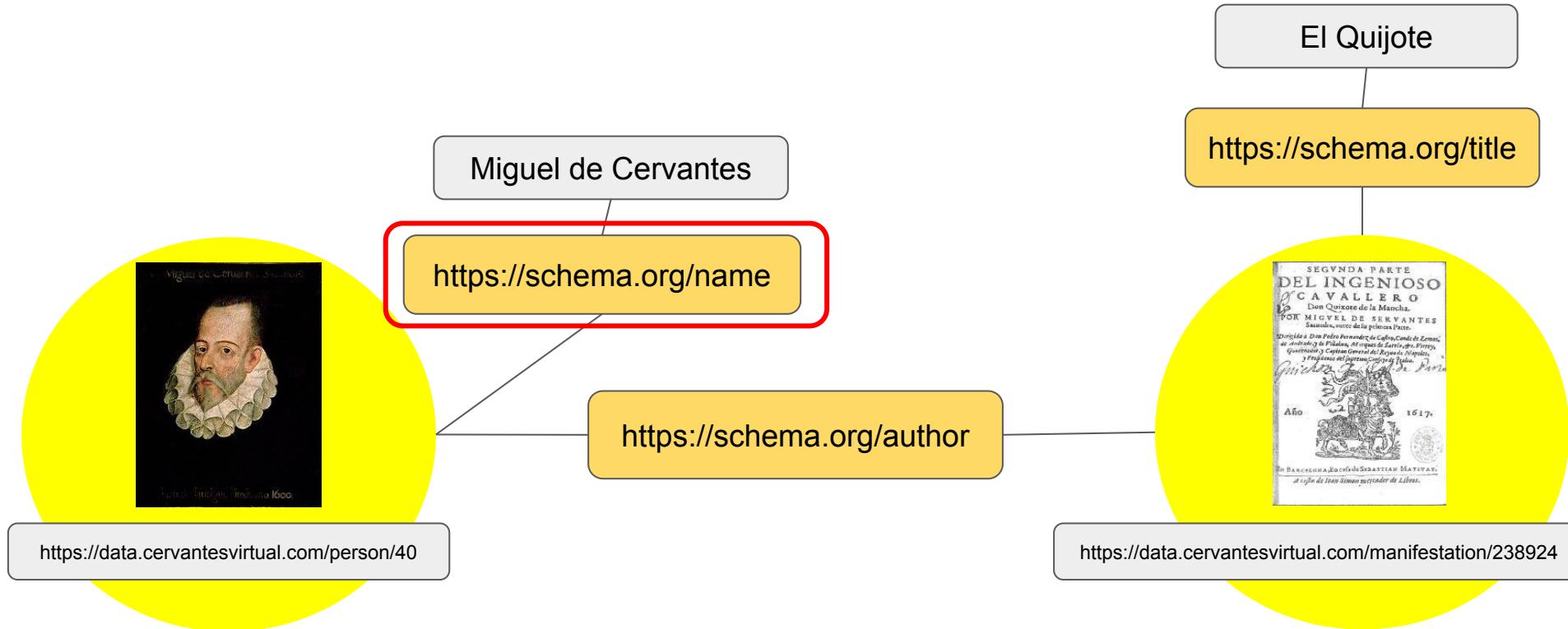
<https://data.cervantesvirtual.com/manifestation/238924>

# Web semántica y Linked Open Data



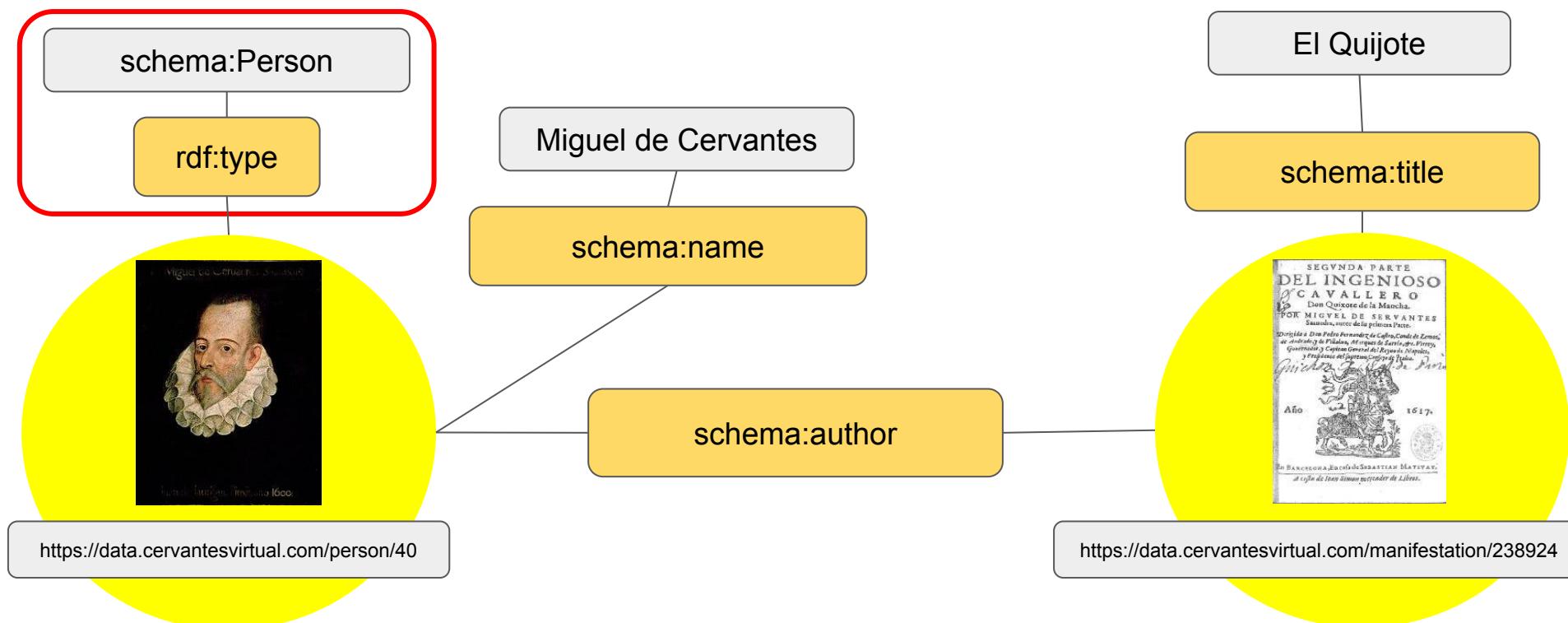
# Web semántica y Linked Open Data

Podemos describir **los recursos** usando ontologías como Schema.org

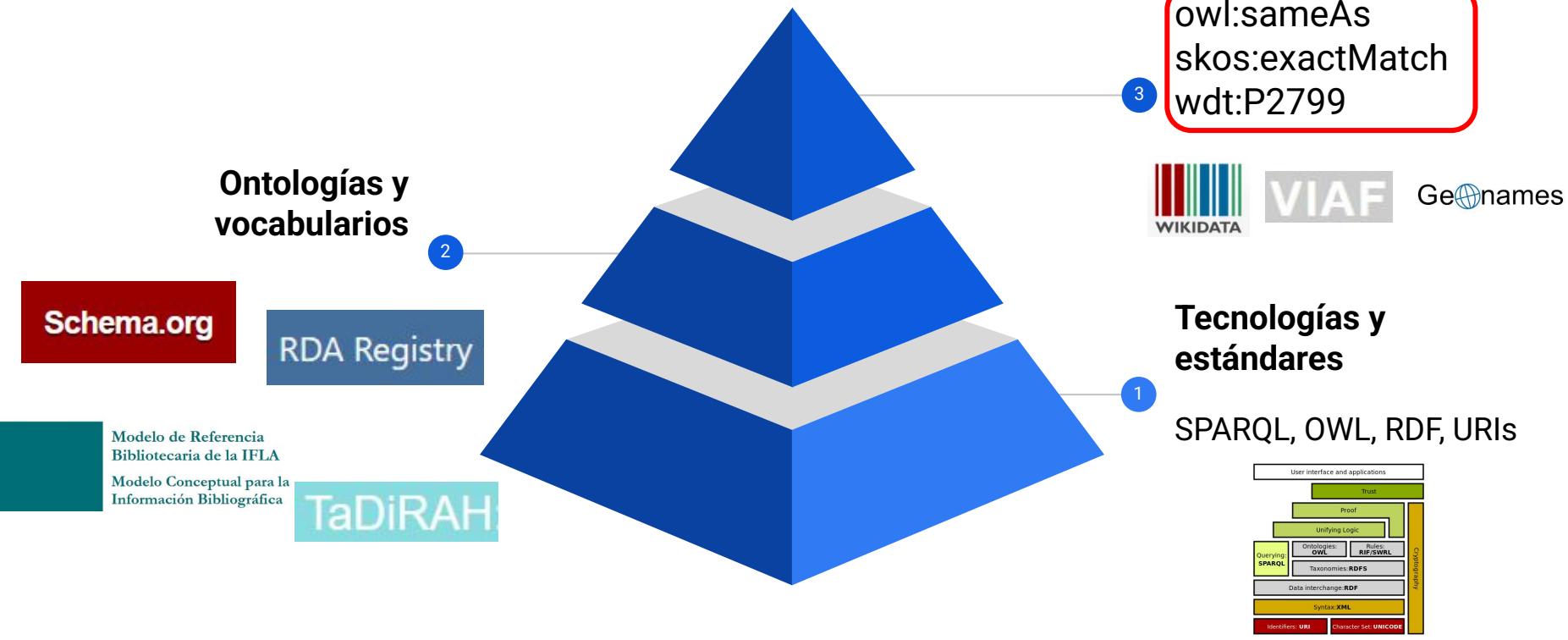


# Web semántica y Linked Open Data

Podemos especificar el tipo de los recursos y usar espacios de nombres (p.ej. schema en lugar de <https://schema.org/>)

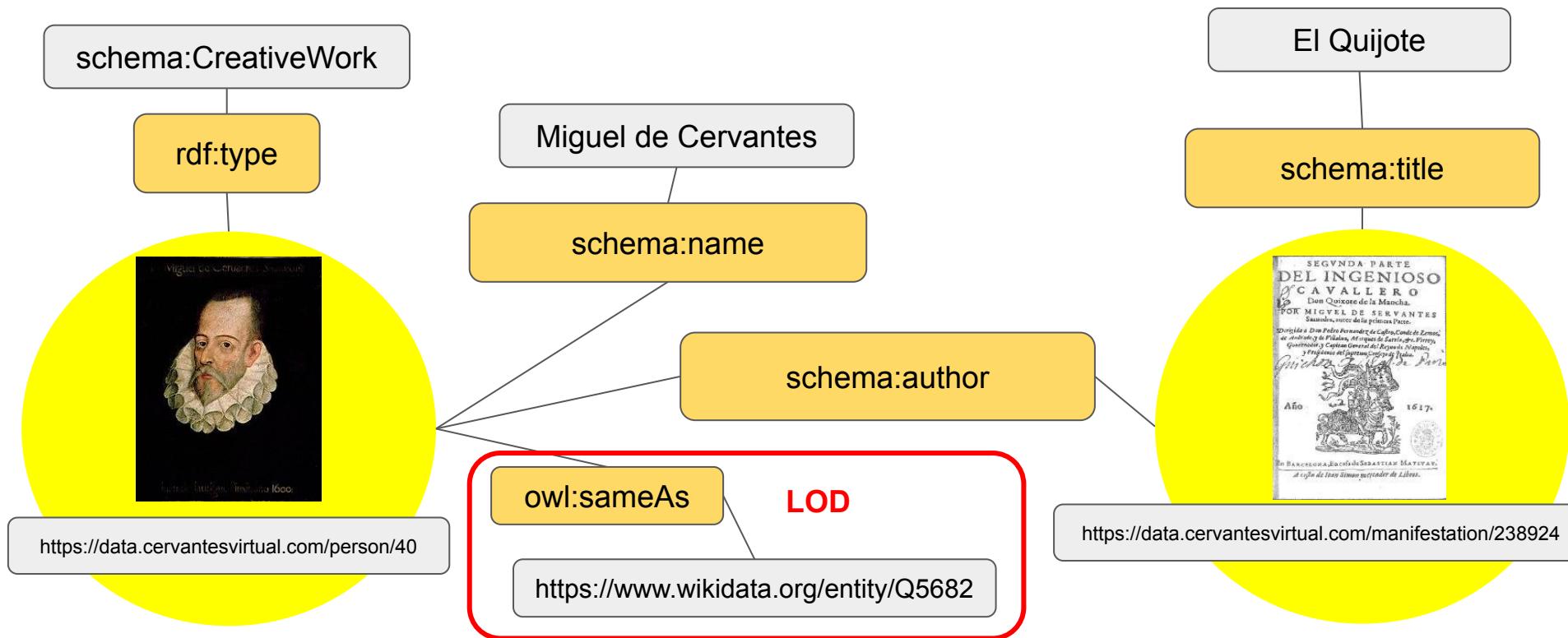


# Web semántica y Linked Open Data



# Web semántica y Linked Open Data

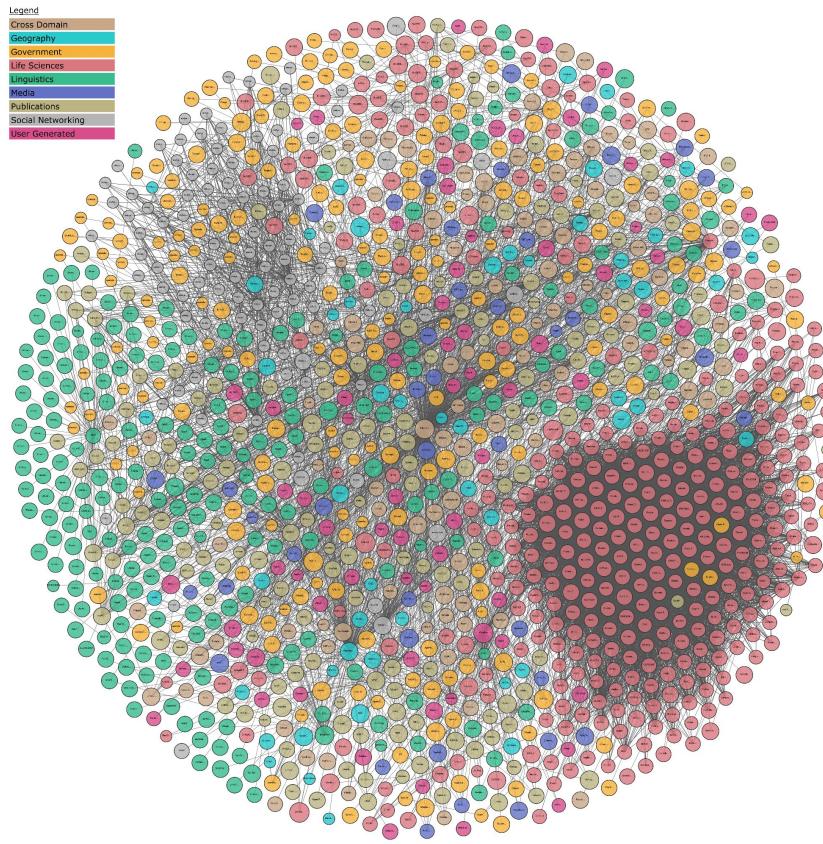
Podemos **enriquecer** los recursos con la propiedad **owl:sameAs**



# Web semántica y Linked Open Data

<https://lod-cloud.net/>

Nov, 2025  
1678 repositorios



# Web semántica y Linked Open Data



<https://doi.org/10.5281/zenodo.8051036>



<https://id.loc.gov>



<https://libris.kb.se/sparql>



<https://americanart.si.edu/about/lod>



<https://data.nationallibrary.fi>



<https://bl.natbib-lod.org/>



<https://data.cervantesvirtual.com/>



Bibliothèque nationale du Luxembourg  
Open Data

<https://data.bnl.lu/>



<https://www.dnb.de/EN/ids>



<https://labs.onb.ac.at/en/dataset/lod>

# Wikidata

Datos estructurados



Basado en estándares

Multilingüe



Dominio público



Editado por la comunidad

# Wikidata



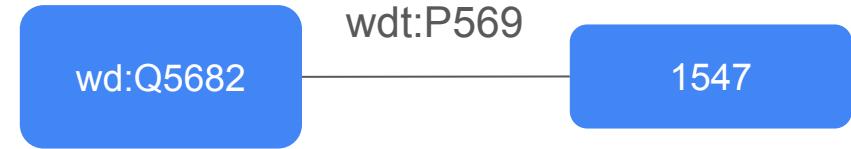
## Representación en forma de tripleta

Statements	
instance of	human + 2 references
image	 Miguel de Cervantes portrait.jpg 1,443 × 1,393; 413 KB + 0 references
signature	 Miguel de Cervantes signature.svg 303 × 214; 11 KB + 1 reference
sex or gender	male + 3 references



Miguel de Cervantes nació en 1547

Sujeto - Predicado - Objeto



<https://www.wikidata.org/wiki/Q5682>

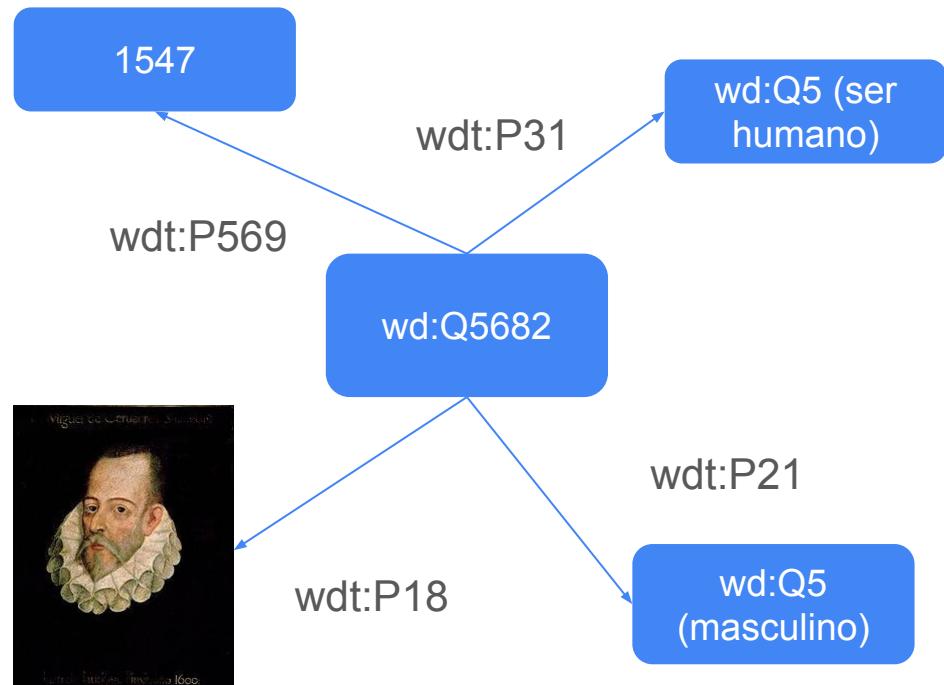
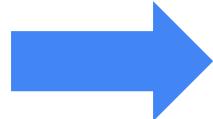
Uso de identificadores únicos para entidades y propiedades



# Wikidata

## Representación en forma de grafo

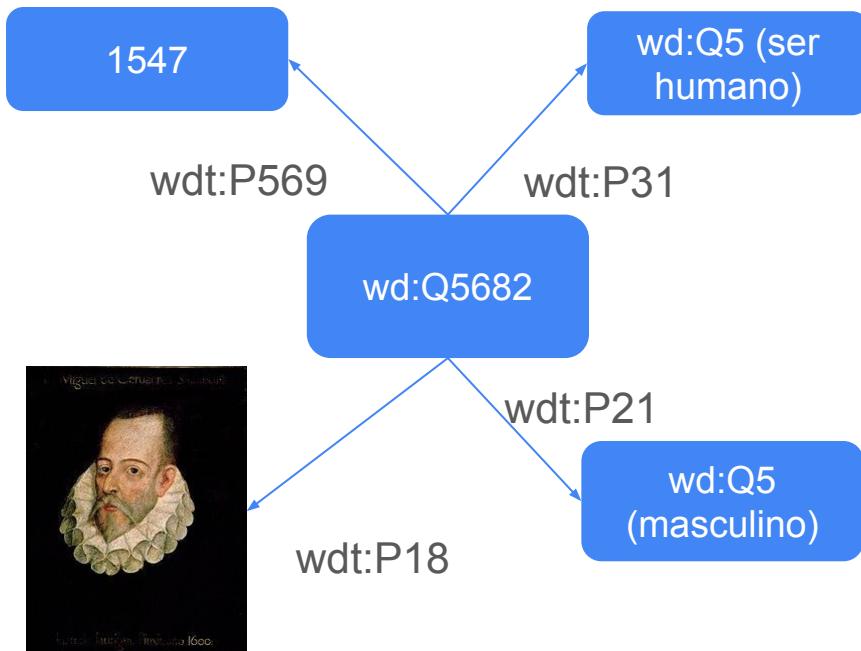
Statements	
instance of	human + 2 references
Image	 Miguel de Cervantes.jpg 1,443 × 1,393; 415 KB + 0 references
signature	 Miguel de Cervantes signature.svg 383 × 216; 11 KB + 1 reference
sex or gender	male + 3 references



<https://www.wikidata.org/wiki/Q5682>

# Wikidata

## Espacios de nombre



`wd` = <http://www.wikidata.org/entity/>  
`wd:Q5682` = <http://www.wikidata.org/entity/Q5682>

`wdt` = <http://www.wikidata.org/prop/direct/>  
`wdt:P31` = <http://www.wikidata.org/prop/direct/P31>

**Podemos usar los prefijos `wd:` y `wdt:` en lugar de la URL completa**

[https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF\\_Dump\\_Format#Full\\_list\\_of\\_prefixes](https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format#Full_list_of_prefixes)

# Consultas con SPARQL

## Introducción

```
PREFIX ex: <example.org>
SELECT *
WHERE {
    ?s ?p ?o
}
```

Espacios de nombre a usar en la sentencia

Selección de columnas. Un \* se refiere a todas

Tripletas para restricciones.  
Las que empiezan con "?" las interpreta como variables

# Consultas con SPARQL

## ¿Qué estructura tienen los datos?

- Revisar las propiedades utilizadas para describir los metadatos
- Obtener las clases y propiedades mediante sentencias SPARQL

```
SELECT DISTINCT ?class
WHERE {
  ?subject a ?class .
  FILTER( REGEX (?class, "rdaregistry"))
}
```

LIMIT 100

<https://data.cervantesvirtual.com/sparql>

<https://datos.bne.es/sparql>

The screenshot shows the Wikidata interface for an entity. At the top right is the WIKIDATA logo with a barcode icon. The entity is identified as a 'statue' (highlighted by a red box) and an 'instance of' (highlighted by a red box). It is part of 'Paseo del Prado and Buen Retiro, a landscape of Arts and Sciences' (highlighted by a red box), which is a 'heritage designation' and a 'UNESCO World Heritage Site'. A thumbnail image of a statue (highlighted by a red box) is shown, labeled 'Monumento a Cervantes (1835) Madrid 03.jpg'. Below the image, its dimensions are listed as 2,848 x 3,520; 1.91 MB. The entity was 'inception' in 1835 (highlighted by a red box). There are sections for '0 references' and '2 references'.

[https://www.wikidata.org/wiki/Wikidata:List\\_of\\_properties](https://www.wikidata.org/wiki/Wikidata:List_of_properties)

# Consultas con SPARQL

## Introducción



?autor devuelve el identificador del recurso en Wikidata

```
SELECT ?autor ?idbvmc  
WHERE {  
    ?autor wdt:P2799 ?idbvmc  
}  
LIMIT 10
```

Propiedad que enlaza autores de la BVMC en Wikidata



<https://w.wiki/88yd>

# Consultas con SPARQL

## Introducción

Si añadimos **Label** al nombre de la variable automáticamente nos recupera la etiqueta del recurso. ?autor devuelve el identificador del recurso

```
SELECT ?autor ?autorLabel ?idbvmc
WHERE {
    ?autor wdt:P2799 ?idbvmc .
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es".
    }
LIMIT 10
```



<https://w.wiki/85wT>

Añadimos esta instrucción para poder tener acceso a las etiquetas de forma más sencilla. AUTO\_LANGUAGE se refiere al idioma del navegador

# Consultas con SPARQL

## Introducción



<https://w.wiki/89LA>

```
SELECT ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
    ?autor wdt:P2799 ?idbvmc .
    ?autor wdt:P569 ?fechaNacimiento
    FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
          && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime).
```

Instrucción  
FILTER permite  
aplicar filtros

```
SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es".
}
LIMIT 10
```

# Consultas con SPARQL

## Uso avanzado

- Operadores Values y Bind
  - asignar uno o más valores específicos a una variable
  - asignar el resultado de una expresión a una variable
- Service
  - consultas federadas entre repositorios de datos SPARQL
- Visualizaciones en Wikidata
  - #defaultView:Map, #defaultView:Graph
- Expresiones regulares
  - FILTER regex(?x, "pattern" [, "flags"])
  - FILTER regex(?nombre, "g", "i"): la variable ?nombre contiene el texto "g" o "G"
- Limit y offset: paginación resultados

[Más ejemplos de SPARQL](#)

# Ontologías y vocabularios

¿Qué es una ontología?

## ontología

Artículo

Del lat. mod. *ontología*, de *onto-* 'onto-' y *-logía* '-logía'.

Sinónimos o afines

1. f. *Fil.* Parte de la metafísica que trata del ser en general y de sus propiedades trascendentales.  
SIN.: [metafísica](#), [filosofía](#).
2. f. En ciencias de la comunicación y en inteligencia artificial, red o sistema de datos que define las relaciones existentes entre los conceptos de un dominio o área del conocimiento.



Diccionario de la  
lengua española

Clases y propiedades para describir los recursos

# Ontologías y vocabularios

Ontologías para modelar y describir la información

**Schema.org**

<https://schema.org/>



<https://www.loc.gov/bibframe/>

 europeana

**EUROPEANA DATA MODEL**

<https://pro.europeana.eu/page/edm-documentation>

**Data Catalog Vocabulary (DCAT)**

<https://www.w3.org/TR/vocab-dcat-3/>

**RDA Registry**

<http://www.rdaregistry.info/>



International Federation of  
Library Associations and Institutions

**IFLA Library  
Reference Model**

<https://repository.ifla.org/handle/123456789/40>

**CIDOC  
CRM**

CONCEPTUAL  
REFERENCE  
MODEL



<https://www.cidoc-crm.org/>

# Ontologías y vocabularios

¿Qué es una vocabulario o taxonomía?

## taxonomía

### Artículo

Del gr. τάξις 'ordenación' y -nomía.

### Sinónimos o afines

1. f. Ciencia que trata de los principios, métodos y fines de la clasificación. Se aplica en particular, dentro jerarquizada y sistemática, con sus nombres, de los grupos de animales y de vegetales.
2. f. **clasificación** (|| acción de clasificar).

SIN.: [clasificación](#), [categorización](#), [sistemática](#).



Diccionario de la  
lengua española

Términos para describir y clasificar los recursos

# Ontologías y vocabularios



skos.um.es/unescod/00/html

## SKOS

### Nomenclatura de Ciencia y Tecnología de la UNESCO

Introducción    [Español](#) [English](#) [Français](#)    Texto a buscar

RDF/XML | N-Triples | N3/Turtle | JSON | JSON-LD

#### Proyecto de nomenclatura internacional normalizada relativa a la ciencia y la tecnología

*Proposed international standard nomenclature for fields of science and technology (en)  
Projet de nomenclature internationale type des domaines de la science et de la technologie (fr)*

##### Conceptos cabecera (skos:topConcept)

- 11 Lógica
- 12 Matemáticas
- 20 Astronomía y Astrofísica
- 22 Física
- 23 Química
- 24 Ciencias de la Vida
- 25 Ciencias de la Tierra y del Espacio
- 31 Ciencias Agrarias
- 32 Ciencias Médicas
- 33 Ciencias Tecnológicas
- 51 Antropología
- 52 Demografía
- 53 Ciencias Económicas
- 54 Geografía
- 55 Historia
- 56 Ciencias Jurídicas y Derecho
- 57 Lingüística
- 58 Pedagogía
- 59 Ciencia Política
- 61 Psicología
- 62 Ciencias de las Artes y las Letras
- 63 Sociología
- 71 Ética
- 72 Filosofía

The LIBRARY of CONGRESS [ASK A LIBRARIAN](#) [DIGITAL COLLECTIONS](#) [LIBRARY CATALOGS](#) Search

The Library of Congress > Cataloging, Acquisitions > Library of Congress Subject Headings PDF Files

### CATALOGING AND ACQUISITIONS

- [Cataloging and Acquisitions Home](#)
- [About the Organization](#)
- [Contact](#)
- [FAQs](#)
- [News](#)
- [Acquisitions](#)
- [Resources for Cataloging](#)
- [Catalogs, Authority Records](#)
- [Classification and Shelflisting](#)
- [Cooperative Cataloging Programs](#)
- [Descriptive Cataloging](#)
- [Products for Purchase](#)
- [Publications, Reports](#)
- [Subject Headings & Genre/Form Terms](#)

About LCSH - Introduction to LCSH - Individual Files - Free-floating Subdivisions - Children's Subject Headings - Genre/Form Terms for Library and Archival Materials - Demographic Group Terms - Medium of Performance Thesaurus for Music -

### Library of Congress Subject Headings PDF Files

About the Library of Congress Subject Headings PDF File

This page provides print-ready PDF files for the 43rd Edition of the Library of Congress Subject Headings (LCSH). Data for the 43rd edition was selected in April 2021. For users desiring enhanced functionality, LCSH will continue as part of the web-based subscription product, Classification Web.

Earlier editions are available [here](#) but should not be used for cataloging.

[Back to Top](#)

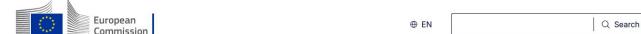
### Introduction to Library of Congress Subject Headings

> [Introduction to Library of Congress Subject Headings \(PDF, 11 p., 337 KB\)](#)

[Back to Top](#)

### Individual PDF Files: By Beginning Letter

<a href="#">Numerals (12 p., 143 KB)</a>	<a href="#">(563 p., 3.14 MB)</a>	<a href="#">(490 p., 2.74 MB)</a>
<a href="#">(842 p., 4.65 MB)</a>	<a href="#">(328 p., 1.86 MB)</a>	<a href="#">(323 p., 1.84 MB)</a>
<a href="#">(322 p., 1.84 MB)</a>	<a href="#">(323 p., 1.83 MB)</a>	<a href="#">(334 p., 1.89 MB)</a>



Live, work, travel in the EU  
Employment, Social Affairs and Inclusion

[Home](#) [Policies and activities](#) [News](#) [Events](#) [Publications](#) [Contact](#) #EUSocialForum

Employment, Social Affairs and Inclusion > ... > Skills for jobs > European Skills/Competences, Qualifications and Occupations (ESCO)

## European Skills/Competences, Qualifications and Occupations (ESCO)

ESCO (European Skills, Competences, Qualifications and Occupations) is the European multilingual classification of skills, competences, qualifications and occupations.

ESCO works like a dictionary, describing, identifying and classifying professional occupations and skills relevant for the EU labour market and education and training area and systematically showing the relations between those occupations and skills. It is available in an online portal where its dataset of occupations and skills can be consulted and downloaded free of charge.

Its common reference terminology helps make the European labour market more effective and integrated, and allows the worlds of work and education/training to communicate more effectively with each other.

What is ESCO?

---

**Schema.org** is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. As of 2024, over 45 million web domains markup their web pages with over 450 billion Schema.org objects. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community process, using the [public-schemaorg@w3.org](mailto:public-schemaorg@w3.org) mailing list and through [GitHub](#).



# Place

A Schema.org Type

Thing > Place

[more...]

Entities that have a somewhat fixed, physical extension.

Property	Expected Type	Description
<strong>Properties from Place</strong>		
<code>additionalProperty</code>	<code>PropertyValue</code>	A property-value pair representing an additional characteristic of the entity, e.g. a product feature or another characteristic for which there is no matching property in schema.org.  Note: Publishers should be aware that applications designed to use specific schema.org properties (e.g. <a href="https://schema.org/width">https://schema.org/width</a> , <a href="https://schema.org/color">https://schema.org/color</a> , <a href="https://schema.org/gtin13">https://schema.org/gtin13</a> , ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism.
<code>address</code>	<code>PostalAddress</code> or <code>Text</code>	Physical address of the item.
<code>aggregateRating</code>	<code>AggregateRating</code>	The overall rating, based on a collection of reviews or ratings, of the item.
<code>amenityFeature</code>	<code>LocationFeatureSpecification</code>	An amenity feature (e.g. a characteristic or service) of the Accommodation. This generic property does not make a statement about whether the feature is included in an offer for the main accommodation or available at extra costs.
<code>branchCode</code>	<code>Text</code>	A short textual code (also called "store code") that uniquely identifies a place of business. The code is typically assigned by the parentOrganization and used in structured URLs.  For example, in the URL <a href="http://www.starbucks.co.uk/store-locator/etc/detail/3047">http://www.starbucks.co.uk/store-locator/etc/detail/3047</a> the code "3047" is a branchCode for a particular branch.
<code>containedInPlace</code>	<code>Place</code>	The basic containment relation between a place and one that contains it. Supersedes <code>containsPlace</code> . Inverse property: <code>containsPlace</code>
<code>containsPlace</code>	<code>Place</code>	The basic containment relation between a place and another that it contains. Inverse property: <code>containedInPlace</code>
<code>event</code>	<code>Event</code>	Upcoming or past event associated with this place, organization, or action. Supersedes <code>events</code> .

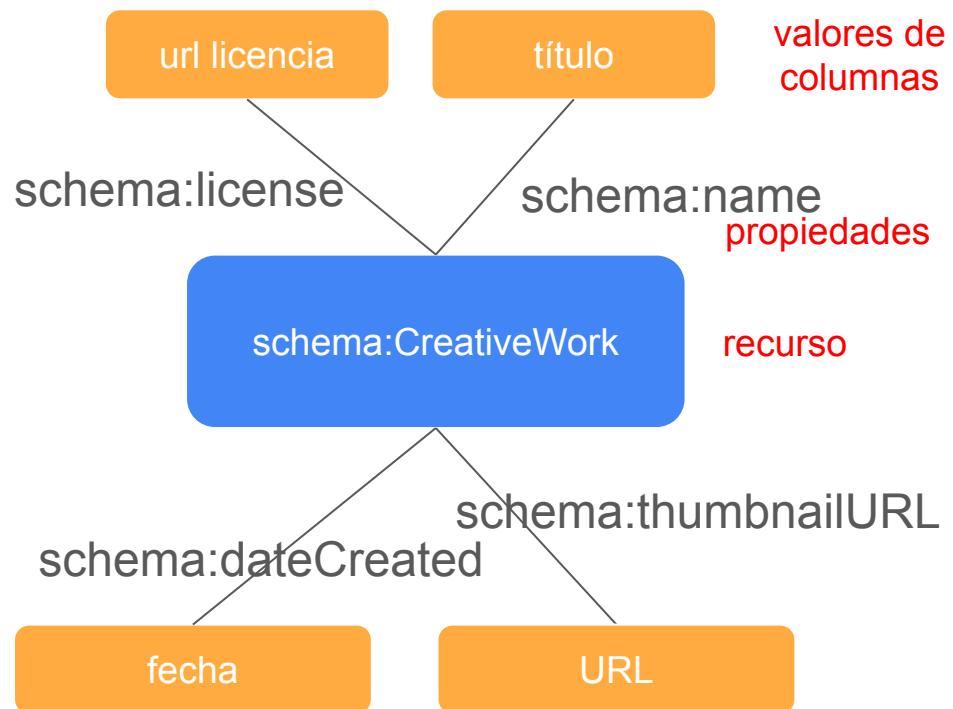
<https://schema.org/Place>

# Ontologías

CSV metadatos:

- título
- licencia
- imagen
- fecha

Schema.org



# Ontologías



## The KnowWhereGraph ontology

<https://knowwheragraph.org/>

At the project's core is the KnowWhere Graph, a geo-knowledge graph that is based on existing standards like RDF, OWL and GeoSPARQL, incorporates custom ontologies, and uses a hierarchical grid for spatial representations. The integrated [KG schema](#) provides a holistic view of the graph modeling. Its current size exceeds 12 billion information triples, and the covered data support pilot scenarios in disaster relief, agricultural land use and food-related supply chains. These data include observations of natural hazards (e.g., hurricanes, wildfires, smoke plumes) and spatial characteristics related to climate (e.g., temperature, precipitation, air quality), soil properties, crop and land-cover types, demographics, and human health, among others. The table below shows a summary of the raw datasets, their providers, their spatiotemporal scopes etc.

Thematic Datasets					Place-Centric Datasets		
Dataset Name/Theme	Source Agency	Key Attributes	Spatial Coverage	Temporal Coverage	Place-Centric Dataset	Defining Authority	Spatial Coverage
Soil Properties	USDA	soil type, farmland class	Targeted regions in US	Current	S2 Cells	Google	Lvl 9 (Global), Lvl 13 (US),
Wildfires	USGS, USDA, USFS, NIFC	wildfire type, burn severity, num. acres burned, contained date	US	1984–current	Global Administrative Regions	University of Berkeley, Museum of Vertebrate Zoology and the International Dino Research	Global
Earthquakes	USGS	magnitude, length, width, geometry	Global (mag. over 4.5)	2011-01-01 to 2022-01-18			

# Ontologías

[GeoNames Home](#) | [Postal Codes](#) | [Download](#) / [Webservice](#) | [About](#)

Alicante   [\[advanced search\]](#)

2802 records found for "Alicante"

	Name	Country	Feature class	Latitude	Longitude
1 ⓘ	<a href="#">Alicante</a> 🌍 A-li-kham-thit,ALC,Akra Leuke,Alacant,Alacante,Alacanti,Alakanto,Alicant,Alicant,Alikante,Alikanteh...	<a href="#">Spain</a> , Valencia Alicante > Alicante	seat of a second-order administrative division population 348,901	N 38° 20' 42"	W 0° 28' 53"
2 ⓘ	<a href="#">Elche</a> 🌍 Eh'ch,Eh'l'che,El'che,Elce,Elch,Elche,Elig,Eix,Elčé,Ilici,ai er qie,alchh,alsh,elche,eruche,xelche,...	<a href="#">Spain</a> , Valencia Alicante > Elche	seat of a third-order administrative division population 234,765, elevation 96m	N 38° 15' 43"	W 0° 42' 3"
3 ⓘ	<a href="#">Alicante-Elche Airport</a> 🌍 ALC,Aeroport d'Alicante-Elche,Aeroporto di Alicante-Elche,Aeropuerto Internacional de Alicante - El ...	<a href="#">Spain</a> , Valencia Alicante > Elche	airport elevation 43m	N 38° 16' 55"	W 0° 33' 29"
4 ⓘ	<a href="#">Alicante Canyon</a>		canyon	N 37° 59' 0"	E 0° 6' 0"
5 ⓘ	<a href="#">Orihuela</a> 🌍 Auraiola,Aurariola,Orihuela,Oriola,Oriouela,Oriuebla,Oriuela,Orivela,ao li wei la,aryhwyla,awrywyla,...	<a href="#">Spain</a> , Valencia Alicante > Orihuela	seat of a third-order administrative division population 101,321	N 38° 5' 5"	W 0° 56' 38"
6 ⓘ	<a href="#">Torrevieja</a> 🌍 Malinovturo,Toreviekha,Toreviena,Torevijekha,Torrev'ekha,Torrevekha,Torrevella,Torrevella de la Mata,...	<a href="#">Spain</a> , Valencia Alicante > Torrevieja	seat of a third-order administrative division population 82,599	N 37° 58' 43"	W 0° 40' 56"
7 ⓘ	<a href="#">Benidorm</a> 🌍 Benidorm,Benidormo,bei ni duo er mu,benidoleum,benidorumu,bnydwrn,Бенидорн,Бенідорн,ବେନିଦର୍ମ,...	<a href="#">Spain</a> , Valencia Alicante > Benidorm	seat of a third-order administrative division population 70,450	N 38° 32' 17"	W 0° 7' 51"
8 ⓘ	<a href="#">Alcoy</a> Al'ka,Al'koj,Alco,Alcodium,Alcoi,Alcoy,Alkoj,Alkojus,a er ke yi,alkwy,arukoi,Alkoj,Алька,Алькој,...	<a href="#">Spain</a> , Valencia Alicante > Alcoy	seat of a third-order administrative division population 61,552	N 38° 42' 19"	W 0° 28' 27"

Genames

<https://www.geonames.org/>

# Ontologías

Creando nuestra propia ontología....

The image shows the homepage of the Protégé website. At the top, there is a navigation bar with the Protégé logo (a stylized blue and purple geometric shape), followed by the word "Protégé". To the right of the logo are four menu items: "SOFTWARE", "SUPPORT", "COMMUNITY", and "ABOUT". Below the navigation bar is a large, dark purple banner. On this banner, the text "A free, open-source ontology editor and framework for building intelligent systems" is displayed in a large, white, sans-serif font. Below the banner, there is a paragraph of text: "Protégé is supported by a strong community of academic, government, and corporate users, who use Protégé to build knowledge-based solutions in areas as diverse as biomedicine, e-commerce, and organizational modeling." At the bottom of the page, there are two call-to-action buttons: a red button on the left labeled "DOWNLOAD NOW" and a white button on the right labeled "USE WEBPROTÉGÉ".

A free, open-source ontology editor  
and framework for building  
intelligent systems

Protégé is supported by a strong community of academic,  
government, and corporate users, who use Protégé to build  
knowledge-based solutions in areas as diverse as  
biomedicine, e-commerce, and organizational modeling.

DOWNLOAD NOW

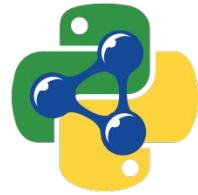
USE WEBPROTÉGÉ

<https://protege.stanford.edu/>

# Transformación

## Tecnologías

- [Apache Jena](#) (RDF Java framework)
- [Virtuoso open-source edition](#)
- [Spring Boot application](#)
- [Apache Maven](#)
- [RDFLib](#)
- [Jupyter](#)
- [Neo4J](#)



# Transformación

En 2015 la BVMC publicó su catálogo como datos abiertos y enlazados

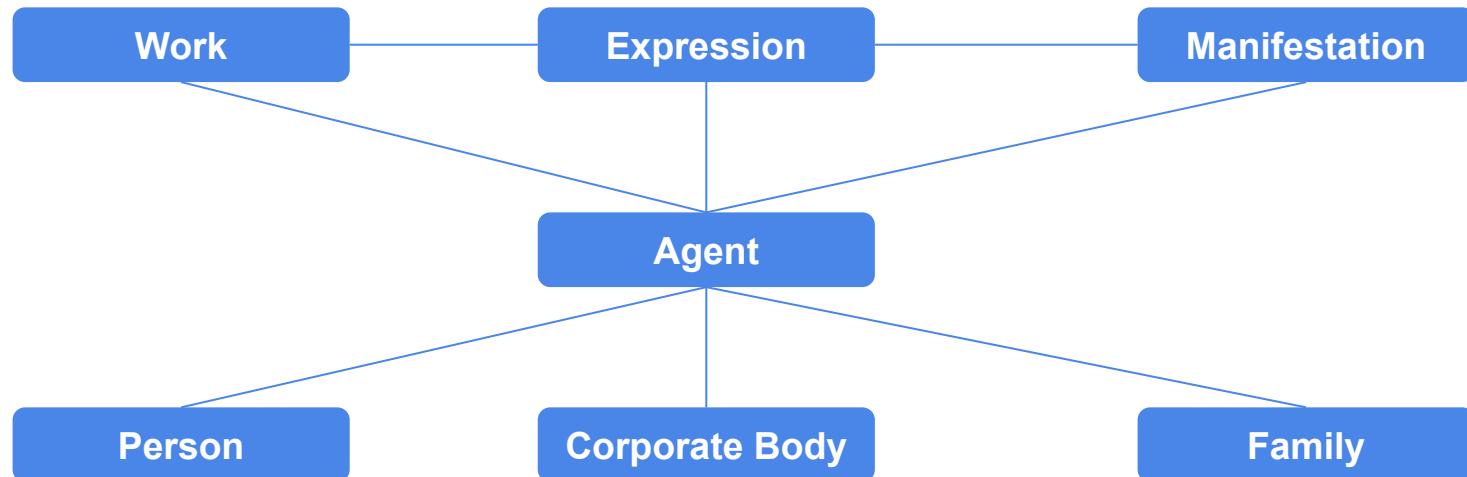
- Modelado de datos basado en RDF y [\*\*RDA registry\*\*](#)
- Enriquecimiento con repositorios externos como **Wikidata** y **GeoNames**
- Disponible a través de un punto de acceso [\*\*SPARQL\*\*](#)
- Candela, G., Escobar, P., Carrasco, R. and Marco-Such, M. Migration of a library catalogue into RDA linked open data.

Semantic Web 9(4): 481-491 (2018). <http://hdl.handle.net/10045/65427>



# Transformación

Modelado de datos



# Transformación

Varios puntos de acceso (URIs) usados como identificadores únicos:

- <https://data.cervantesvirtual.com/date/1616>
- <https://data.cervantesvirtual.com/language/es>
- <https://data.cervantesvirtual.com/manifestation/722875>
- <https://data.cervantesvirtual.com/person/40>
- <https://data.cervantesvirtual.com/corporatebody/7802>
- <https://data.cervantesvirtual.com/work/2904>
- <https://data.cervantesvirtual.com/expression/152342>
- <https://data.cervantesvirtual.com/item/134373>

Nuevas formas de  
acceso al catálogo

# Transformación

Wikidata y la Biblioteca Virtual Miguel de Cervantes

Propiedad	Descripción	Elementos
<a href="#"><u>P2799</u></a>	BVMC person ID	<a href="#"><u>15054</u></a>
<a href="#"><u>P3976</u></a>	BVMC work ID	<a href="#"><u>764</u></a>
<a href="#"><u>P10834</u></a>	BVMC organization ID	<a href="#"><u>240</u></a>



[https://www.wikidata.org/wiki/Wikidata:Property\\_proposal](https://www.wikidata.org/wiki/Wikidata:Property_proposal)

# Transformación

## Wikidata y la Biblioteca Virtual Miguel de Cervantes

Vega, Lope de, 1562-1635



sacerdote, poeta y dramaturgo español del Siglo de Oro  
Información extraída de [Wikidata](#) / [\(CC-BY-SA 3.0\)](#)

**Identificador:** 72  
**Nombre:** Vega, Lope de  
**Fecha de nacimiento:** [1562](#)  
**Fecha de fallecimiento:** [1635](#)  
**VIAF:** <https://viaf.org/viaf/89773778>

**Variantes de nombre:**

- Burguillos, Tomé de
- Carrera, Luis de la
- Flórez, Antonio
- Lope de Vega Carpio, Félix
- Padecopio, Gabriel
- Puente, Juan de la
- Un Ingenio de esta Corte
- Vega Carpio, Lope de
- Vega y Carpio, Lope de
- Lope de Vega

**Lugar de nacimiento:** [Madrid](#)  
**Lugar de fallecimiento:** [Madrid](#)  
**Nacionalidad:** España  
**Nombre de pila:** Lope  
**Género:** masculino

**EXPORTAR:**  
[RDF](#) [JSON](#)

WWW

[Sociedad de América](#)

**Cervantes : Bulletin of the Cervantes Society of America**

**PUBLICACIÓN PERIÓDICA**

**Título:** [Cervantes : Bulletin of the Cervantes Society of America - Registro bibliográfico](#)

**Portales:** Literatura [\(+\) Miguel de Cervantes](#) [\(+\) Cervantes : Bulletin of the Cervantes Society of America](#)

**Mat.\_aut.:** Cervantes Saavedra, Miguel de (1547-1616) -- Crítica e interpretación [\(+\) Cervantes Saavedra, Miguel de, 1547-1616 -- Publicaciones periódicas](#)

**Fondo:** [\(+\) tomos](#)

The [Cervantes](#) [...]

[...] Society of America Bulletin of the [Cervantes](#) Society of America Página personal de Daniel Eisenberg et [...]

[...] Daniel Eisenberg dentro del portal temático Figuras del Hispanismo de la Biblioteca Virtual [Miguel de Cervantes](#) [...]

[...] de [Cervantes](#) Anuario Bibliográfico Cervantino H-Cervantes IV Centenario del Quijote Instituto [Cervantes](#) et [...]

[...] Centro de Estudios Cervantinos Proyecto [Cervantes](#) 2001 et [...]

[Ver más](#)



[Ver más]



[Ver más]

<https://data.cervantesvirtual.com/person/72>

<https://www.cervantesvirtual.com/buscador/?q=miguel+de+cervantes>

# Transformación - Java y Jena

## Proyecto biblioteca de Spring Boot en version 3.4.2

Este proyecto se ha creado para la asignatura [Tecnologías de Internet Orientadas al Navegador del Máster Universitario en Desarrollo de Aplicaciones y Servicios Web de la Universidad de Alicante](#).

Para poder editar el código es posible instalar un IDE como [Eclipse](#) o [Idea](#).

### Configuración

En el fichero `src/main/resources/application.properties` se puede configurar el acceso a la base de datos mysql para el almacenamiento de los libros del catálogo. También se puede configurar el puerto para ejecutar la aplicación.

El proyecto incluye la [librería Jena](#) como dependencia en el fichero `pom.xml` para el uso de SPARQL.

El fichero `pom.xml` incluye las versiones de las librerías utilizadas en este proyecto.

### API REST

El proyecto proporciona los siguientes patrones de URL para dar acceso al API REST:

- Crear libro (POST): `localhost:8081/api/book`
- Modificar libro (UPDATE): `localhost:8081/api/book/{ID}`
- Eliminar libro (DELETE): `localhost:8081/api/book/{ID}`
- Recuperar libro (GET): `localhost:8081/api/book/{ID}`

Para poder realizar pruebas se puede utilizar la herramienta [Postman](#)



```
<dependencies>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-data-jpa</artifactId>
    </dependency>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-web</artifactId>
    </dependency>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-thymeleaf</artifactId>
    </dependency>

    <!-- https://mvnrepository.com/artifact/com.itextpdf/itextpdf -->
    <dependency>
        <groupId>com.itextpdf</groupId>
        <artifactId>itextpdf</artifactId>
        <version>5.5.13.4</version>
    </dependency>

    <!-- java version 4.1.0 Java 11 -->
    <dependency>
        <groupId>org.apache.jena</groupId>
        <artifactId>apache-jena-libs</artifactId>
        <version>5.3.0</version>
        <type>pom</type>
    </dependency>
```

<https://github.com/hibernator11/biblioteca-springboot>



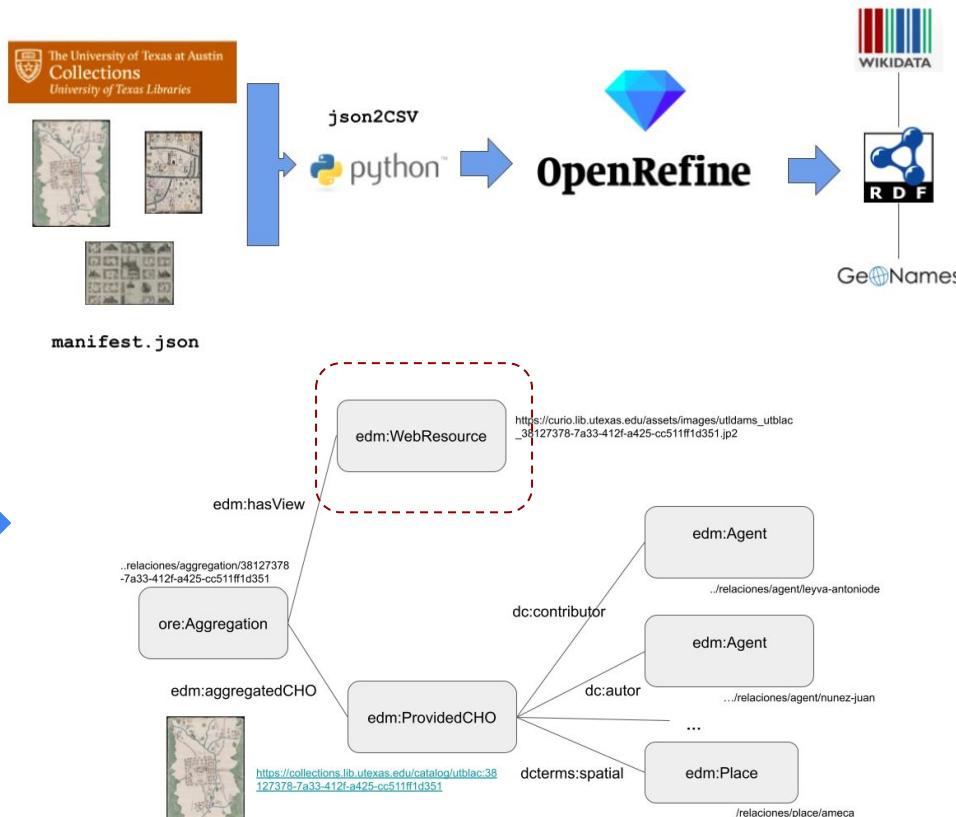
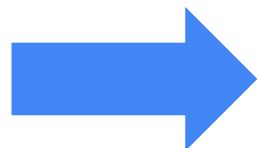
# Transformación

## Unlocking the Colonial Archive



Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2023. An ontological approach for unlocking the Colonial Archive. *J. Comput. Cult. Herit.* Just Accepted (April 2023).

<https://doi.org/10.1145/3594727>



<https://github.com/hibernator11/UCA-relacionesgeograficas>

# Transformación

## Datos originales



`manifest.json`

## Extracción

`json2csv`



`JSON`

## Modelado



**OpenRefine**

`Schema.org`

`CSV`

## Enriquecimiento



`Geonames`

`RDF`

<https://github.com/hibernator11/uned-unlocking-workshop>



A power tool for working with messy data.

New Version Download

Create project

Open project

Import project

Language settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data do

Get data from

This Computer

Locate one or more files on your computer to upload

Examinar... No se han seleccionado archivos.

Next »

Clipboard

OpenRefine

A power tool for working with messy data.

New version

Download OpenRefine v5.0.1 now.

Create project

Open project

Import project

Language settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats

Get data from

This Computer

Enter one or more web addresses (URLs) pointing to data to download:

`https://raw.githubusercontent.com/hibernator11/uned-unlocking-workshop/main/output/metadatos.csv`

Web Addresses (URLs)

Add another URL

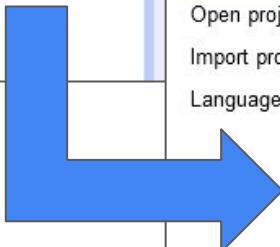
Next »

Clipboard

Database

Google Data

# Abrimos el fichero CSV



[Create project](#)
[Open project](#)
[Import project](#)
[Language settings](#)

« start over Configure parsing options

Project name UNED metadatos csv Tags [Create project »](#)

título	autores	fecha	lugar	coordenadas	thumbnail	licencia	urimanifest
1. Pintura de Ameca	unknown (artist) Leyva, Antonio de (contributor) Moras, Pedro de (scribe) Bejarano, Pedro (signer) Núñez, Juan (signer) Cortes, Martín	1579-10-02	MX  Diocese of Guadalajara Ameca, Jalisco, Mexico	20.547647, -104.04662	<a href="https://curio.lib.utexas.edu/assets/DAMS/utblac_38127378-7a33-412f-a425-cc51ff1d351/TN/utblac_38127378-7a33-412f-a425-cc51ff1d351.jpg">https://curio.lib.utexas.edu/assets/DAMS/utblac_38127378-7a33-412f-a425-cc51ff1d351/TN/utblac_38127378-7a33-412f-a425-cc51ff1d351.jpg</a>	<a href="https://creativecommons.org/publicdomain/mark/1.0/">https://creativecommons.org/publicdomain/mark/1.0/</a>	<a href="https://curio.lib.utexas.edu/assets/DAMS/utblac/manifests/2/utblac_38127378-7a33-412f-a425-cc51ff1d351.json">https://curio.lib.utexas.edu/assets/DAMS/utblac/manifests/2/utblac_38127378-7a33-412f-a425-cc51ff1d351.json</a>

Parse data as Character encoding

[Update preview](#) [Disable auto preview](#)

**CSV / TSV / separator-based files**

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

Columns are separated by  commas (CSV)  tabs (TSV)  custom ,

Use character " " to enclose cells containing column separators  Trim leading & trailing whitespace from strings Escape special characters with \

Ignore first 0 line(s) at beginning of file  Attempt to parse cell text into numbers

Parse next 1 line(s) as column headers  Store blank rows  Store blank cells as nulls

Column names (comma separated)  Store file source  Store archive file

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

# Le damos un nombre al proyecto

## Le damos un nombre y creamos el proyecto desde el menú de la parte superior



Version 3.8.0 [ff7de4d]

[Preferences](#)

**OpenRefine UNED metadatos csv** [Permalink](#)

Facet / Filter [Undo / Redo](#) 0 / 0

**10 rows**

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

	<input type="checkbox"/> All	<input type="checkbox"/> título	<input type="checkbox"/> autores	<input type="checkbox"/> fecha	<input type="checkbox"/> lugar	<input type="checkbox"/> coordenadas	<input type="checkbox"/> thumbnail	<input type="checkbox"/> licencia	<input type="checkbox"/> urlmanifest
1.		Pintura de Ameca	unknown (artist) Leyva, Antonio de (contributor) Moras, Pedro de (scribe) Bejarano, Pedro (signer) Nuñez, Juan (signer) Cortes, Martin (signer) Mesa, Francisco de (signer) Vázquez, Juan (signer) García Icazbalceta, Joaquín, 1825-1894 (collector)	1579-10-02	MX  Diocese of Guadalajara  Ameca, Jalisco, Mexico	20.547647, -104.04662	<a href="https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_38127378-7a33-412f-a425-cc511ff1d351/THN/utblac_38127378-7a33-412f-a425-cc511ff1d351.jpg">https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_38127378-7a33-412f-a425-cc511ff1d351/THN/utblac_38127378-7a33-412f-a425-cc511ff1d351.jpg</a>	<a href="https://creativecommons.org/publicdomain/mark/1.0/">https://creativecommons.org/publicdomain/mark/1.0/</a>	<a href="https://curio.lib.utexas.edu/assets/utblac_e38127378-7a33-412f-a425-cc511ff1d351.json">https://curio.lib.utexas.edu/assets/utblac_e38127378-7a33-412f-a425-cc511ff1d351.json</a>
2.		Pintura de Cuzcatlan	unknown (artist) Castañeda León, Juan de (signer) García, Perú (interpreter) Cortes,	1580-10-26	MX  Diocese of Tlaxcala  Coxcatán, Puebla, Mexico	18.266262, -97.147278	<a href="https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3/THN/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3.jpg">https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3/THN/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3.jpg</a>	<a href="https://creativecommons.org/publicdomain/mark/1.0/">https://creativecommons.org/publicdomain/mark/1.0/</a>	<a href="https://curio.lib.utexas.edu/assets/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3.json">https://curio.lib.utexas.edu/assets/utblac_e3818f71-f967-46fb-9493-1edec8cbc1b3.json</a>

**Extensions** [RDF Transform](#) [Wikidata](#)

Vista de creación del proyecto, desde el menú superior podemos configurar la extensión para generar el formato RDF

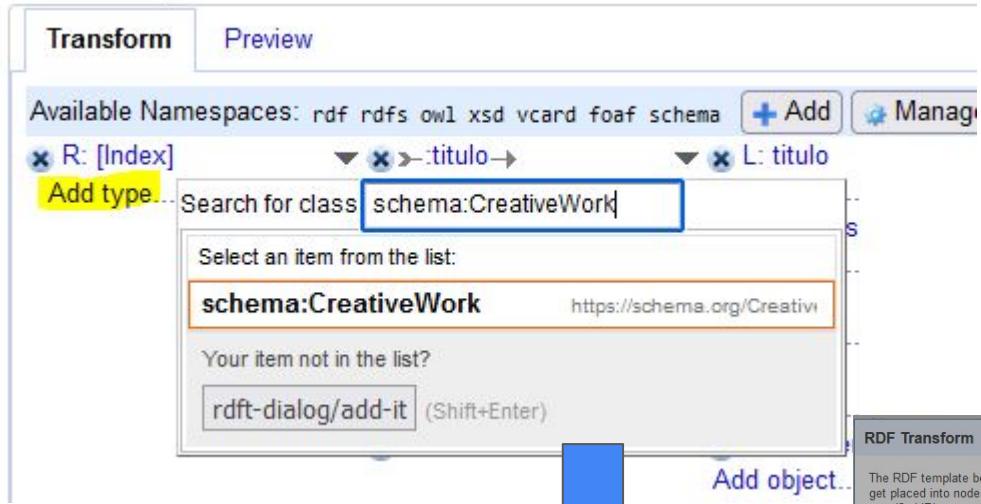
## RDF Transform

The RDF template below specifies how the RDF data is generated from your tabular data. The cells in each record of your data will get placed into nodes within the transform. Configure the transform by using column names and values, computed strings, or specified IRI as a subject, property, and object resources or literals. Compute strings using GREL. See the docs.

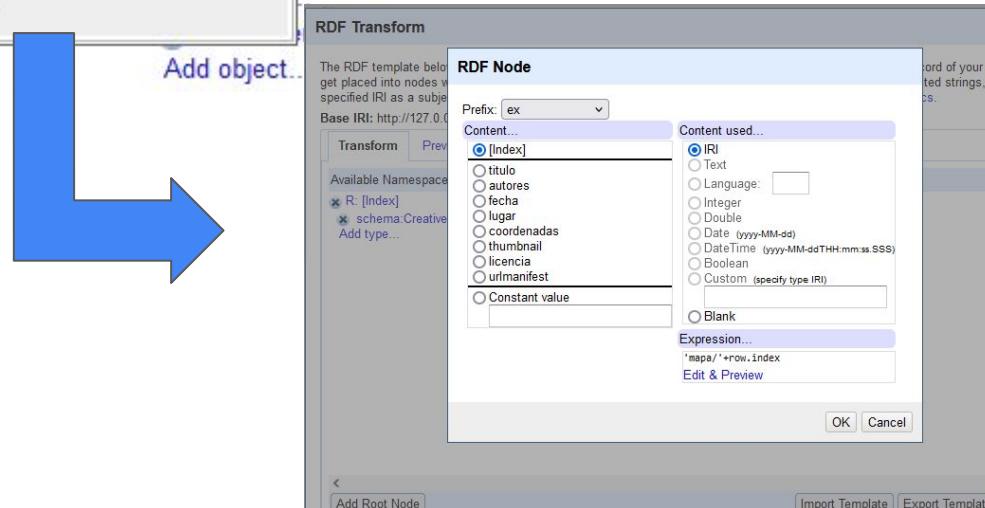
Base IRI: <http://127.0.0.1:3333/> [Edit](#)

The screenshot shows the 'RDF Transform' interface. On the left, there's a tree view of available namespaces under 'Available Namespaces'. A red box highlights the 'Add' and 'Manage' buttons at the top of this section. A large blue arrow points from the 'Manage' button area to a modal dialog box titled 'Add new prefix'. This dialog has 'Prefix:' set to 'schem' and 'IRI:' set to '<http://schema.org/>'. The main interface also displays the base IRI as 'http://127.0.0.1:3333/' and includes tabs for 'Transform' and 'Preview'.

Añadimos el espacio de nombres Schema.org y <https://example/> para crear las URIs (<http://example/mapa/{xyz}>)



Añadimos la clase y definimos la URL



## RDF Transform

The RDF template below specifies how the RDF data is generated from your tabular data. The cells in each row get placed into nodes within the transform. Configure the transform by using column names and values, or specify a specified IRI as a subject, property, and object resources or literals. Compute strings using GREL. See the [Base IRI](#): <http://127.0.0.1:3333/> [Edit](#)

Transform Preview

Available Namespaces: rdf rdfs owl xsd vcard foaf schema ex [+ Add](#) [Manage](#)

[R: \[Index\]](#) [schema:CreativeWork](#) [Add type...](#)

[Add property...](#)

- ✖ R: [Index]
- ✖ schema:CreativeWork
- Add type...
- Add property...

- ✖ ➔ schema:name ➔ L: titulo  
Add object...
- ✖ ➔ schema:dateCreated ➔ L: fecha  
Add object...
- ✖ ➔ schema:thumbnailURL ➔ L: thumbnail  
Add object...
- ✖ ➔ schema:license ➔ L: licencia  
Add object...

Añadimos las propiedades

propiedad de Schema.org

Columna CSV

## RDF Transform

The RDF template below specifies how the RDF data is generated from your tabular data. The cells in each record of your data will get placed into nodes within the transform. Configure the transform by using column names and values, computed strings, or specified IRI as a subject, property, and object resources or literals. Compute strings using GREL. [See the docs.](#)

Base IRI: <http://127.0.0.1:3030/> [Edit](#)

Transform Preview

Sample Data Preview: 20 rows. Shown below:  Turtle (Stream)  Turtle (Pretty)

```
@prefix schema: <https://schema.org/> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example/mapa/3>
    rdf:type schema:CreativeWork ;
    schema:dateCreated "1579-11-22" ;
    schema:license "https://creativecommons.org/publicdomain/mark/1.0/" ;
    schema:name "Pintura de Zapotitlan" ;
    schema:thumbnailURL "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_32713c87-b23c-474c-acb9-0757bc1d2dd3/TN/utblac_32713c87-b23c-474c-acb9-0757bc1d2dd3.jpg" .

<http://example/mapa/8>
    rdf:type schema:CreativeWork ;
    schema:dateCreated "1580-04-29" ;
    schema:license "https://creativecommons.org/publicdomain/mark/1.0/" ;
    schema:name "Pintura de Coatzacualco" ;
    schema:thumbnailURL "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_9a78720b-d927-4635-bdc0-9993a3024f9a/TN/utblac_9a78720b-d927-4635-bdc0-9993a3024f9a.jpg" .

<http://example/mapa/1>
    rdf:type schema:CreativeWork .
```

Podemos ver el resultado desde la opción Preview

Podemos exportar los datos una vez finalizado el proceso de transformación



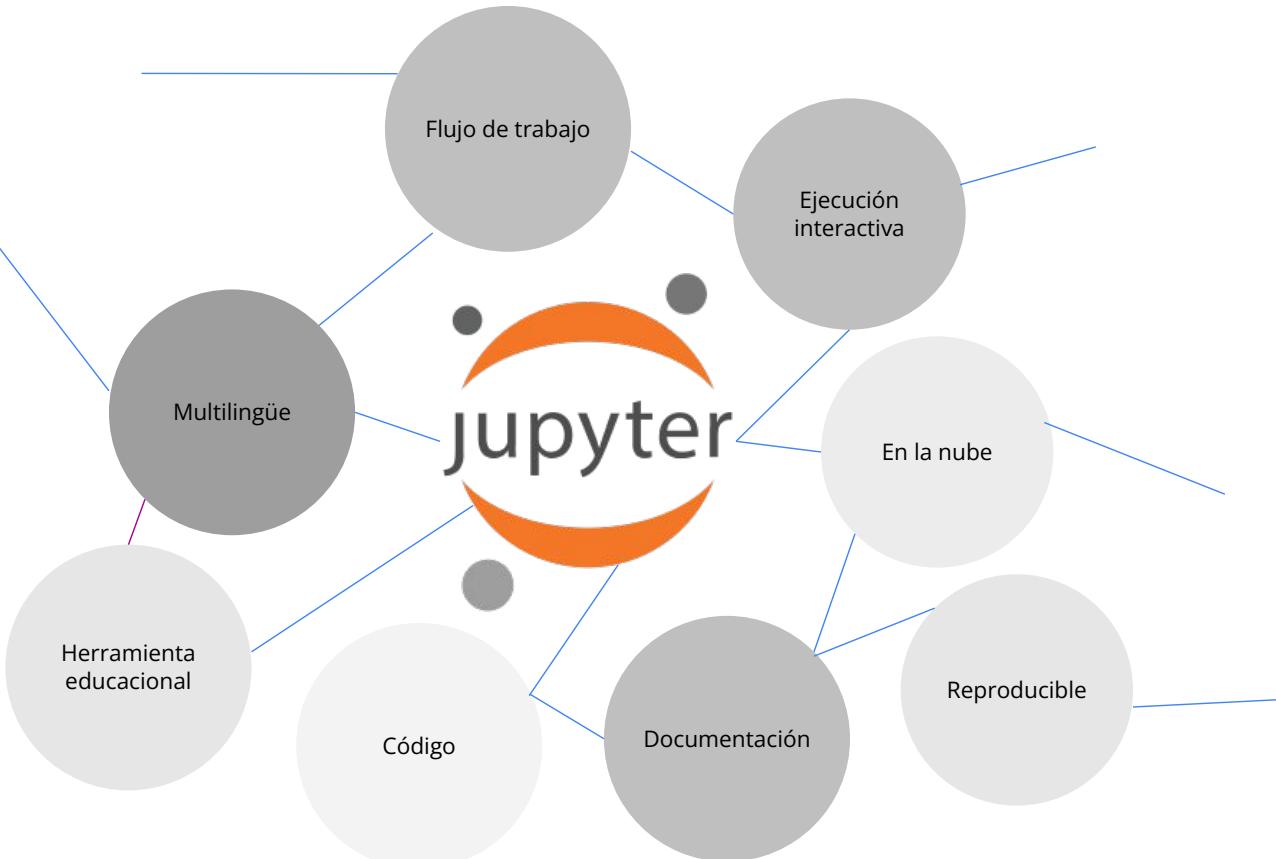
```
7 @prefix schema: <https://schema.org/> .
8 @prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
9 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
10
11 <http://example/mapa/3>
12     rdf:type          schema:CreativeWork ;
13     schema:dateCreated "1579-11-22" ;
14     schema:license    "https://creativecommons.org/publicdomain/mark/1.0/" ;
15     schema:name       "Pintura de Zapotitlan" ;
16     schema:thumbnailURL "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_32713c87-b23c-474c-ac
17
18 <http://example/mapa/8>
19     rdf:type          schema:CreativeWork ;
20     schema:dateCreated "1580-04-29" ;
21     schema:license    "https://creativecommons.org/publicdomain/mark/1.0/" ;
22     schema:name       "Pintura de Coatzacualco" ;
23     schema:thumbnailURL "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_9a78720b-d927-4635-bd
24
25 <http://example/mapa/1>
26     rdf:type          schema:CreativeWork ;
27     schema:dateCreated "1580-10-26" ;
28     schema:license    "https://creativecommons.org/publicdomain/mark/1.0/" ;
29     schema:name       "Pintura de Cuzcatlan" ;
30     schema:thumbnailURL "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_e38f8f71-f967-46fb-94
~
```

El fichero RDF en formato TTL se encuentra en la carpeta output del proyecto GitHub

[https://github.com/hibernator11/uned-unlocking-workshop/blob/main/output/UNED\\_metadata\\_csv.ttl](https://github.com/hibernator11/uned-unlocking-workshop/blob/main/output/UNED_metadata_csv.ttl)

# Transformación - Código reproducible

[GLAM WorkBench](#)  
[LC Maps for Robots](#)  
[BL Labs](#)  
[National Library of Scotland](#)  
[Archives Unleashed Notebooks](#)  
[LC - Newspapers navigator](#)  
[Biblioteca Virtual Miguel de Cervantes](#)



**dhlab-nova-lisboa**

**1**

[launch binder](#)

This project was originally created for a workshop [de Lisboa](#).



This project provide several examples of reuse based on different techniques: Language Processing.

## Map representing the nationalities of the authors of *El Quijote* by Cervantes

This example is based on the [Linked Open Data version of the Biblioteca Virtual Miguel de Cervantes](#) from Wikidata, enabling the creation of visualisations using the Wikidata SPARQL API.

<https://github.com/hibernator11/dhlab-nova-lisboa>

Thanks to OVH, GESIS Notebooks and Curvendo for supporting us! 🎉 mybinder.org has updated the base image to Ubuntu 22.04! See the [upgrade guide](#) for details.



**2**

Getting repository: hibernator11/uned-unlocking-workshop

**3**

The screenshot shows a Jupyter Notebook interface with the following content:

```
File Edit View Run Kernel Tabs Settings Help  
Launcher x https://hub.binder.eu.org/user/hibernator11/unlocking-workshop-04xneddb/lab/tree/notebooks/  
Filter files by name  
/ notebooks /  
Name Last Modified  
Extraccion... 2 minutes ago  
SPARQLana... 2 minutes ago
```

**Introducción**

Este ejemplo se encarga de extraer los metadatos necesarios de la colección. En concreto, vamos a extraer los datos de los libros de la colección.

En primer lugar cargamos las librerías necesarias

```
[1]: import urllib.request, json  
import csv
```

Función que extrae los campos requeridos

```
[47]: def getCampos(urlmanifest):  
    manifest = json.load(urllib.request.urlopen(urlmanifest))  
  
    titulo = ''  
    autores = ''  
    fecha = ''  
    lugar = ''  
    coordenadas = ''  
    thumbnail = ''  
    licencia = ''
```

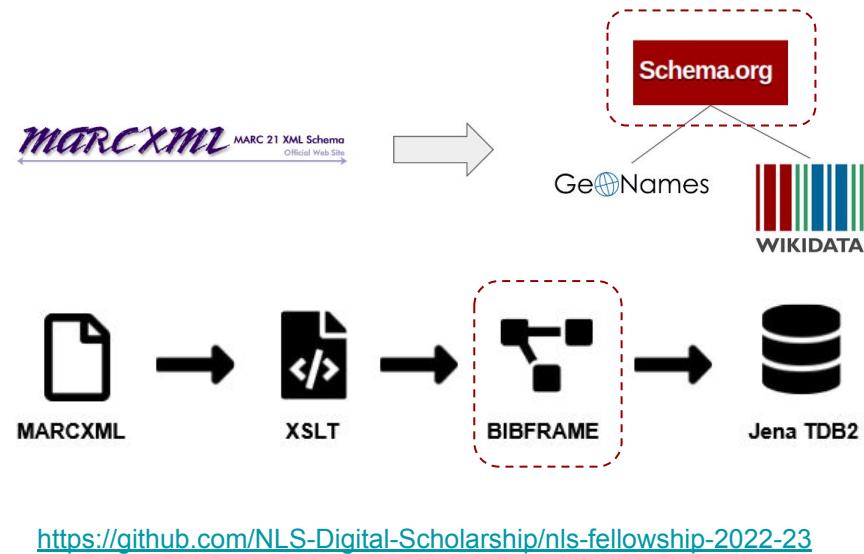


# Transformación

## National Librarian's Research Fellowship in Digital Scholarship 2022-23



Candela, G. (2023). Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland. *Journal of Information Science*.  
<https://doi.org/10.1177/01655515231174386>



<https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23>

# Transformación

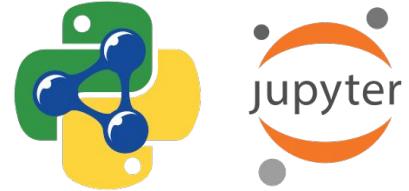
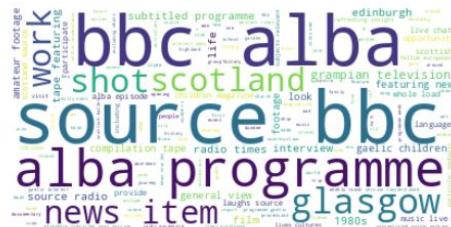
## Moving Image Archive

This dataset represents the descriptive metadata from the Moving Image Archive catalogue, which is Scotland's national collection of moving images.

- Data format: metadata available as MARCXML and Dublin Core
- Data source: <https://data.nls.uk/data/metadata-collections/moving-image-archive/>

The Jupyter Notebooks include a set of examples to reproduce the transformation to RDF and enrichment with external repositories:

- [Data extraction](#)
- [Exploring the CSV text file](#)
- [Transformation to LOD](#)
- [Enrichment](#)
- [Exploring with SPARQL](#)
- [Exploring geographic locations](#)
- [Data Quality assessment](#)



## Preparation

Import the libraries required to explore the summary of each record included in the dataset to present a word cloud

```
In [3...]:  
from rdflib import Graph, URIRef, Literal, Namespace  
from rdflib.namespace import FOAF, RDF, DCTERMS, VOID, DC, SKOS  
import pandas as pd
```

## Transformation to RDF

\*Note: The variable domain could be updated to the domain of the organisation (e.g., <https://data.nls.uk>).

```
In [3...]:  
domain = 'https://example.org/'
```

First, we instantiate all the namespaces that we will use when defining the RDF data

```
In [3...]:  
g = Graph()  
g.bind("foaf", FOAF)  
g.bind("rdf", RDF)  
g.bind("dcterms", DCTERMS)  
g.bind("dc", DC)  
g.bind("void", VOID)  
g.bind("skos", SKOS)  
  
schema = Namespace("https://schema.org/")  
g.bind("schema", schema)  
  
edm = Namespace("http://www.europeana.eu/schemas edm/")  
g.bind("edm", edm)
```

We define the resource National Library of Scotland

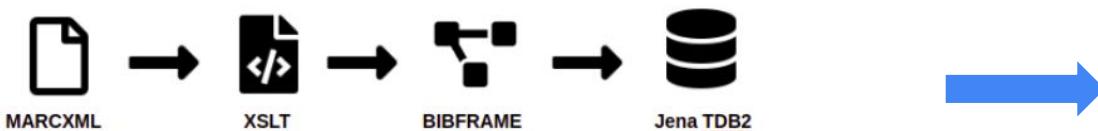
```
In [3...]:  
nls = URIRef(domain + "organisation/nls")  
g.add((nls, RDF.type, schema.Organization))  
g.add((nls, schema.url, URIRef("https://www.nls.uk/")))  
g.add((nls, schema.logo, URIRef("https://www.nls.uk/images/nls-logo.png")))  
g.add((nls, schema.name, Literal("National Library of Scotland")))  
g.add((nls, DC.title, Literal("National Library of Scotland")))
```

<https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23>

# Transformación

## National Bibliography of Scotland and BOSLIT

The transformation process is based on the tool [marc2bibframe](#) that uses BIBFRAME as main vocabulary to describe the resources.



<https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23>

### Loading RDF

The class `JenaTDBLoad` is in charge of loading the RDF into the RDF Jena TDB storage system. The RDF files must be placed in the folder `rdf`. The following code shows the process:

```
// Create dataset
Path path = Paths.get(".").toAbsolutePath().normalize();
String dbDir = path.toFile().getAbsolutePath() + "/db/";
Location location = Location.create(dbDir);
Dataset dataset = TDB2Factory.connectDataset(location);

dataset.begin(ReadWrite.WRITE);
Model model = dataset.getDefaultModel();

// update path
Files.walk(Paths.get(path.toFile().getAbsolutePath() + "/rdf"))
.filter(p -> p.toString().endsWith(".gz"))
.forEach(p -> {
    logger.info(p.toFile().getAbsolutePath());
    try {
        RDFDataMgr.read(model, p.toFile().getAbsolutePath(), Lang.RDFXML);
    }catch (Exception e){
        logger.error(p.toFile().getAbsolutePath() + e.getMessage());
    }
});
dataset.commit();

// Releasing dataset resources
dataset.close();
```

<https://github.com/hibernator11/nls-jena-tdb>



**Maven**<sup>TM</sup>

 Apache Jena



# Transformación



<https://atrium-research.eu/>

Facilitating access to digital research infrastructures and advancing frontier knowledge in the arts and humanities — across disciplines, languages and media.



Bibliographic records  
4 777 867



Persons  
202 040



Places  
9514



Corporates  
7454



Magazines  
843



Events  
319

[literarybibliography.eu](http://literarybibliography.eu)



# Transformación

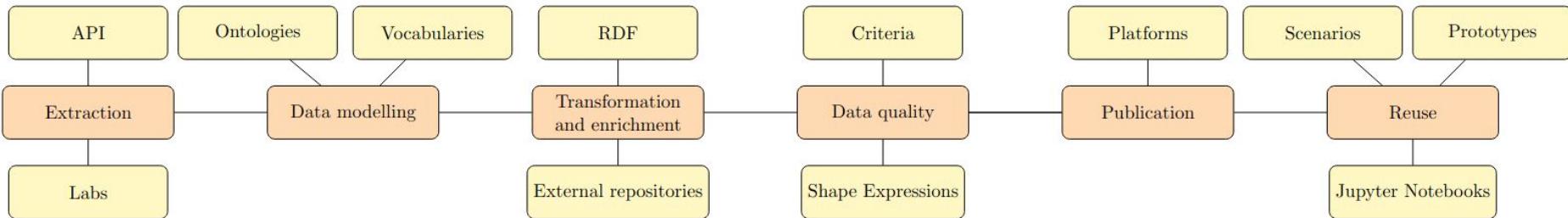


<https://atrium-research.eu/>

Facilitating access to digital research infrastructures and advancing frontier knowledge in the arts and humanities — across disciplines, languages and media.



[literarybibliography.eu](http://literarybibliography.eu)



# ATRIUM project

## A reproducible approach

The **Jupyter Notebooks** include a set of examples to reproduce the transformation to RDF.

- [Data extraction](#): extraction of the data from the biblio.json file to create a CSV file.
- [Data exploration](#): exploration of the original biblio.json file.
- [Transformation to LOD](#): transformation of the original data using the CSV file to RDF using as main vocabulary Schema.org.
- [Data exploration using SPARQL](#): exploration using SPARQL of the RDF data generated.
- [Data quality ShEx](#): validation of the RDF data generated using Shape Expressions. In order to check how to test the ShEx against the RDF generated you can see the [Data quality assess notebook](#).



Candela, G., Rosiński, C., & Margraf, A. (2024). A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography.  
<https://doi.org/10.5281/zenodo.14106707>



# Transformación



## Data

European  
Literary  
Bibliography



API, ZIP files



## Infrastructure

Data curation and  
management

Code  
repository

Cloud  
computing

Jupyter  
Notebooks

Sustainability

Schema.org



## Publication & reuse

Vocabularies

Best practices

Dissemination

Research  
scenarios

Data platforms



Architecture defined to apply the framework proposed in this work. It focuses on three main elements: data, infrastructure, and publication



# Transformación

Dataset	Classes	Properties	No. triples	No. external links
Adam Mickiewicz	7	19	47,631	863
Miguel de Cervantes	5	19	44,389	749
National Library of Spain	7	19	10,515,016	67,563

**Scenario 1:** Comparative analysis of provincial Spanish novels featuring vampires

**Scenario 2:** Republican writers who migrated during the Spanish Civil War

**Scenario 3:** Geographical distribution of publications dedicated to particular Spanish writers

**Article**

## A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography

Gustavo Candela (1), Cezary Rosiński (2), Arkadiusz Margraf (3)

(1) University of Alicante

(2) Institute of Literary Research of the Polish Academy of Sciences

(3) Institute of Bioorganic Chemistry of the Polish Academy of Sciences

### Abstract

EN GLAM (Galleries, Libraries, Archives and Museums) institutions host rich content that is provided in the form of digital collections. Bibliographic databases are collections of references focused on a particular topic that can be used to apply Digital Humanities (DH) methods. Recent approaches such as Collections as Data and Labs promote the publication of digital collections supporting computational use. This work aims to provide a framework for publishing and reusing digital collections based on literary bibliographies published by GLAM institutions in order to make them suitable for computational use in the form of Collections as Data, in particular, in the context of the European Literary Bibliography. It also describes the infrastructure used and DH research scenarios to illustrate how the results can be reused for different goals. Digital curators and DH researchers interested in making their datasets available in the form of Collections as Data are the intended audience of this work.

<https://doi.org/10.46298/transformations.14729>

[Download article](#)[Open on Zenodo](#)[Cite](#)[Metadata](#)[Share](#)

### Publication details

Published on

June 16, 2025

Accepted on

April 3, 2025

Submitted on

November 12, 2024

### Volume 1

Workflows

DOI

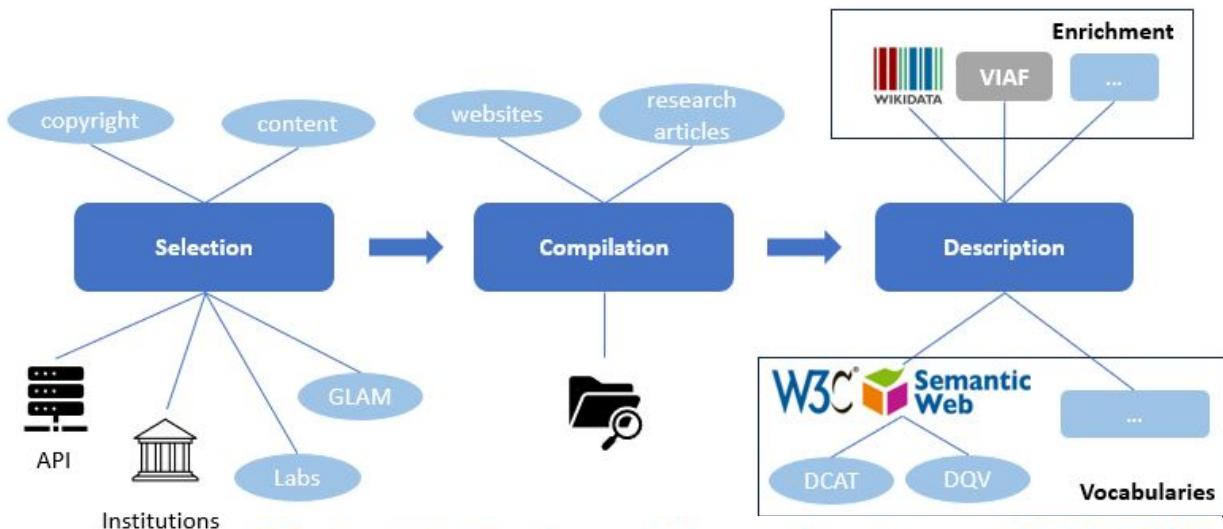
[10.46298/transformations.14729](https://doi.org/10.46298/transformations.14729)



<https://marketplace.sshopencloud.eu/workflow/I3JvP6>

# Transformación

## Espacios de datos de patrimonio cultural



## Data Catalog Vocabulary (DCAT)

<https://github.com/hibernator11/dcat-glam-catalog>

### Workflow steps (10)

- Provide a clear license and terms of use allowing reuse of the dataset without restrictions
  - Provide a suggested citation for the dataset so reusers are aware of how to cite it
- about the dataset
- make available the dataset for the public
  - demonstrate how the dataset can be reused
  - for the dataset for a better understanding of how to reuse
- role metadata about the content provided in the dataset
- ative-edition platform to include the information about the
- means of an existing API
- sent and describe the dataset to encourage its reuse

# Transformación

## Spanish-Civil-War-KGs

Extracting Knowledge Graphs as Collections as Data using the Spanish Civil War as case study

DOI [10.5281/zenodo.1674752](https://doi.org/10.5281/zenodo.1674752)

### Introduction

This project intends to analyse the options and potential of Wikidata to extract Knowledge Graphs as Collections as Data using the Spanish Civil War as case study.

The data was retrieved from the [National Archives in Spain](#), and enriched in Wikidata.

The authors of this work are:

- Gustavo Candela, University of Alicante, Spain
- Meltem Dişli, Hacettepe University, Türkiye
- Paul Spence, King's College London, United Kingdom
- Cezary Rosiński, Institute of Literary Research of Polish Academy of Sciences, Poland
- Arkadiusz Margraf, Poznan Supercomputing and Networking Center affiliated to Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Poland

We provide 4 examples of use covering different topics and content:

- Artists from the period of the Spanish Republic and Spanish Civil War
- Refugee ships on which Spanish exiles embarked during the Spanish Exile and the Spanish Civil War
- French refugee camps, which concentrated the Spanish Republican exiles of 1939
- Retrieving bibliographic records related to the Spanish Civil War from the Biblioteca Virtual Miguel de Cervantes LOD repository

## Casos de uso basados en la extracción de datos y la Guerra Civil Española



# Transformación

## ¿Cómo buscar en varios repositorios y evitar silos de datos?

### Creating Collections as Data Using Federated Queries



This repository provides a series of **Jupyter notebooks** designed for performing **federated queries** across Wikidata and other **linked open data (LOD)** endpoints, with a focus on querying bibliographic and author-related data.

These notebooks support our paper *From Linked Open Data to Collections as Data: A Reproducible Framework Using Federated Queries* and aim to present reproducible examples of query design strategies other than structured real-world use in library and information studies, digital humanities and other domains.

[launch binder](#)

#### Materials in this Repository

The repository is organized into the following sections:

- **Query Building:**
  - 4 examples showcasing a query-building workflow for three different endpoints:
    - [Biblioteca Nacional de España \(BNE\) Query Building](#)
    - [Bibliothèque nationale de France \(BNF\) Query Building](#)
    - [Biblioteca Virtual Miguel de Cervantes \(BVMC\) Query Building](#)
    - [Wikidata\\_map\\_viz](#), an exploratory query visualizing authors' places of birth from the Spanish Golden Age (wd:Q530936).
- **Single Author, Multiple Works:**
  - 2 examples focusing on retrieving multiple works by a single author:
    - [Jorge Juan y Santacilia](#)
    - [Maria Zayas y Sotomayor](#)
- **Multiple Authors, Multiple Works:**
  - 1 example demonstrating queries involving multiple authors and their works,
    - [Spanish Golden Age](#)

### Ejemplos de sentencias SPARQL federadas

# Transformación

¿Cómo buscar en varios repositorios y evitar silos de datos?

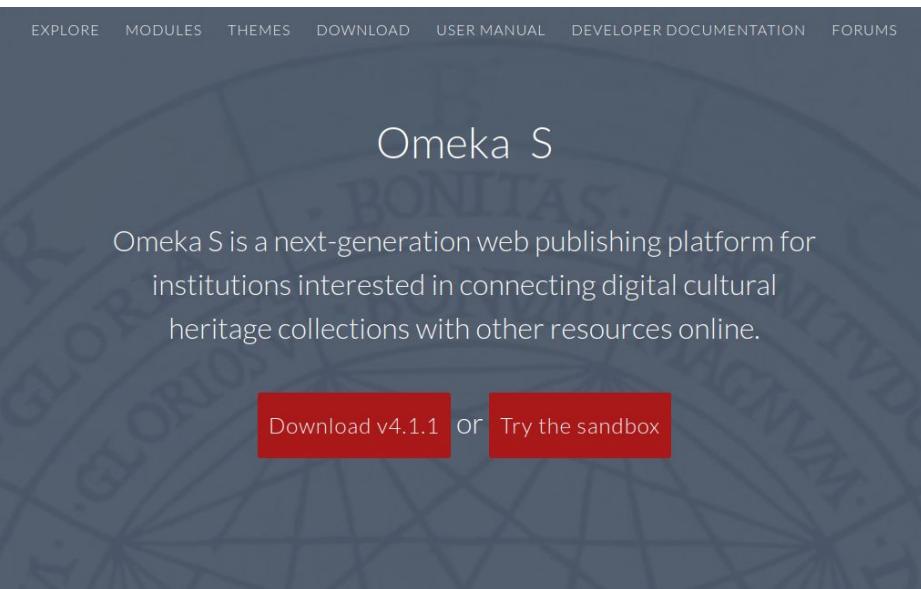
PREFIX bne-def: <https://datos.bne.es/def/>

PREFIX dcterms: <http://purl.org/dc/terms/>

```
SELECT ?author ?authorLabel ?work ?workLabel ?edition ?placeOfProduction ?yearOfPublication ?langCode
WHERE {
    ?author wdt:P950 ?id .
    ?author wdt:P135 wd:Q530936 .
    ?author rdfs:label ?authorLabel. FILTER(LANG(?authorLabel) = "en").
    BIND(uri(concat("https://datos.bne.es/resource/", ?id)) as ?bneID)
    SERVICE <http://datos.bne.es/sparql> {
        ?bneID bne-def:OP5001 ?work .
        ?work rdfs:label ?workLabel .
        OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition bne-def:P3003 ?placeOfProduction}
        OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition bne-def:P3006 ?yearOfPublication}
        OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition dcterms:language ?langCode}
    }
}
LIMIT 1000
```

Obras de la BNE de autores  
pertenecientes al Siglo de Oro en Wikidata

# Transformación



The screenshot shows the Omeka S homepage. At the top, there is a navigation bar with links: EXPLORE, MODULES, THEMES, DOWNLOAD, USER MANUAL, DEVELOPER DOCUMENTATION, and FORUMS. Below the navigation bar, the title "Omeka S" is displayed in a large, serif font. A descriptive paragraph follows: "Omeka S is a next-generation web publishing platform for institutions interested in connecting digital cultural heritage collections with other resources online." At the bottom of the page are two red buttons: "Download v4.1.1" and "Try the sandbox". The background of the page features a faint watermark of a globe.

Omeka S

Omeka S is a next-generation web publishing platform for institutions interested in connecting digital cultural heritage collections with other resources online.

Download v4.1.1 or Try the sandbox

<https://omeka.org/s/>



The screenshot shows the Wikibase Cloud landing page. At the top, the Wikibase Cloud logo is visible. The main headline reads "Your own Wikibase in five minutes" in large, white, sans-serif font. Below the headline, the text "Simple & Secure" is displayed. At the bottom, there is a blue button labeled "TRY IT OUT".

Your own Wikibase  
in five minutes

Simple & Secure

TRY IT OUT

<https://www.wikibase.cloud/>

# Transformación

DHWIKI

Buscar en DHWiki

## Main Page

Portada Discusión Ver código fuente Historial

Welcome do DHwiki Wikibase. This mediaWiki installation, hosted on Wikibase Cloud, includes a Wikibase instance, and is maintained since September 2024 by the DARIAH-EU WG DHwiki, in the framework of the DARIAH ERIC.

Please read the public announce of the WG creation. You find a short set of slides about our group here: <https://doi.org/10.5281/zenodo.16041468>. You find dates and agendas of our group meetings here: [Group meetings](#).

This Wikibase contains:

- Documentation pages about Wikibase in a Digital Humanities context (in the making.)
- a LOD version of our group bibliography (a Zotero collection available [here](#) (see list of contributors));
- information about group members;
- textual descriptions of Wikibase instances related to the digital humanities (cf. [Category:Wikibase instances](#))
- resources related to Wikibase and its uses in the digital humanities (cf. [Resource directory](#)).

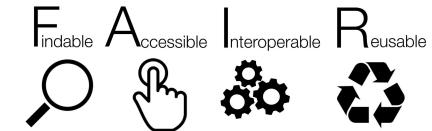
The group has set up a Zenodo community, where you find our papers and presentation slides: [https://zenodo.org/communities/dariah\\_dhwiki/](https://zenodo.org/communities/dariah_dhwiki/)

We also are having experiments with a set of Jupyter notebooks.

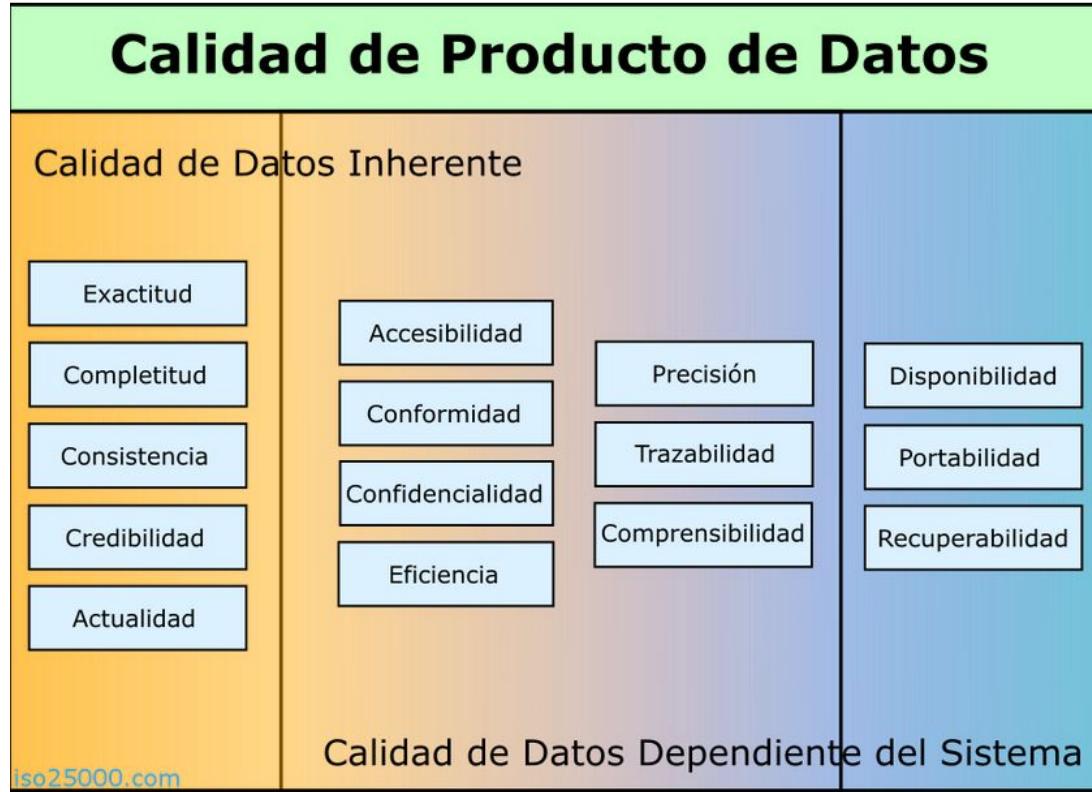
Sumario ocular

1 What's in DHwiki Wikibase?  
1.1 Bibliographical items  
1.1.1 Bibliographical items domains  
1.2 Persons  
1.2.1 Authors  
1.2.2 Group members  
1.2.2.1 Group members image gallery  
1.3 Wikibase instances described on Wikibase World and on DHwiki Wikibase  
1.4 Classes and instances on this Wikibase

[https://dhwiki.wikibase.cloud/wiki/Main\\_Page](https://dhwiki.wikibase.cloud/wiki/Main_Page)



# Calidad de datos



¿Qué medimos?

# Calidad de datos

## Ejemplos



<https://shex.io/>

A Shape Expression approach for assessing the quality of Linked Open Data in libraries

<https://doi.org/10.3233/SW-210441>



A reproducible framework to assess LOD

<https://github.com/hibernator11/lod-quality-reproducible>



Evaluating the quality of linked open data in digital libraries

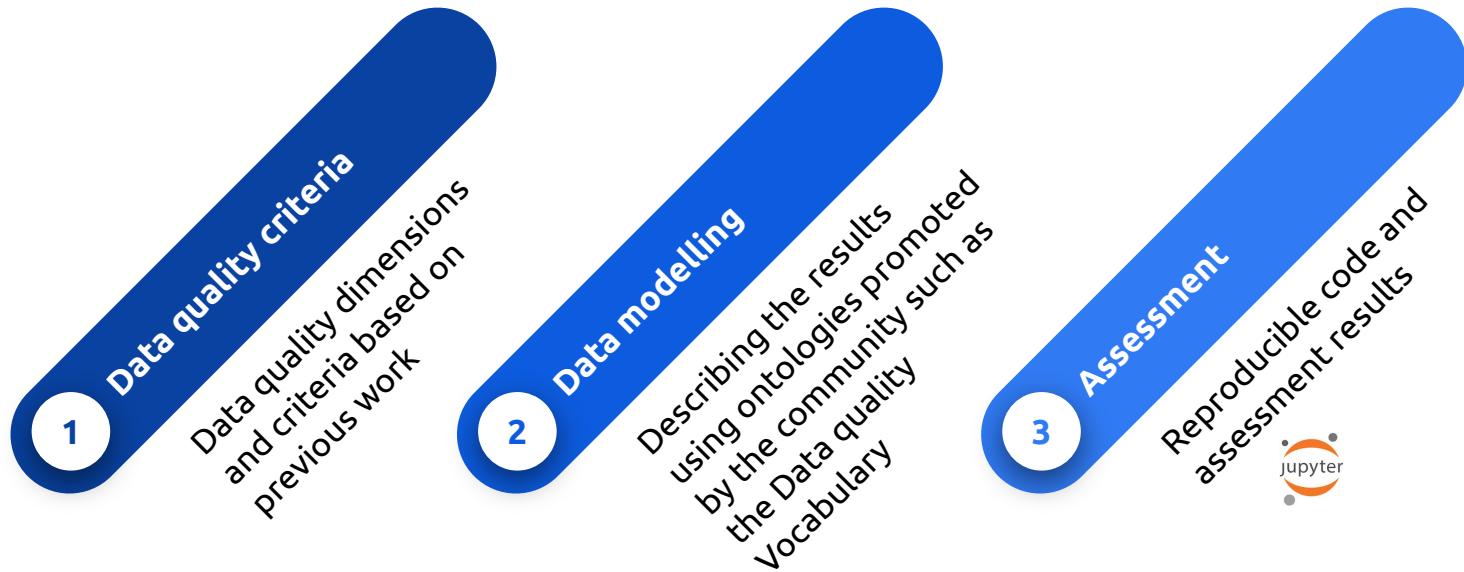
<https://doi.org/10.1177/0165551520930951>



An automatic data quality approach to assess semantic data from cultural heritage institutions

<https://doi.org/10.1002/asi.24761>

# A Reproducible Framework to Assess LOD Quality



# Assessment results

Running in  
the cloud

## Create visual interface

```
children = []
titles = []
tab = widgets.Tab()
children.insert(0,introductionLayout)
children.insert(1,criteriaLayout)
children.insert(2,assessmentResultsLayout)
tab.children = children

titles.insert(0, 'Repository')
titles.insert(1, 'Data quality criteria')
titles.insert(2, 'Assessment results')
tab.titles = titles

tab
```

4

Repository	Data quality criteria	Assessment results
SPARQL: <input type="text" value="https://data.bnf.fr/sparql"/>		
Title: <input type="text" value="National Library of France"/>		
Description: <input type="text" value="Reference information about authors, work"/>		

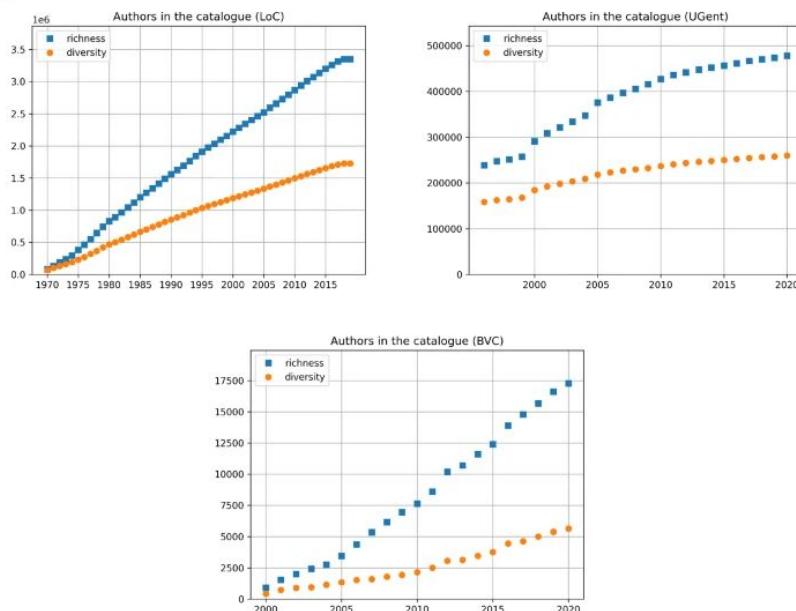
5

Repository	Data quality criteria	Assessment results															
This section shows the data quality assessment results obtained for the LOD repositories selected. The results are retrieved from the ttl files in the data folder using the properties dqv:value.																	
<table border="1"><thead><tr><th>Dimension</th><th>Criterion</th><th>Result</th></tr></thead><tbody><tr><td>Availability</td><td>Availability</td><td>correct</td></tr><tr><td>Licensing</td><td>Licensing</td><td>not satisfied</td></tr><tr><td>Interlinking</td><td>Interlinking</td><td>correct</td></tr><tr><td>Security</td><td>Security</td><td>correct</td></tr></tbody></table>			Dimension	Criterion	Result	Availability	Availability	correct	Licensing	Licensing	not satisfied	Interlinking	Interlinking	correct	Security	Security	correct
Dimension	Criterion	Result															
Availability	Availability	correct															
Licensing	Licensing	not satisfied															
Interlinking	Interlinking	correct															
Security	Security	correct															

6

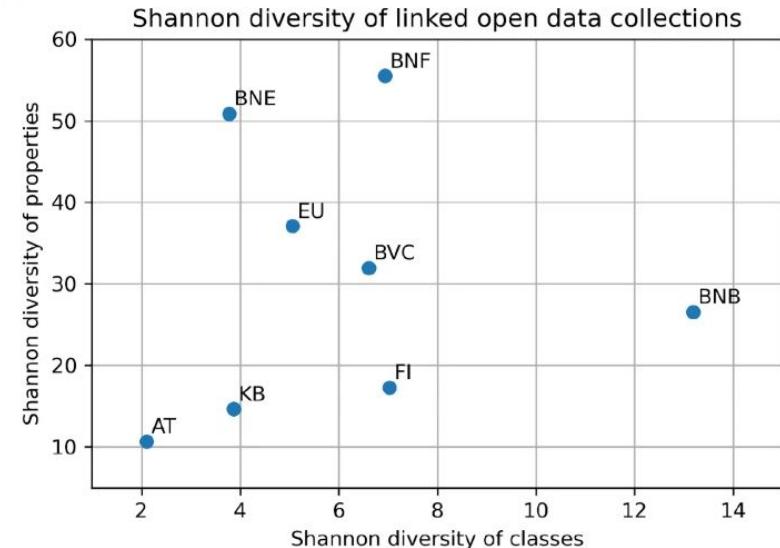
# Calidad

**Fig. 8**



Cumulative number of authors and Shannon diversity of the authors in the catalogue as a function of the year the MARC record was included in the catalogue extracted from the MARC field 008

**Fig. 10**



Shannon diversity of classes and properties in *Linked Open Data* published by libraries

# Visualización

## Visualizaciones

- #defaultView:ImageGrid

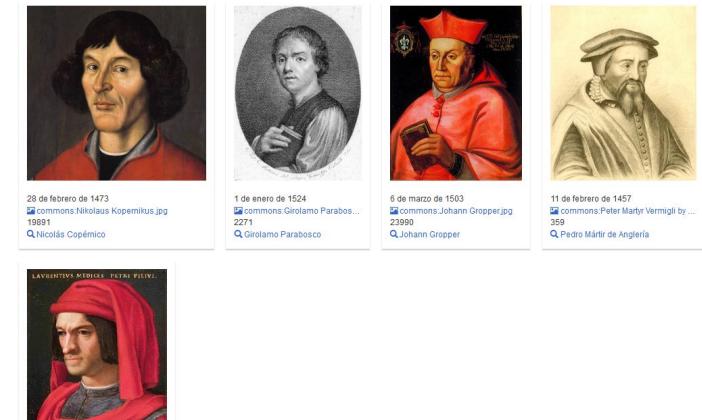
```
#defaultView:ImageGrid
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
    ?autor wdt:P2799 ?idbvmc .
    ?autor wdt:P569 ?fechaNacimiento .
    ?autor wdt:P18 ?imagen

    FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
        && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .

    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es".
    }
}
LIMIT 10
```



<https://w.wiki/8922>



# Visualización

## Visualizaciones

- #defaultView:Timeline

```
#defaultView:Timeline
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
    ?autor wdt:P2799 ?idbvmc .
    ?autor wdt:P569 ?fechaNacimiento .
    ?autor wdt:P18 ?imagen}
```

```
FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
&& ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .
```

```
SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es". }
}
LIMIT 10
```



[https://w.wiki/88\\$y](https://w.wiki/88$y)

# Visualización

A reproducible approach to browse Linked Open Data in Cultural Heritage



Schema.org



<https://github.com/hibernator11/visual-ch-ld>

The screenshot shows a Jupyter notebook interface with several tabs and visualizations:

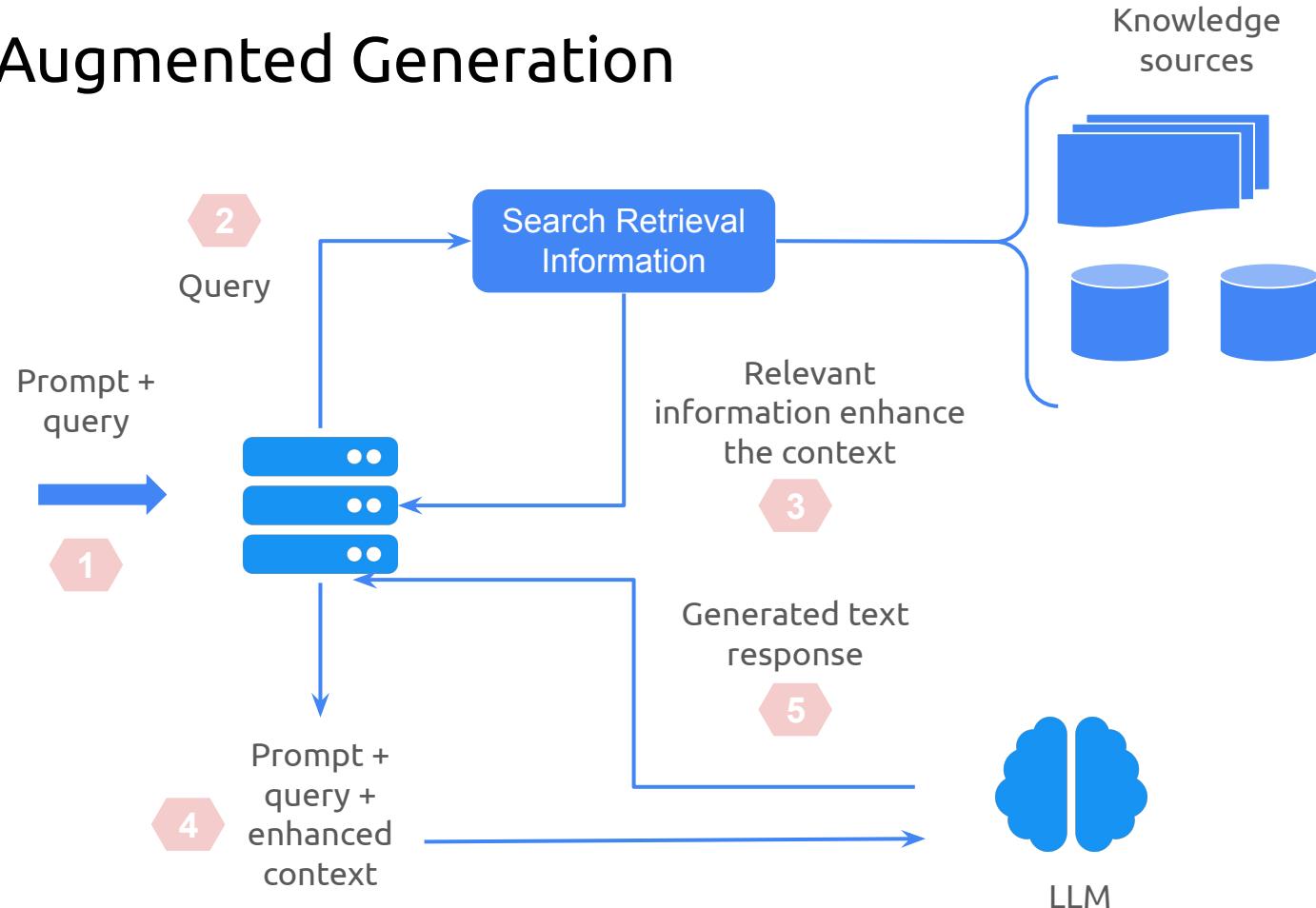
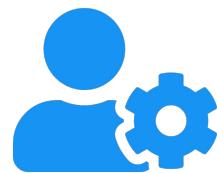
- Top Tab Bar:** SPARQL endpoint, Author, Classes, Properties, Wikidata.
- Author Tab:** Shows results for "author Frederick Sydney Fox - http://c".
- Metadata Panel:** Displays RDF triples:
  - `http://www.w3.org/1999/02/22-rdf-syntax-ns#type: http://xmlns.com/foaf/0.1/Person`
  - `http://www.w3.org/2002/07/owl#sameAs: http://data.bnfr/ark:/12148/cb177487799#foaf:Person`
  - `http://xmlns.com/foaf/0.1/depiction: https://gallica.bnfr/ark:/12148/btv1b53175239d.thumbnail`
  - `http://xmlns.com/foaf/0.1/depiction: https://gallica.bnfr/ark:/12148`
- Image:** A black and white portrait photograph of Frederick Sydney Fox.
- Wikidata Tab:** Shows a world map with red dots representing Wikidata entities and a bar chart of entity counts.
- Properties Tab:** Displays a list of properties:
  - `/data.bnfr/ontology/bnf-onto/ExpositionVirtuelle`
  - `/purl.org/dc/dcmitype/Event`
  - `/purl.org/dc/dcmitype/InteractiveResource`
  - `/purl.org/iso25964/skos-thes#ConceptGroup`
  - `/purl.org/ontology/bibo/Periodical`
  - `/rdaregistry.info/Elements/c/#C10001`
  - `/rdaregistry.info/Elements/c/#C10006`
  - `/rdaregistry.info/Elements/c/#C10007`
- Bottom Tab Bar:** SPARQL endpoint, Author, Classes, Properties, Wikidata.
- Bottom Panel:** Displays a bar chart of entity counts for various properties:

Property	Count
riontology:onto	~1.5
url.org/dc/dcmitype	~2.0
o25964:kos-the#	~1.0
bibo	~1.0
1:FRBRentitiesRDA	~3.0
y.info/Elements/c/#	~3.0
a:FRBRentitiesRDA	~3.0
mischemas:vinted#	~1.0
2/22-rdf-syntax-ns#	~1.0
0/0/1/rdf-schema#	~1.0
3.org/2002/07/owl#	~4.0
004/02/kos/core#	~1.0
1/geonows84_pos#	~1.0
004/02/kos/core#	~3.0
service-description#	~1.0
xmlins.com/foaf/0.1	~3.0

<https://doi.org/10.1145/3707647>

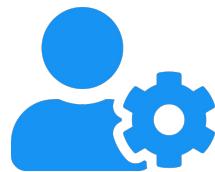
# Retrieval Augmented Generation

RAG method



# Retrieval Augmented Generation

RAG application



LangChain

4

Prompt + query

1

2  
Query



Word embeddings

Search Retrieval Information

Relevant information enhance the context

3

Generated text response

5

Prompt + query + enhanced context



LLM



Hugging Face



Knowledge sources

Local/Cloud environment

