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

W3C

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 corersponds 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 processes 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

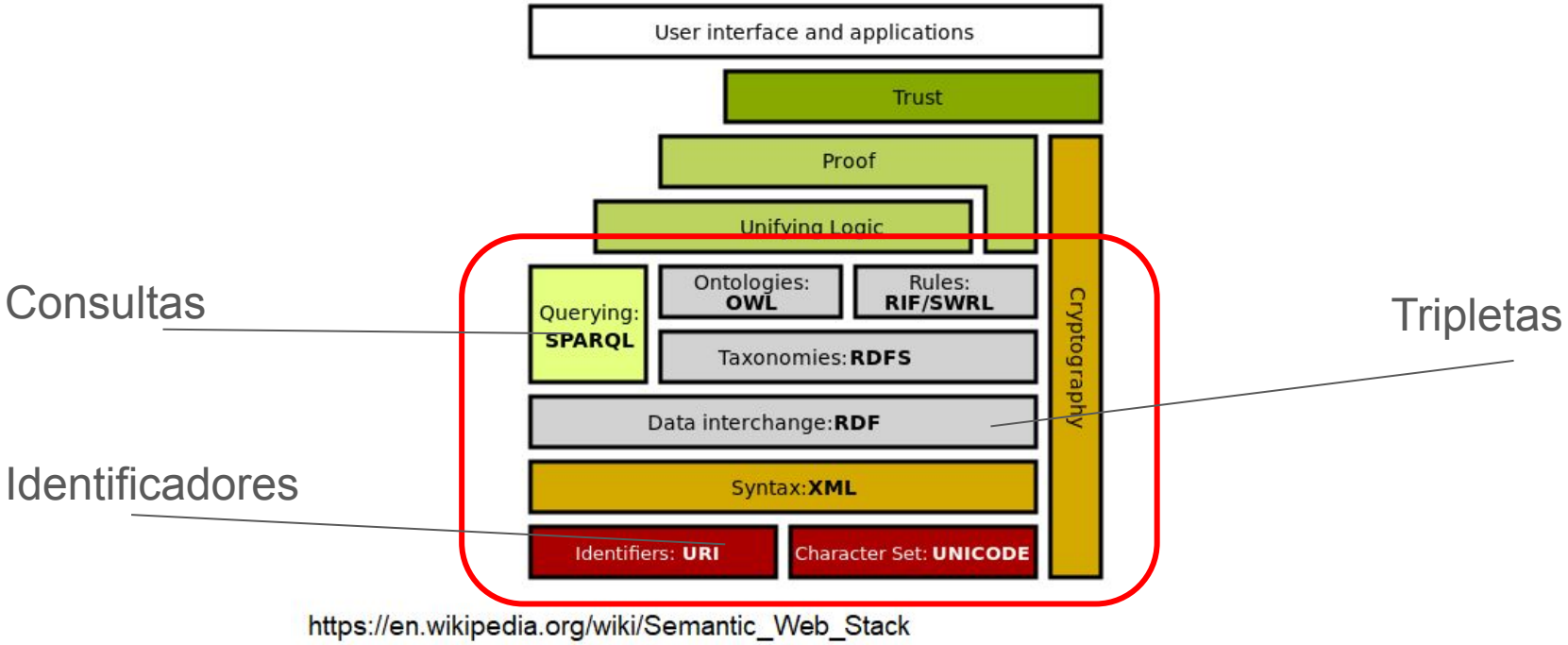
Two important technologies for the semantic web are already in place. The XML language for expressing information gives the community a common way of creating structured documents, thoug it doesn't have a way of saying in any sense what they mean. However, a starting point for talking about meaning comes 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



https://en.wikipedia.org/wiki/Semantic_Web_Stack

Web semántica y Linked Open Data

Enriquecimiento

3

owl:sameAs
skos:exactMatch
wdt:P2799

Ontologías y vocabularios

2

Vocabularios promovidos por instituciones como IFLA, W3C, DARIAH

Tecnologías y estándares

1

SPARQL, OWL, RDF, URIs

https://5stardata.info/es/

Web semántica y Linked Open Data

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

es_autor_de

Sujeto **Predicado** **Objeto**

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

Enriquecimiento

owl:sameAs
skos:exactMatch
wdt:P2799

Ontologías y vocabularios

Schema.org

RDA Registry

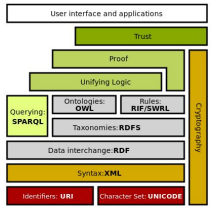
Modelo de Referencia Bibliotecaria de la IFLA

Modelo Conceptual para la Información Bibliográfica

TaDiRAH

Tecnologías y estándares

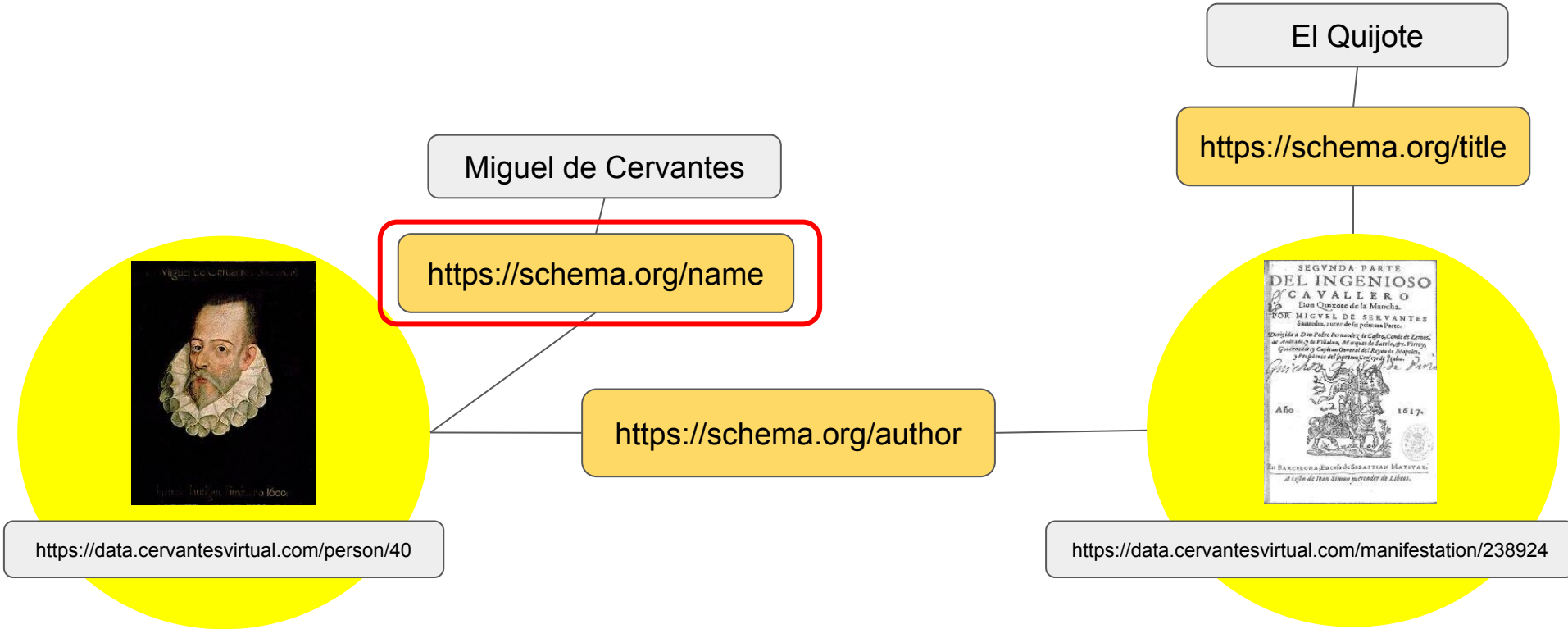
SPARQL, OWL, RDF, URIs



https://en.wikipedia.org/wiki/Semantic_Web_Stack

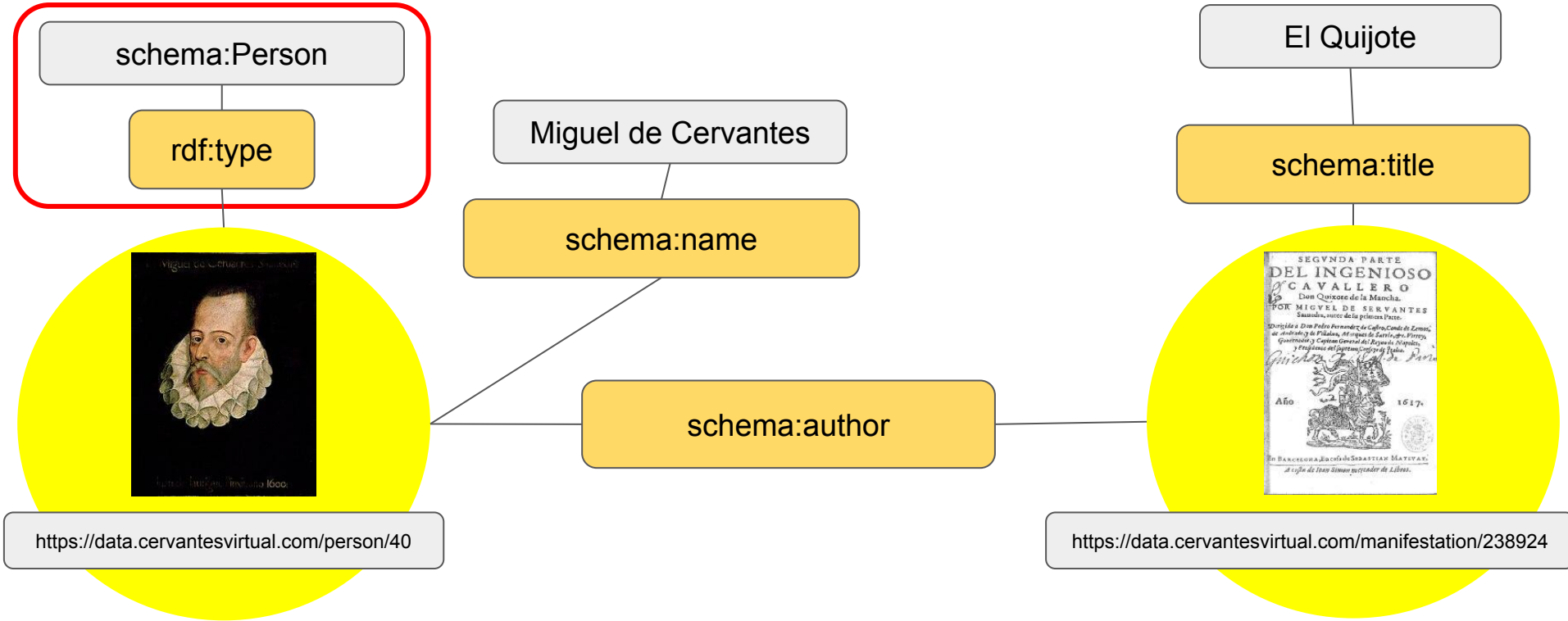
Web semántica y Linked Open Data

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

Enriquecimiento

owl:sameAs
skos:exactMatch
wdt:P2799

3

Ontologías y vocabularios

2

Schema.org

RDA Registry

Modelo de Referencia Bibliotecaria de la IFLA

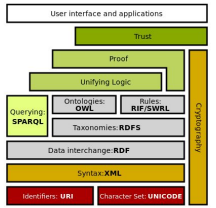
Modelo Conceptual para la Información Bibliográfica

TaDiRAH

Tecnologías y estándares

1

SPARQL, OWL, RDF, URIs



https://en.wikipedia.org/wiki/Semantic_Web_Stack

Web semántica y Linked Open Data

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

schema:CreativeWork

rdf:type

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

Miguel de Cervantes

schema:name

schema:author

owl:sameAs

LOD

https://www.wikidata.org/entity/Q5682

El Quijote

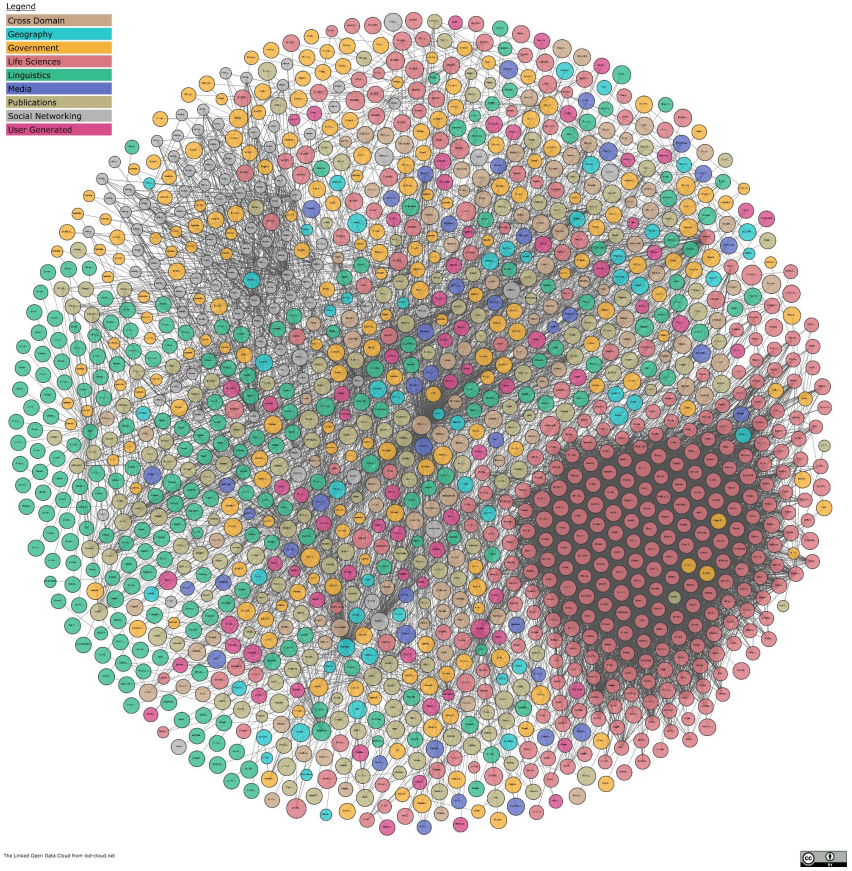
schema:title

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

Web semántica y Linked Open Data

https://lod-cloud.net/

Nov, 2025
1678 repositorios



Web semántica y Linked Open Data

europeana pro

https://pro.europeana.eu/page/apis

Data Foundry

Data collections from the National Library of Scotland

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

LIBRARY OF CONGRESS | LABS

https://id.loc.gov

KUNGL. BIBLIOTEKET

National Library of Sweden

https://libris.kb.se/sparql

SAAM

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

DATOS·BNE·es

https://datos.bne.es/

BnF Data

https://data.bnf.fr/

16 40 THE NATIONAL LIBRARY OF FINLAND

https://data.nationallibrary.fi

KB LAB

https://data.bibliotheken.nl

BRITISH LIBRARY

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

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

https://data.cervantesvirtual.com/

Bibliothèque nationale du Luxembourg

Open Data

https://data.bnl.lu/

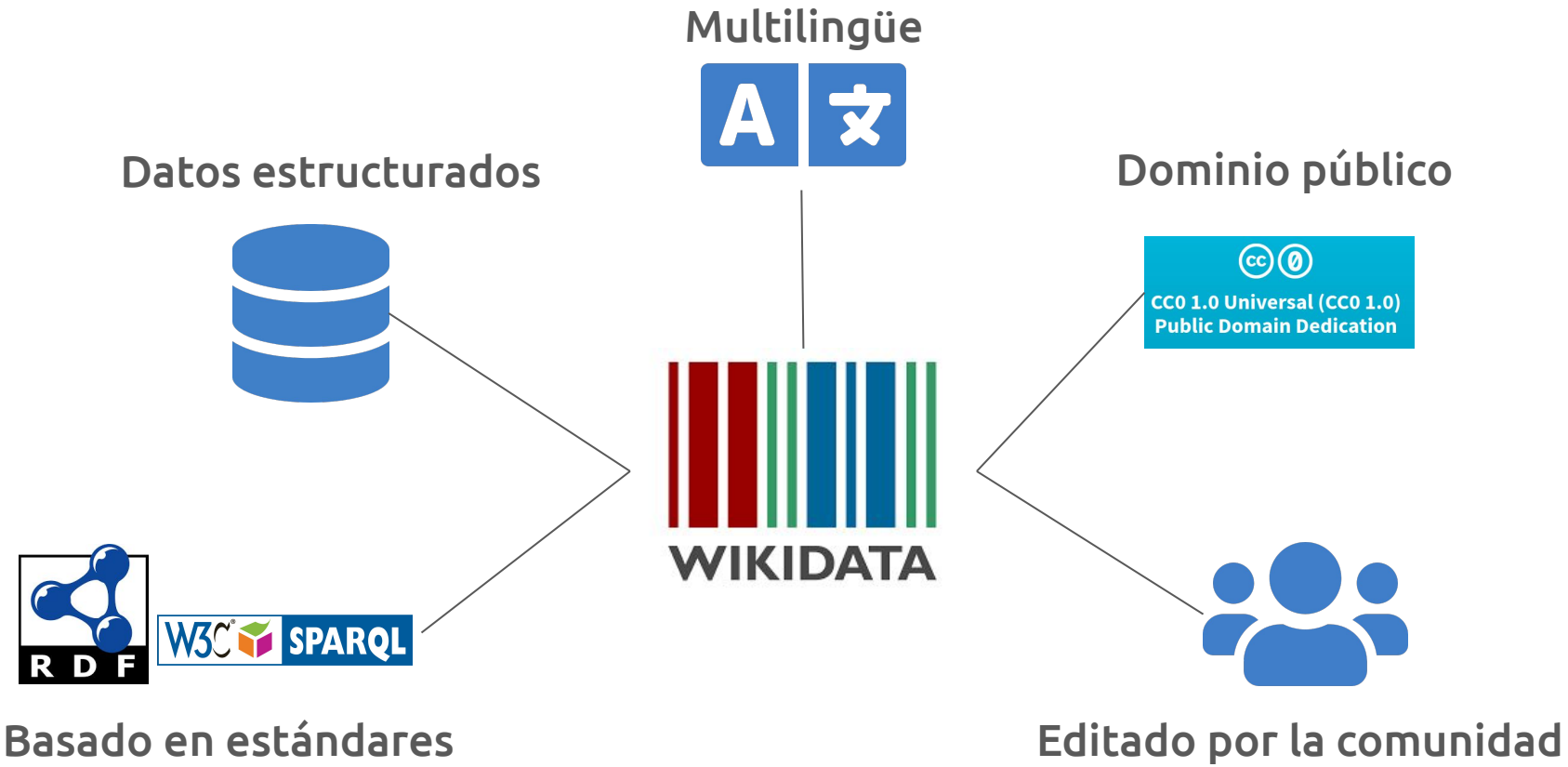
DEUTSCHE NATIONAL BIBLIOTHEK

https://www.dnb.de/EN/lds

Österreichische Nationalbibliothek

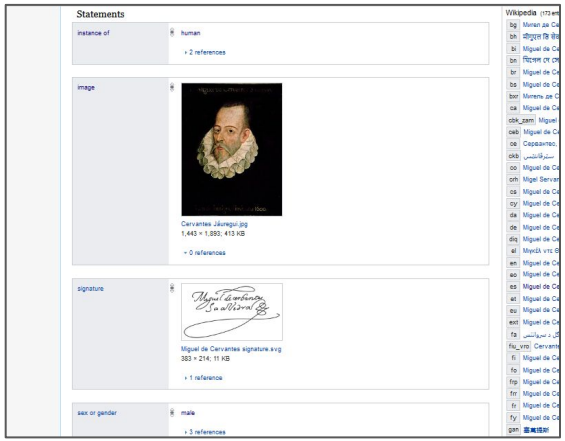
https://labs.onb.ac.at/en/dataset/lod

Wikidata


Multilingüe
Datos estructurados
Dominio público
CC0 1.0 Universal (CC0 1.0)
Public Domain Dedication
WIKIDATA
RDF
W3C SPARQL
Basado en estándares
Editado por la comunidad

Wikidata

Representación en forma de tripleta



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

Miguel de Cervantes nació en 1547

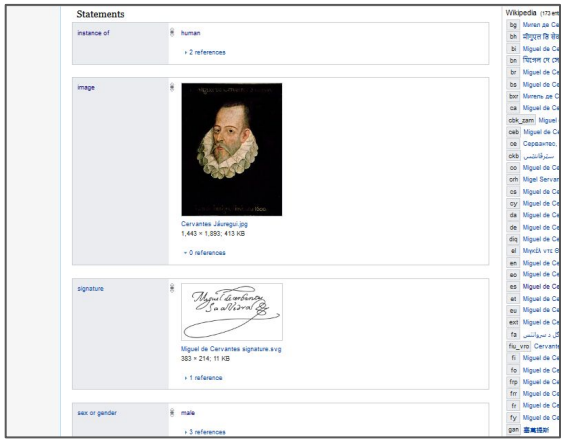
Sujeto - Predicado - Objeto

wd:Q5682 — wdt:P569 — 1547

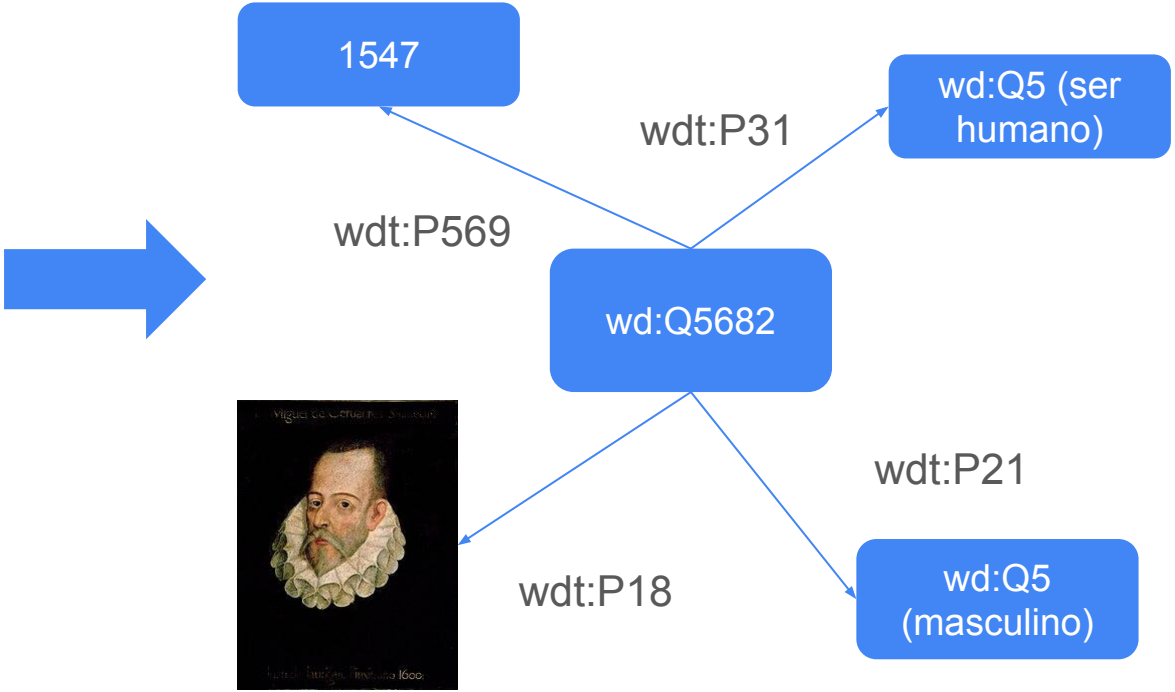
Uso de identificadores únicos para entidades y propiedades

Wikidata

Representación en forma de grafo

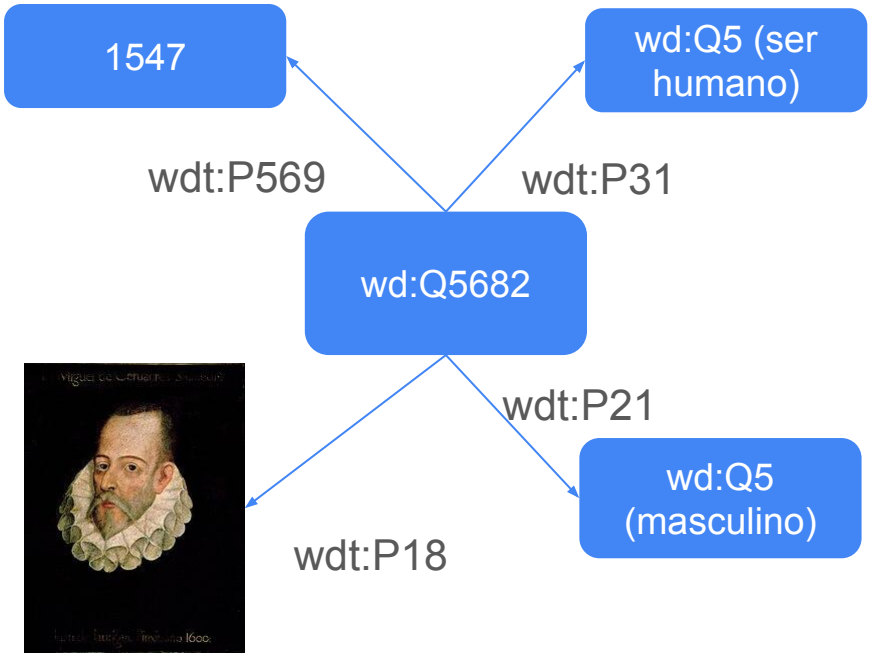


https://www.wikidata.org/wiki/Q5682



Wikidata

Espacios de nombre



wd = http://www.wikidata.org/entity/
wd:Q5682 = http://www.wikidata.org/entity/Q5682

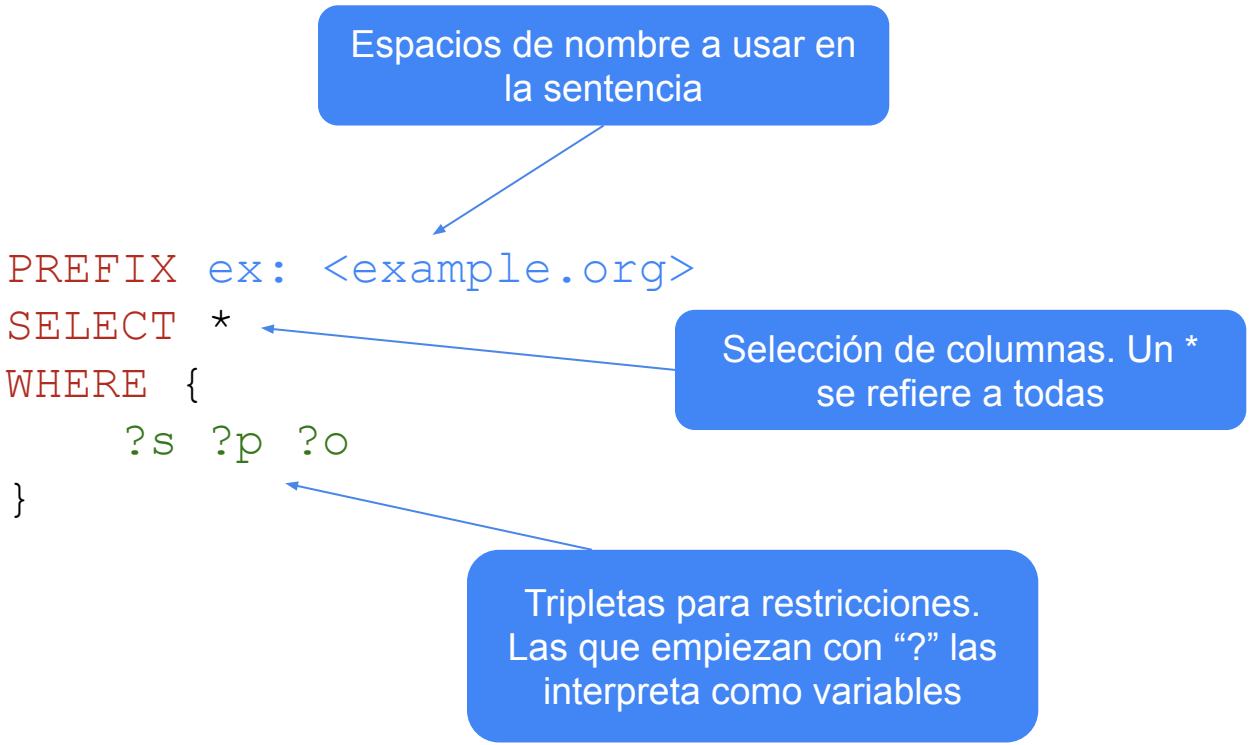
wdt = http://www.wikidata.org/prop/direct/
wdt:P31 = http://www.wikidata.org/prop/direct/P31

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

Consultas con SPARQL

Introducción



Consultas con SPARQL

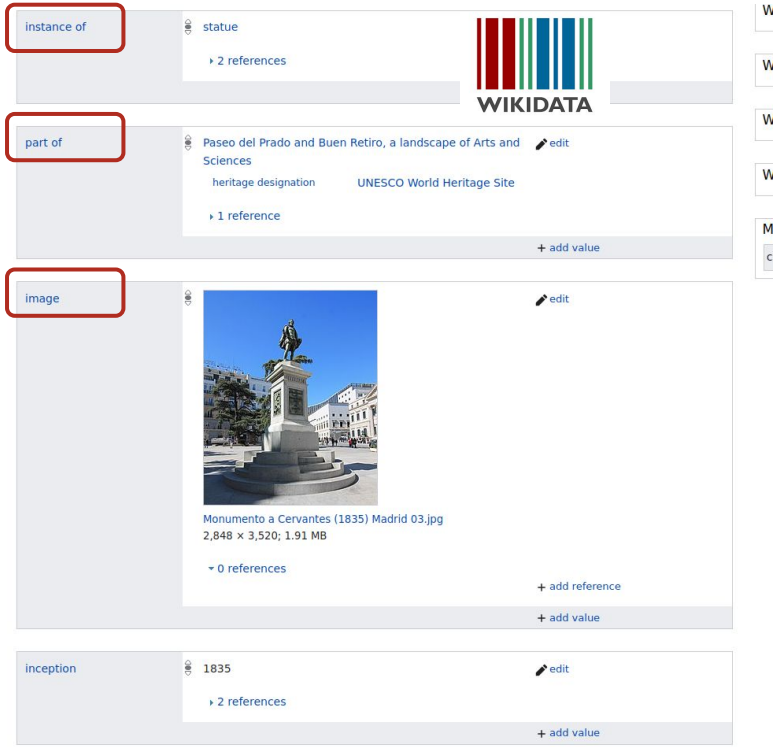
¿Qué estructura tienen los datos?

- 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



https://www.wikidata.org/wiki/Wikidata:List_of_properties

Consultas con SPARQL

Introducción

https://w.wiki/88yd

?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

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

https://w.wiki/85wT

```
SELECT ?autor ?autorLabel ?idbvmc
WHERE {
    ?autor wdt:P2799 ?idbvmc .

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

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

Instrucción FILTER permite aplicar filtros

```
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).

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

Consultas con SPARQL

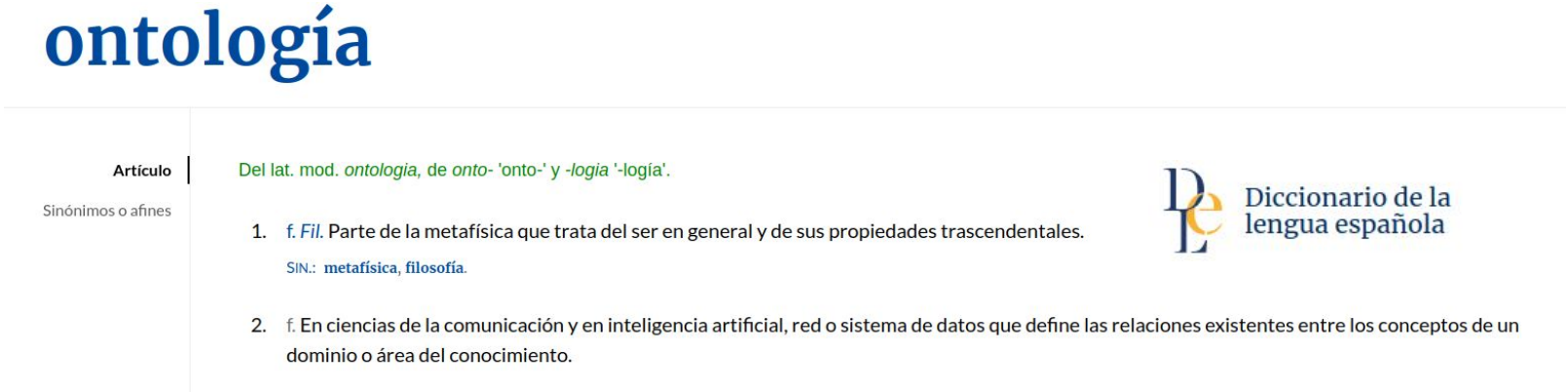
Uso avanzado

Más ejemplos de SPARQL

- 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

Ontologías y vocabularios

¿Qué es una ontología?



ontología

Artículo | Sinónimos o afines

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

Diccionario de la lengua española

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.

Clases y propiedades para describir los recursos

Ontologías y vocabularios

Ontologías para modelar y describir la información

Schema.org

https://schema.org/

bf: BIBFRAME

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

RDA Registry

http://www.rdaregistry.info/

IFLA International Federation of Library Associations and Institutions

IFLA Library Reference Model

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

europeana

EUROPEANA DATA MODEL

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

Data Catalog Vocabulary (DCAT)

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

CIDOC CRM CONCEPTUAL REFERENCE MODEL

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

Ontologías y vocabularios

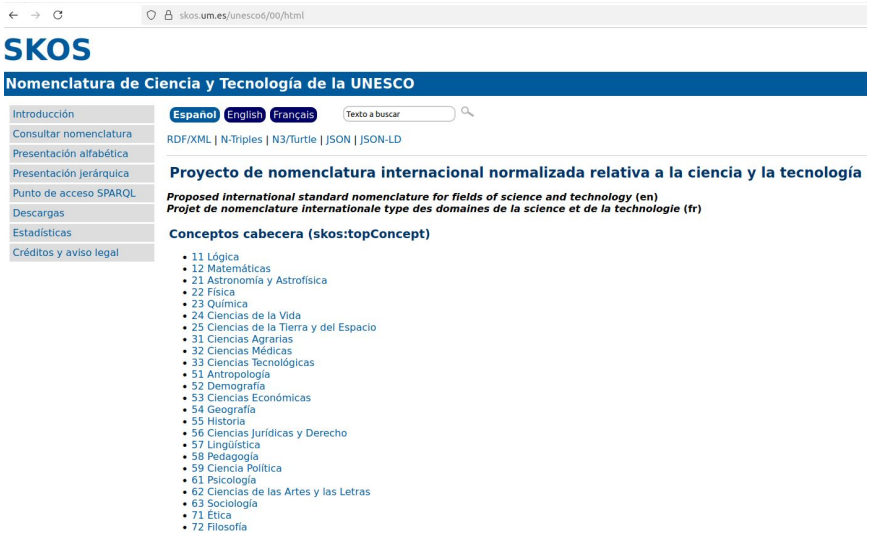
¿Qué es una vocabulario o taxonomía?

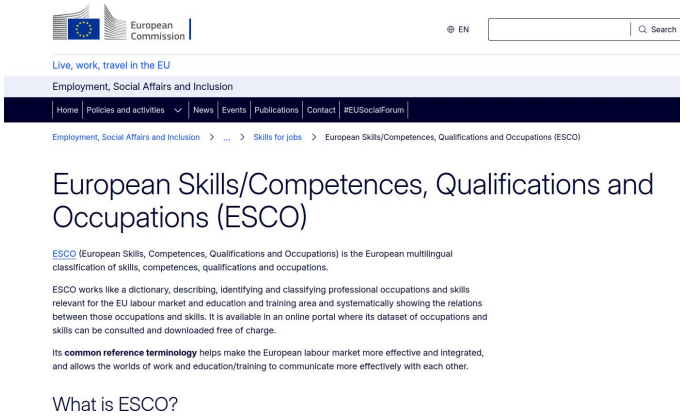
Términos para describir y clasificar los recursos

Ontologías y vocabularios

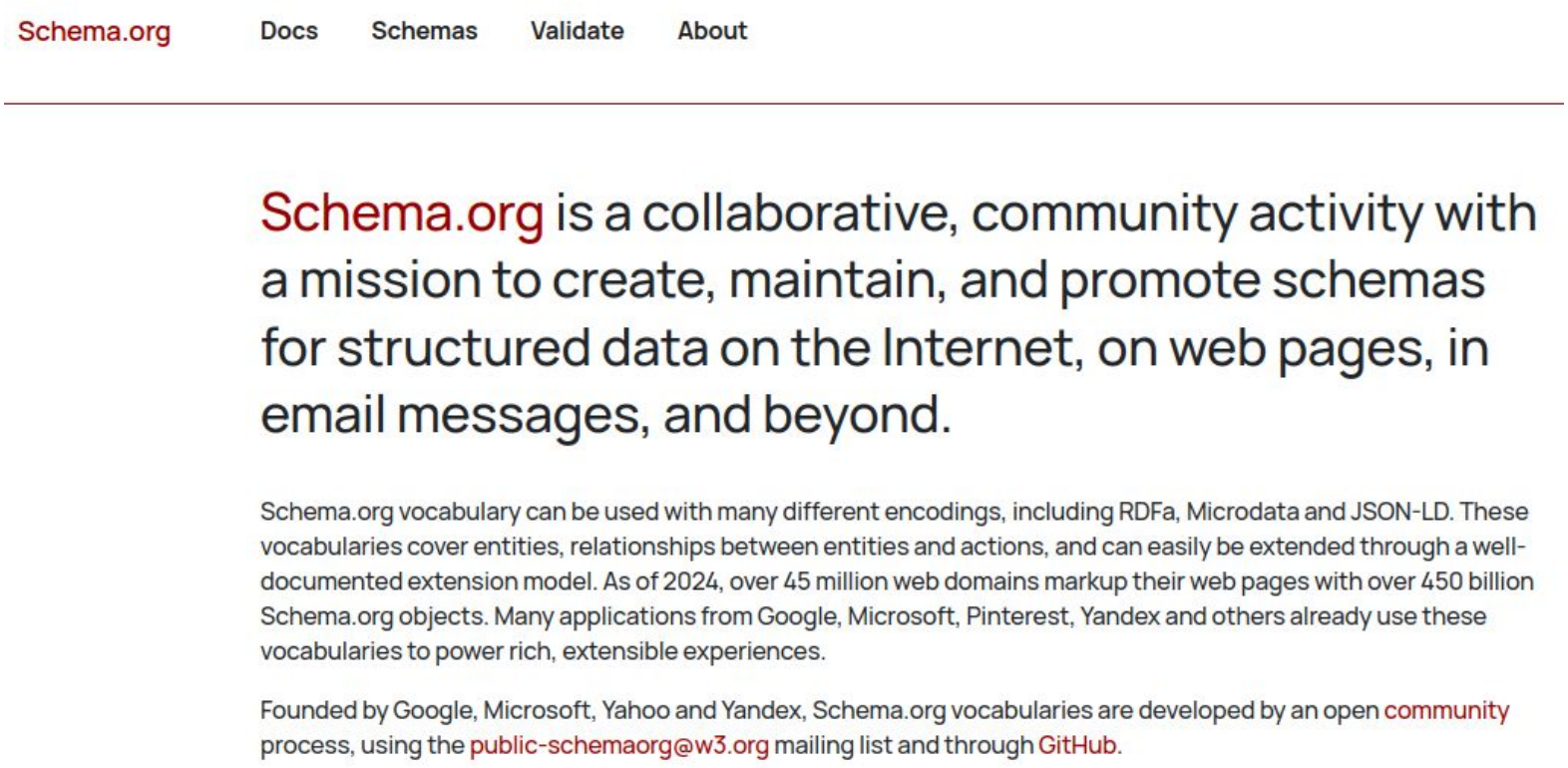
TADIRAH

THE TAXONOMY OF DIGITAL RESEARCH ACTIVITIES IN THE HUMANITIES





Schema.org


Schema.org
Docs
Schemas
Validate
About
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 mailing list and through GitHub.

Place

A Schema.org Type

Thing › Place

Entities that have a somewhat fixed, physical extension. [more...]

Property	Expected Type	Description
Properties from Place		
additionalProperty	PropertyValue	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. https://schema.org/width, https://schema.org/color, https://schema.org/gtin13, ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism.
address	PostalAddress or Text	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
amenityFeature	LocationFeatureSpecification	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.
branchCode	Text	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 http://www.starbucks.co.uk/store-locator/etc/detail/3047 the code "3047" is a branchCode for a particular branch.
containedInPlace	Place	The basic containment relation between a place and one that contains it. Supersedes containedIn. Inverse property: containsPlace
containsPlace	Place	The basic containment relation between a place and another that it contains. Inverse property: containedInPlace
event	Event	Upcoming or past event associated with this place, organization, or action. Supersedes events.

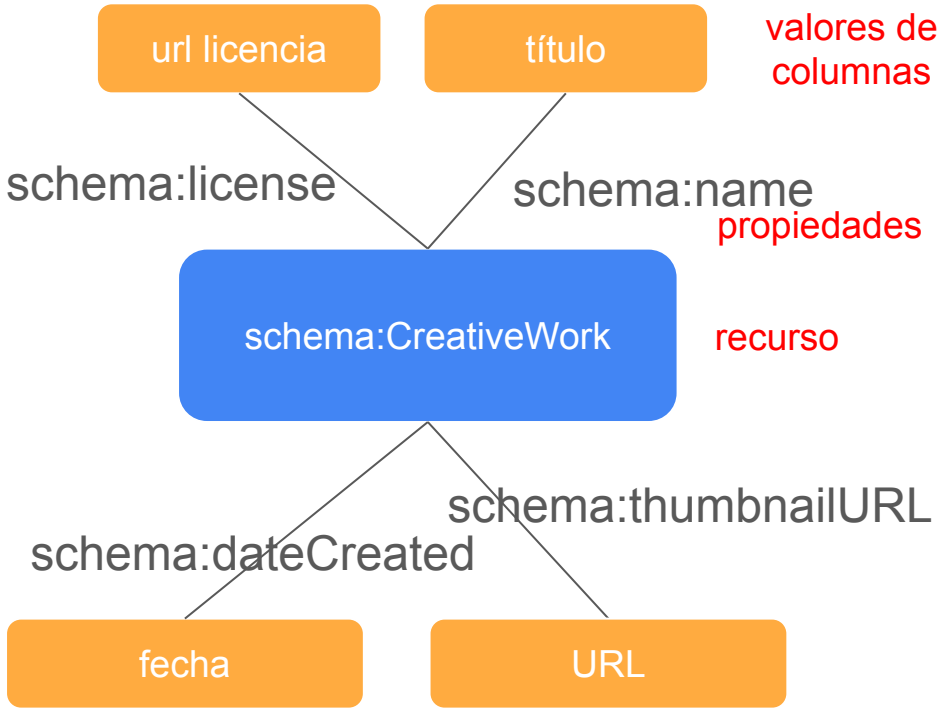
https://schema.org/Place

Ontologías

CSV metadatos:

- título
- licencia
- imagen
- fecha

Schema.org



Ontologías

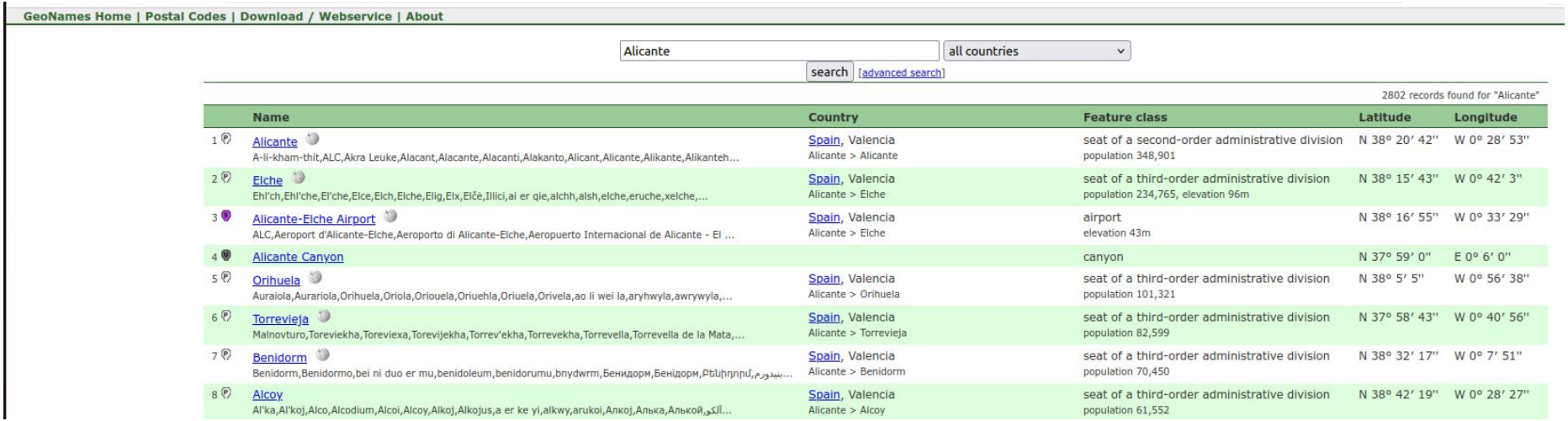
The KnowWhereGraph ontology

https://knowwheregraph.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 KWG 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	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 | all countries | search | [advanced search]

2802 records found for "Alicante"

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

GeoNames

https://www.geonames.org/

Ontologías

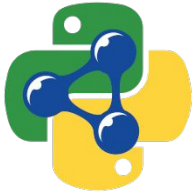
Creando nuestra propia ontología….

https://protege.stanford.edu/

Transformación

Tecnologías

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



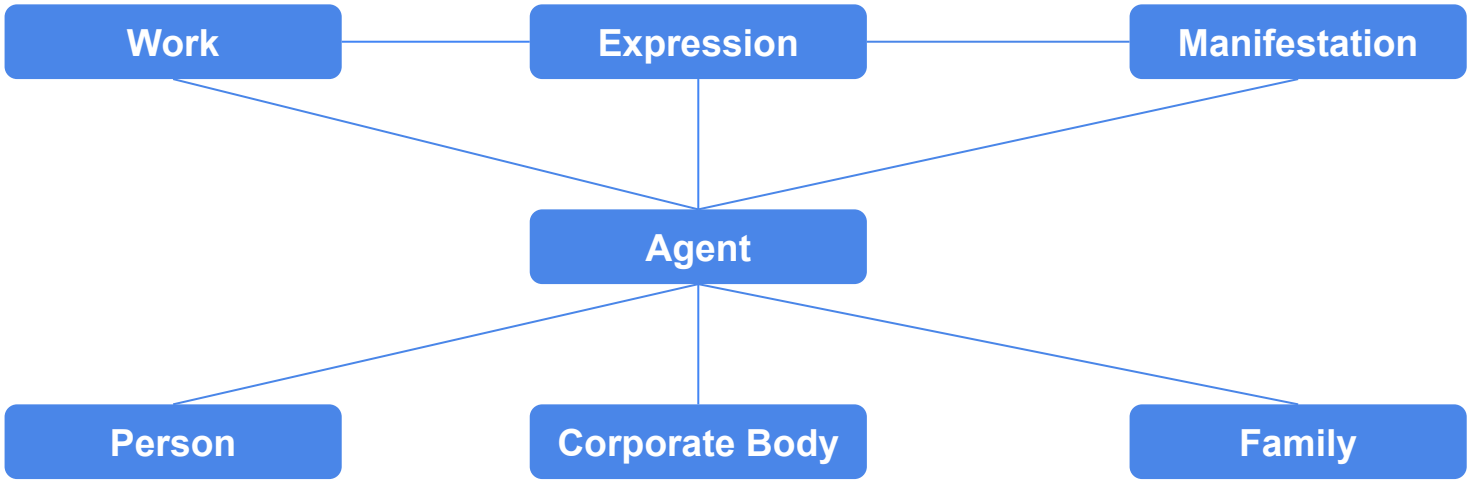
Transformación

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



RDA Registry

http://www.rdaregistry.info/Elements/c/

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

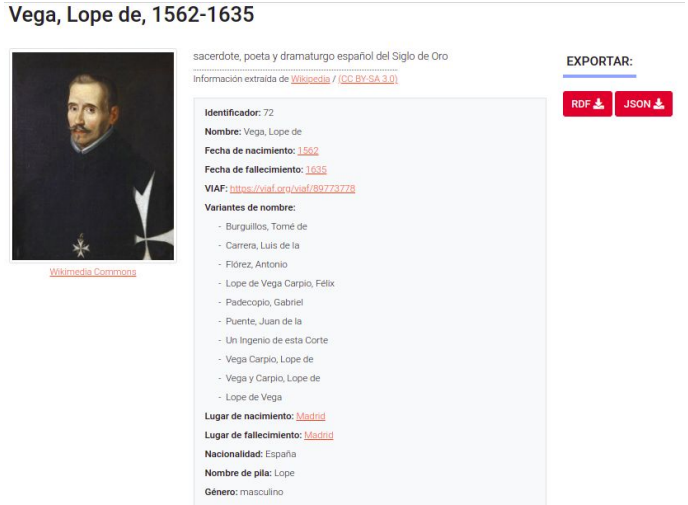
Wikidata y la Biblioteca Virtual Miguel de Cervantes

Propiedad	Descripción	Elementos
P2799	BVMC person ID	15054
P3976	BVMC work ID	764
P10834	BVMC organization ID	240

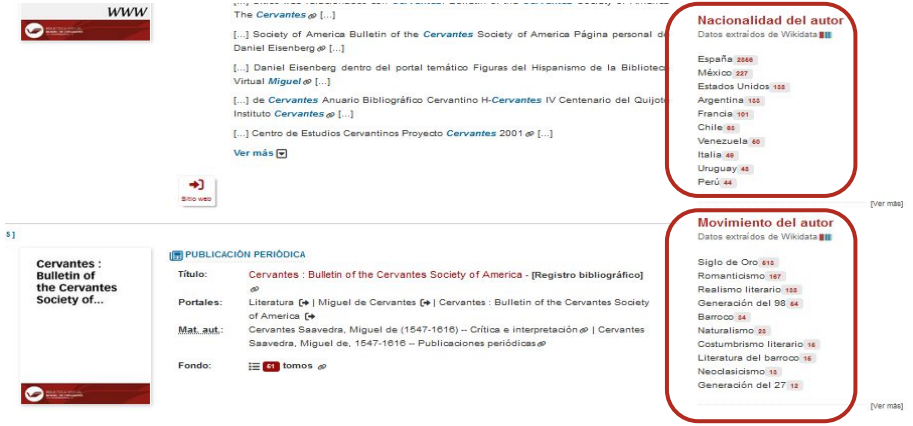
https://www.wikidata.org/wiki/Wikidata:Property_proposal

Transformación

Wikidata y la Biblioteca Virtual Miguel de Cervantes



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 installar 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

```xml
<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>

    <!-- jeva 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>
```

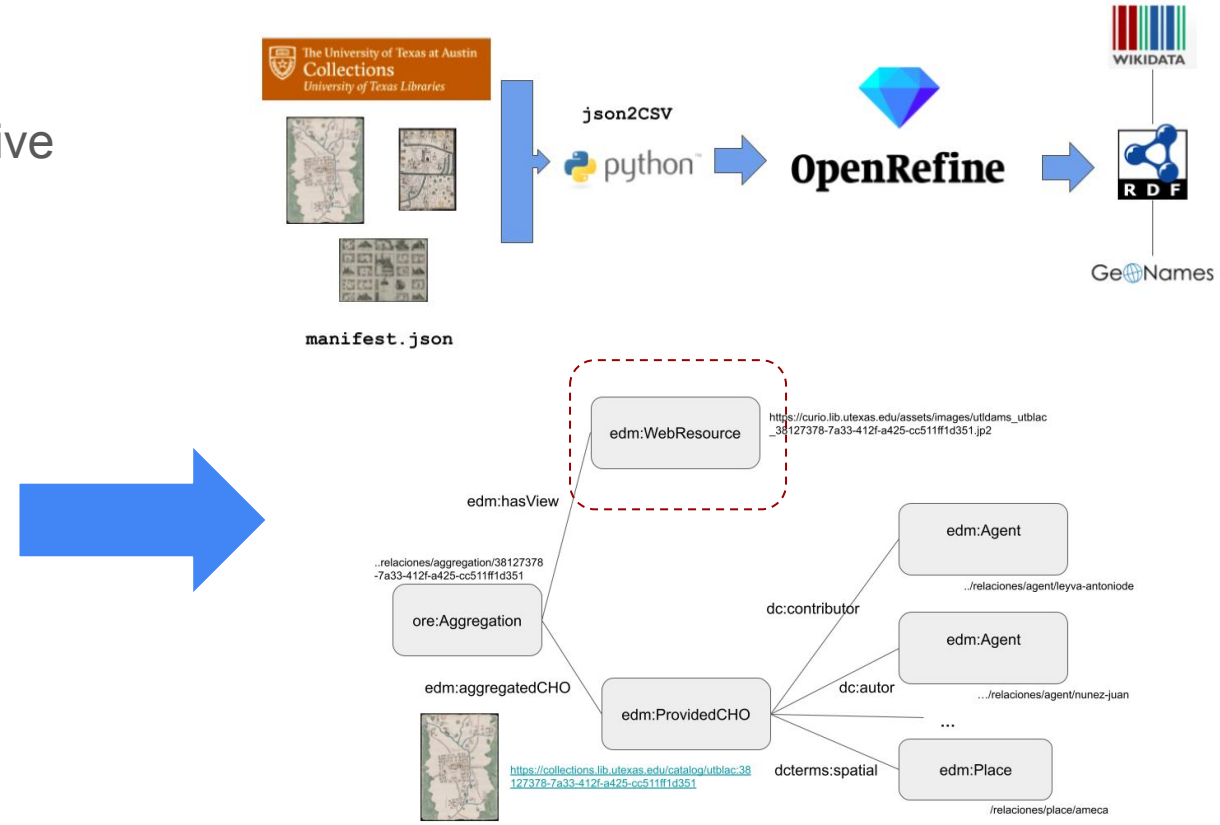
https://github.com/hibernator11/biblioteca-springboot

Transformación

Unlocking the Colonial Archive

Lancaster University

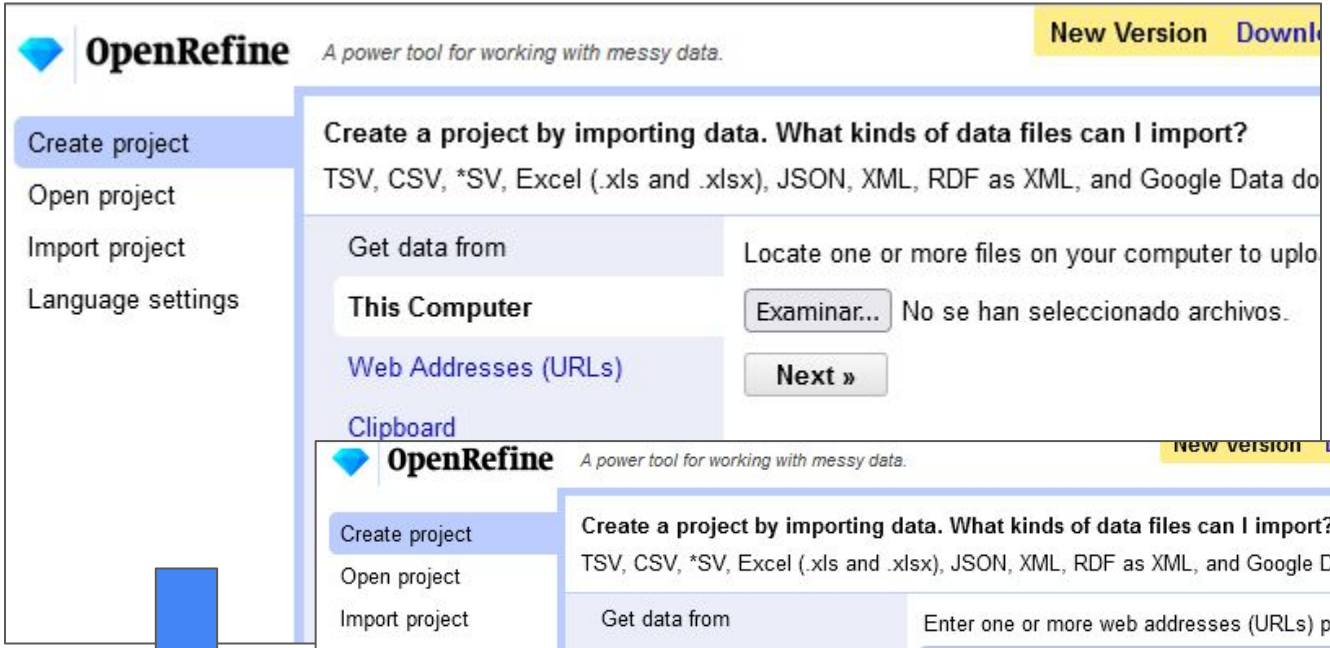
Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2023. An ontological approach for unlocking the Colonial Archive. J. Comput. Cult. Herit. Just Accepted (April 2023). https://doi.org/10.1145/3594727

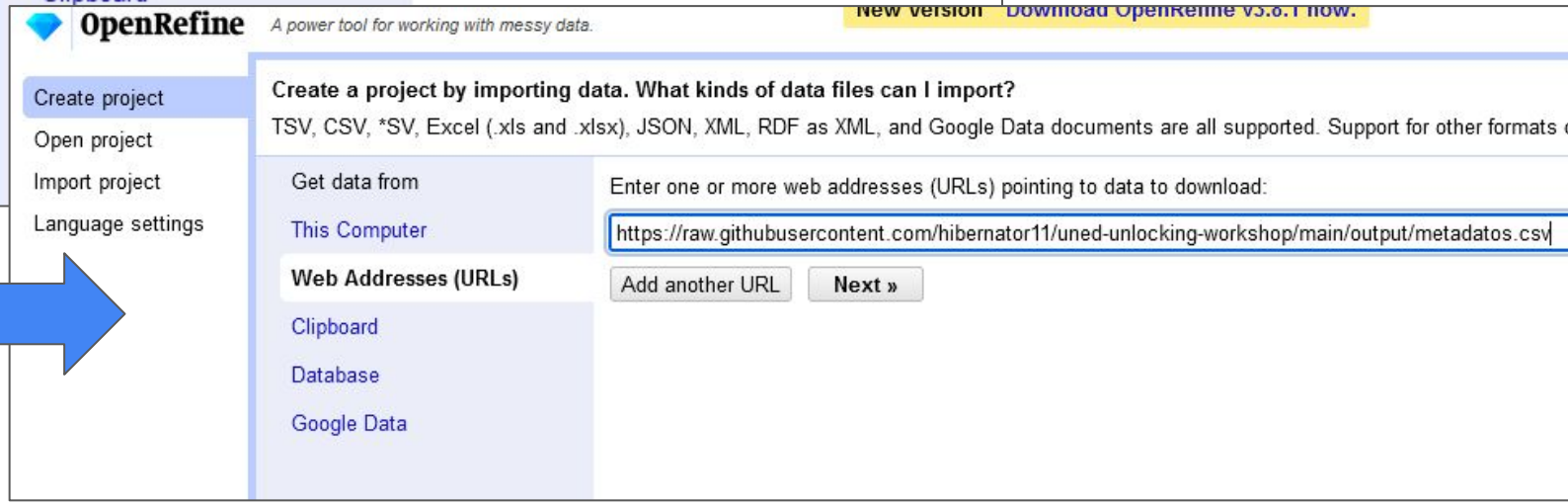


https://github.com/hibernator11/UCA-relacionesgeograficas

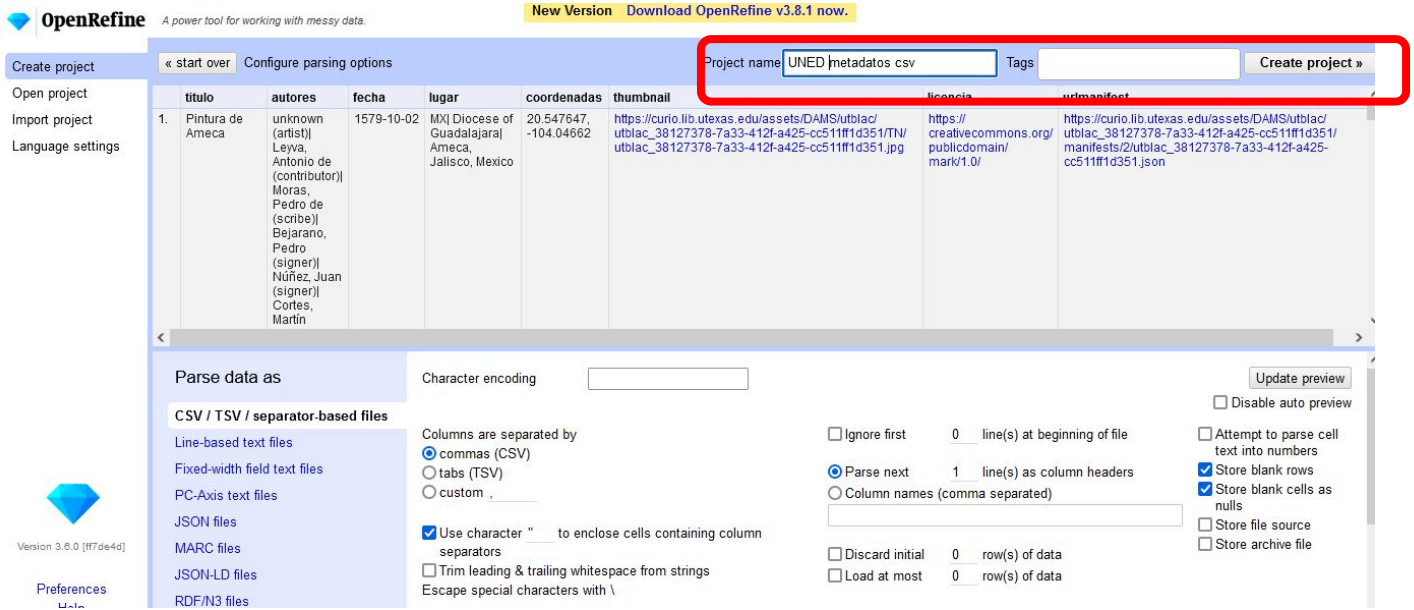
Transformación

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



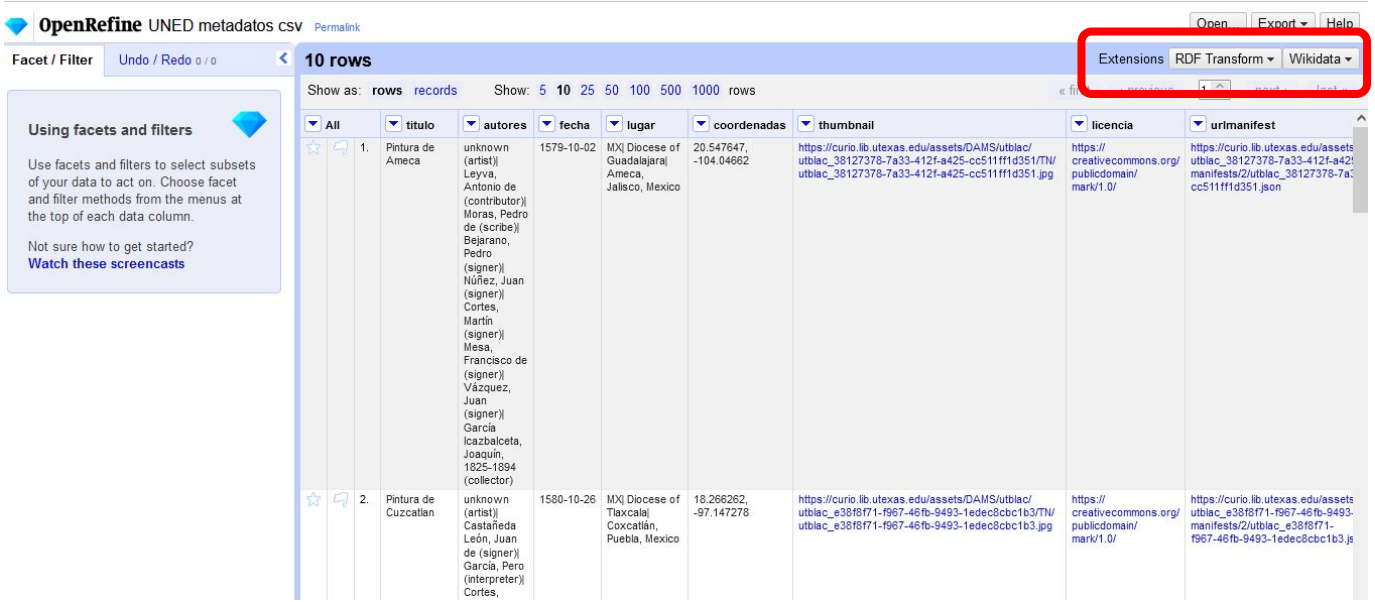


Abrimos el fichero CSV

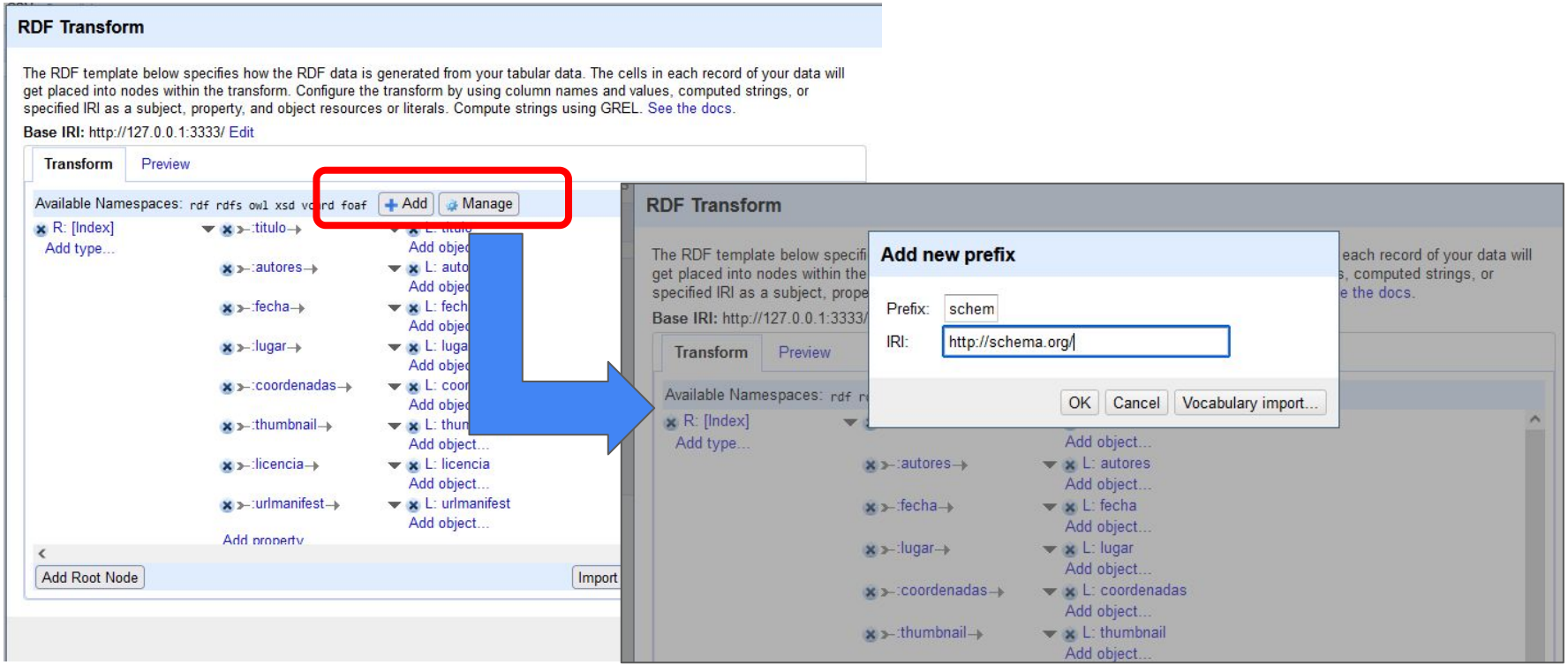


Le damos un nombre al proyecto

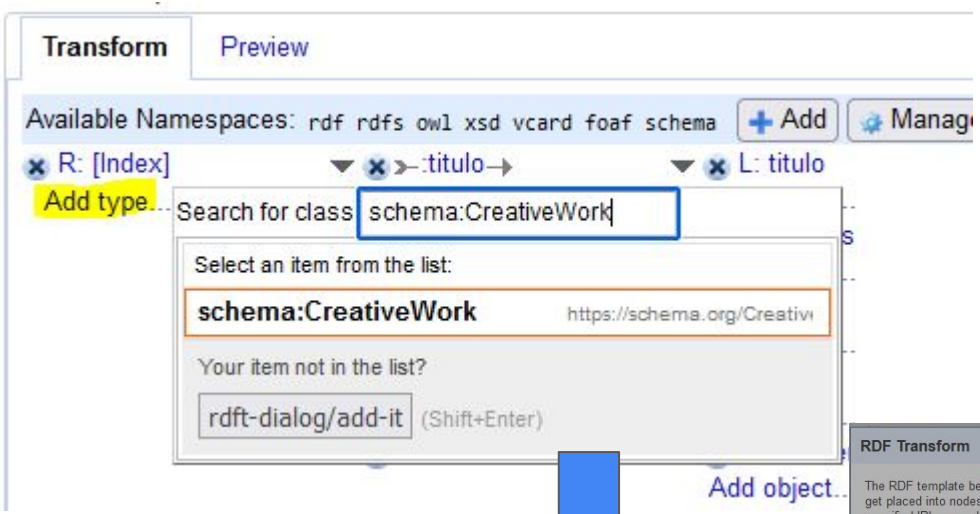
Le damos un nombre y creamos el proyecto desde el menú de la parte superior

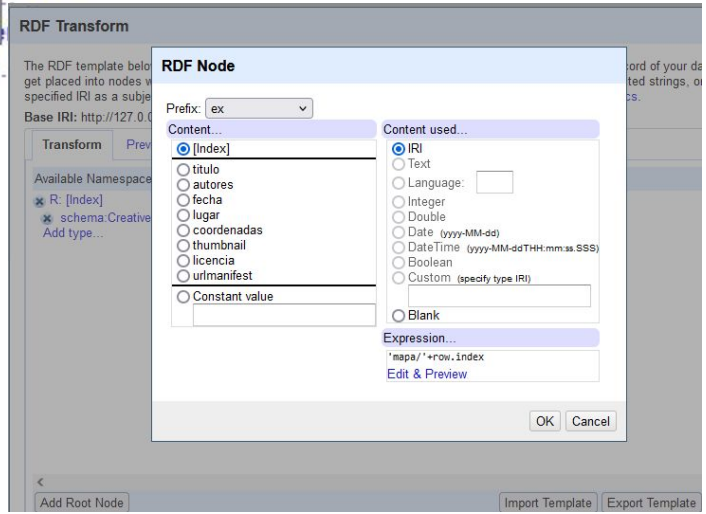


Vista de creación del proyecto, desde el menú superior podemos configurar la extensión para generar el formato RDF

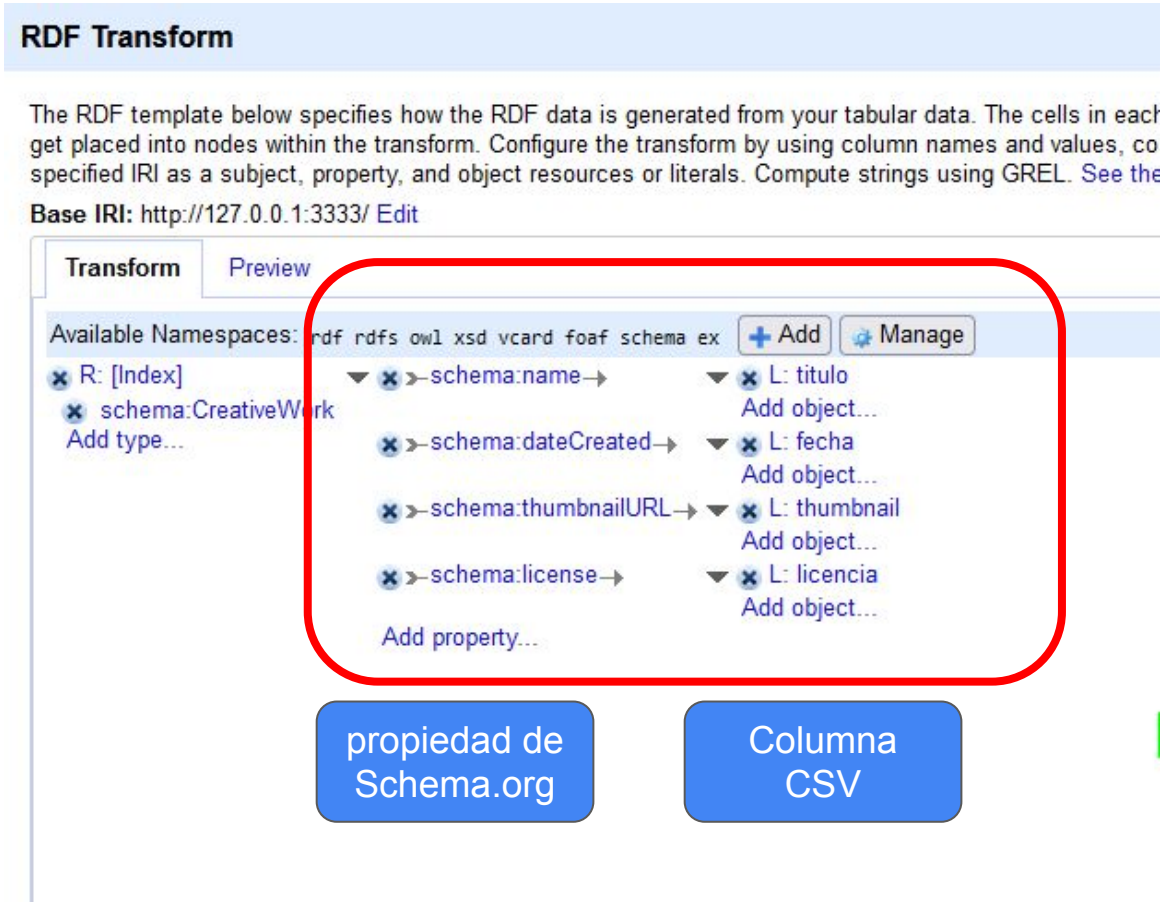


Añadimos el espacio de nombres Schema.org y https://example/ para crear las URIs (http://example/mapa/{xyz})

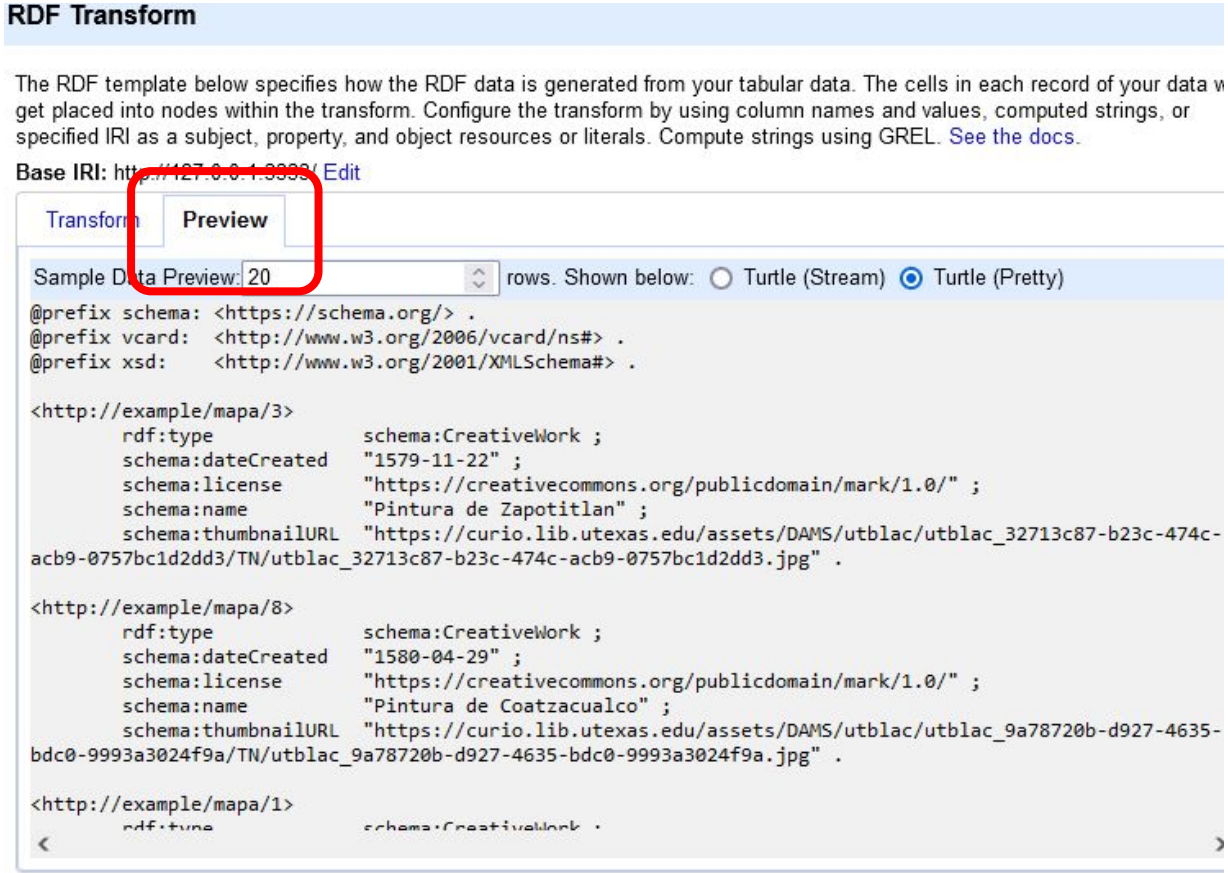




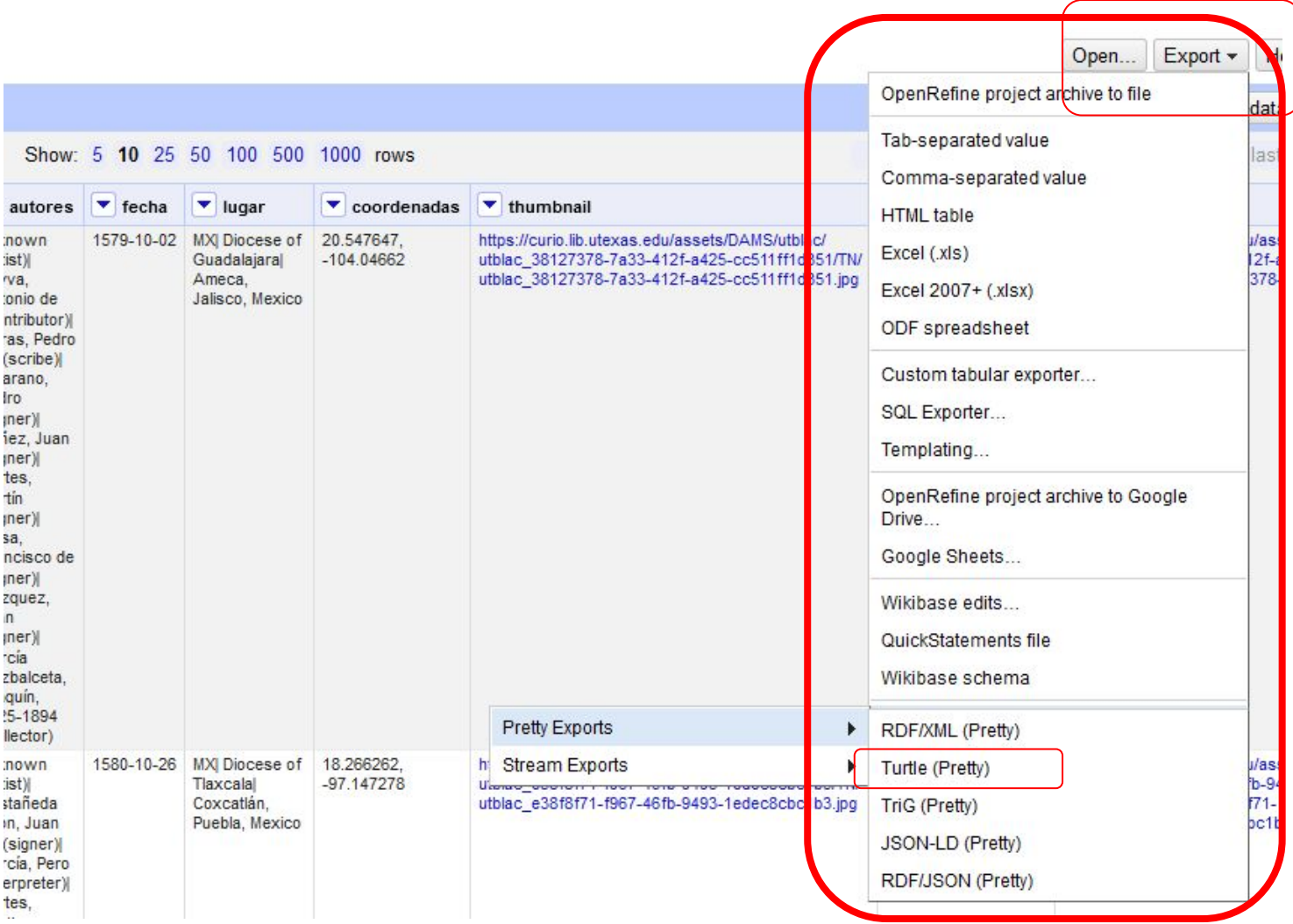
Añadimos la clase y definimos la URL



Añadimos las propiedades



Podemos ver el resultado desde la opción Preview



Podemos exportar los datos una vez finalizado el proceso de transformación

```
@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-ac

<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-bd

<http://example/mapa/1>
        rdf:type              schema:CreativeWork ;
        schema:dateCreated    "1580-10-26" ;
        schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
        schema:name           "Pintura de Cuzcatlan" ;
        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_metadatos_csv.ttl

Transformación - Código reproducible

GLAM WorkBench

LC Maps for Robots

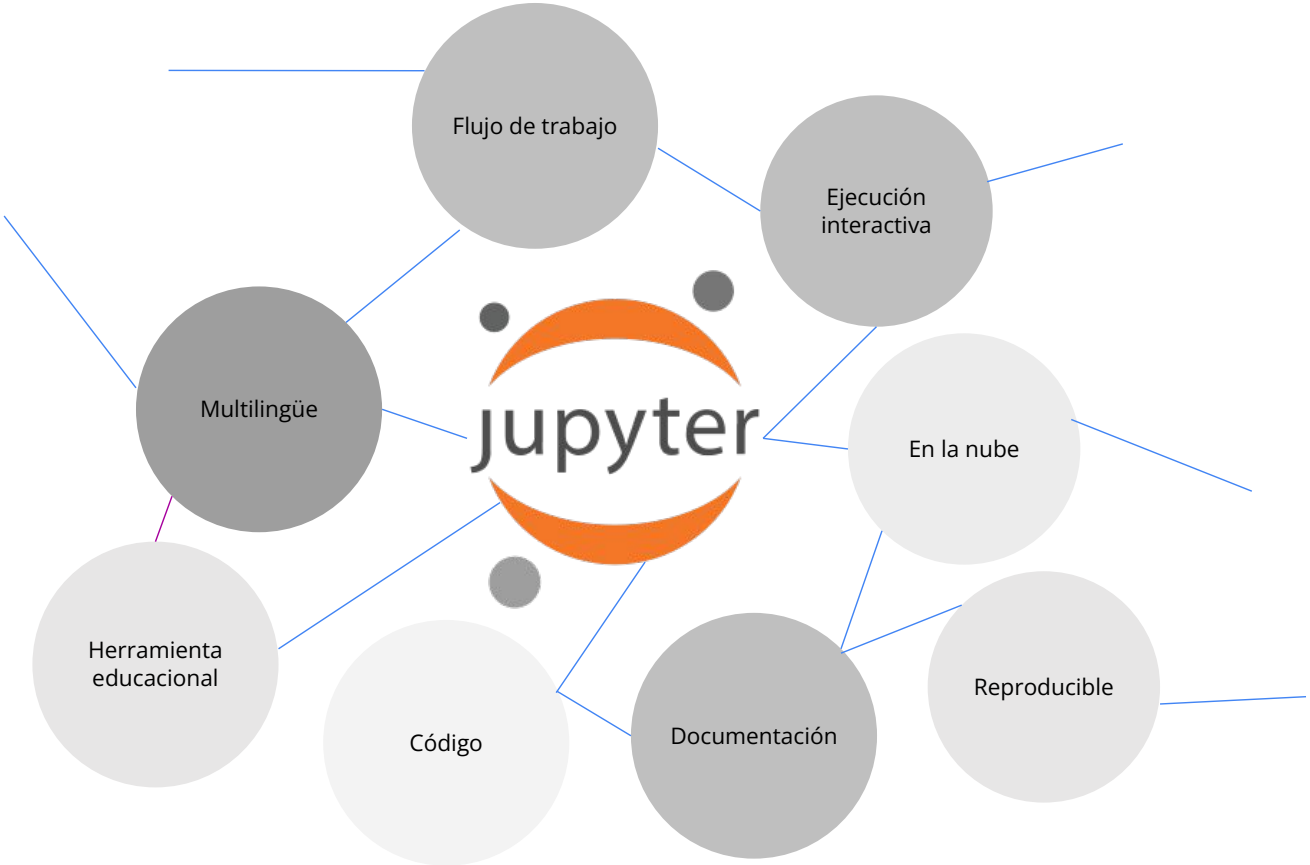
BL Labs

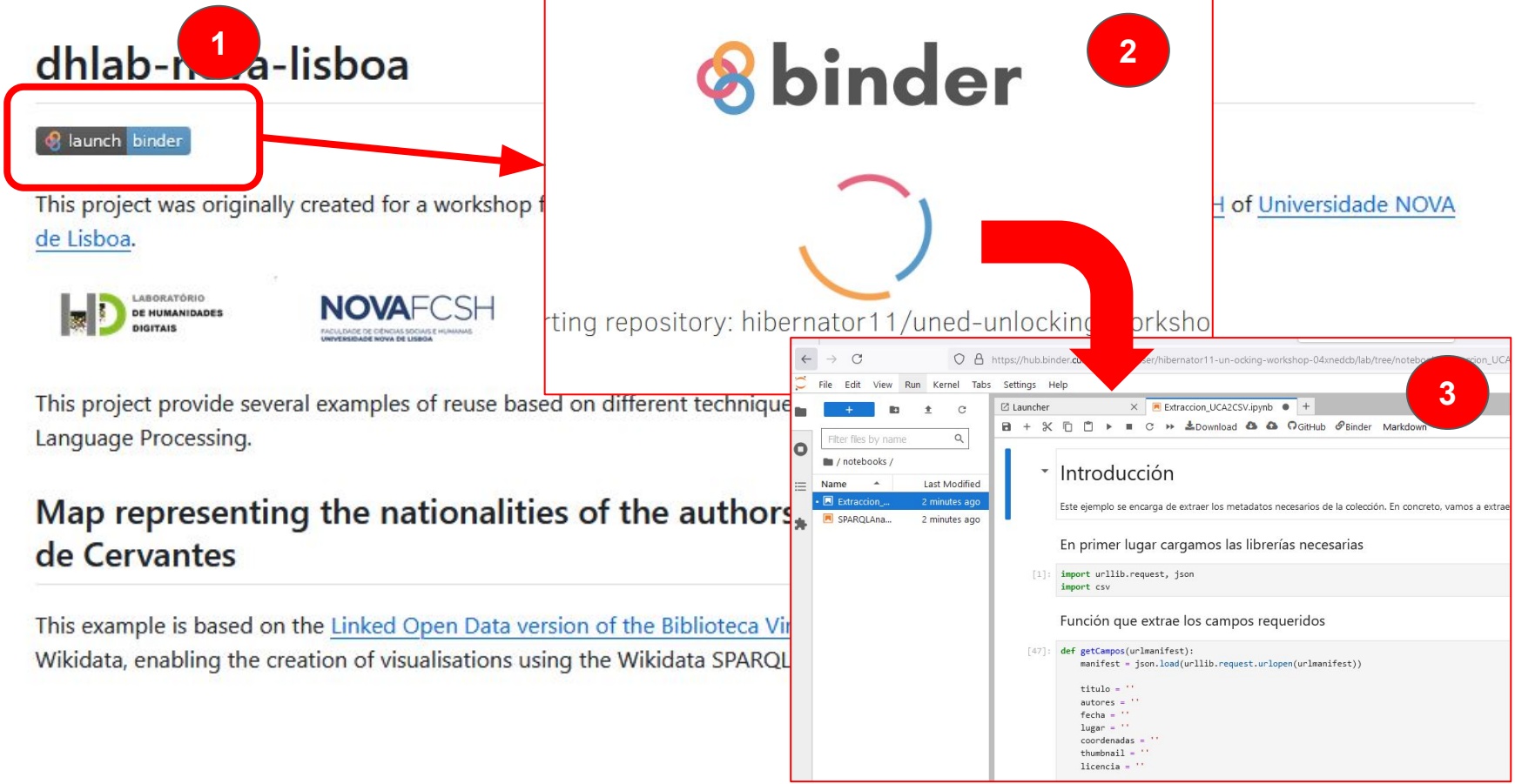
National Library of Scotland

Archives Unleashed Notebooks

LC - Newspapers navigator

Biblioteca Virtual Miguel de Cervantes





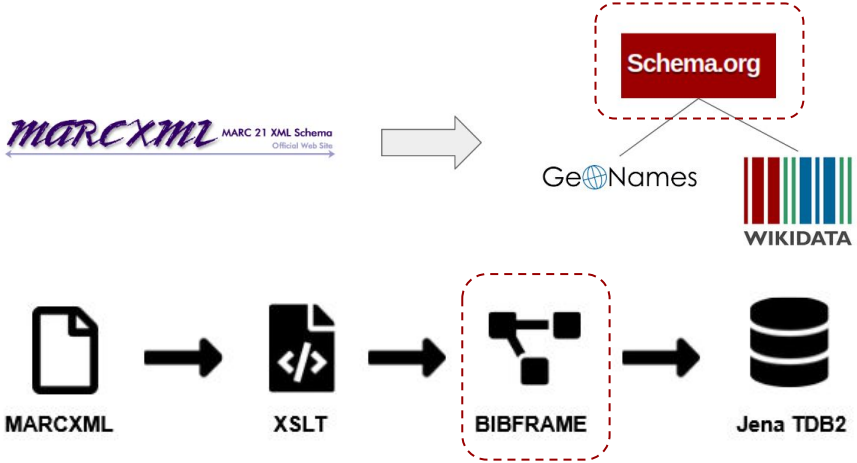
https://github.com/hibernator11/dhlab-nova-lisboa

Transformación

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

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

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

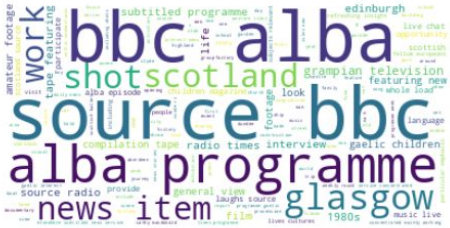
Moving Image Archive

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

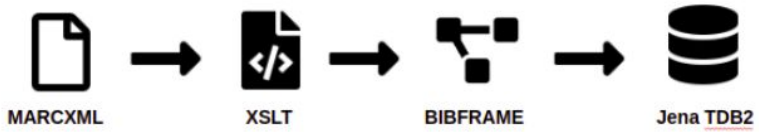
```
In [3…
nls = URIRef(domain + "organisation/nls")
g.add((nls, RDF.type, schema.Organization))
g.add((nls, schema.url, URIRef("https://www.nls.uk/")))
g.add((nls, schema.logo, URIRef("https://www.nls.uk/images/nls-logo.png")))
g.add((nls, schema.name, Literal("National Library of Scotland")))
g.add((nls,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

Java

Maven

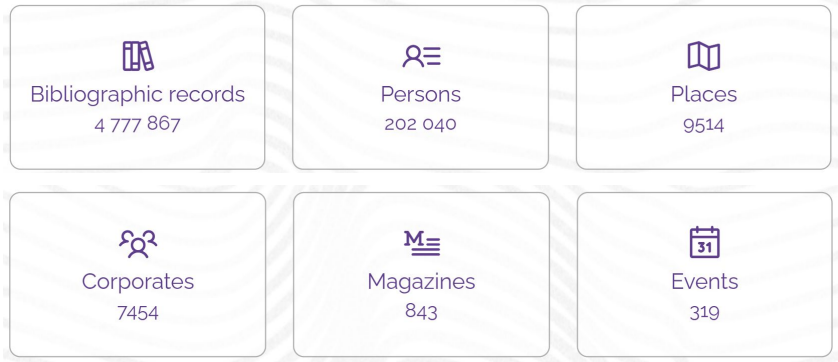
Apache Jena

Transformación

ATRIUM

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

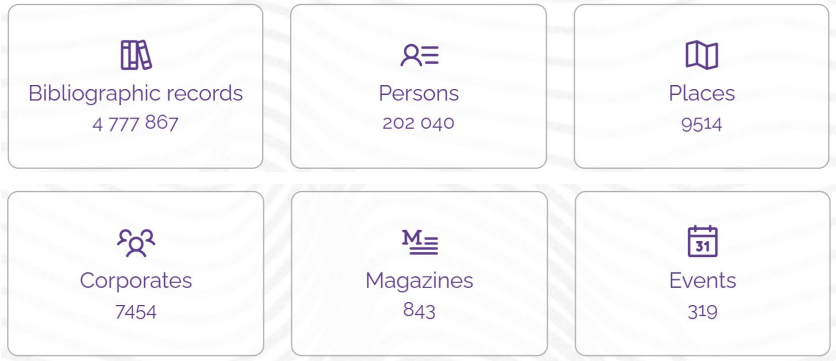
Transformación

DH2025
ACCESSIBILITY & CITIZENSHIP

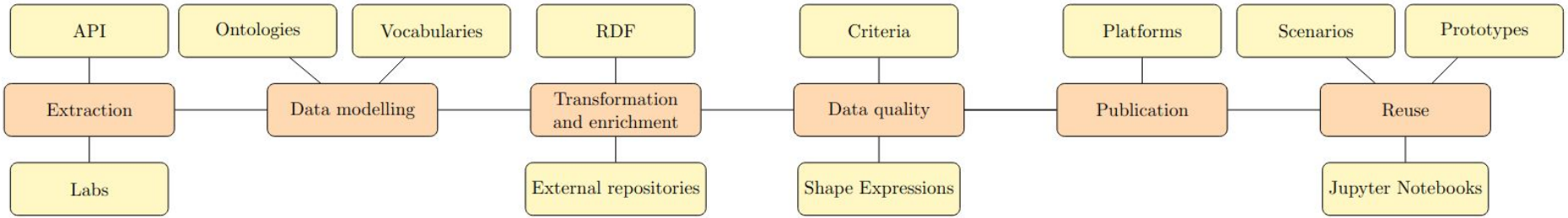
ATRIUM

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



ATRIUM project

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

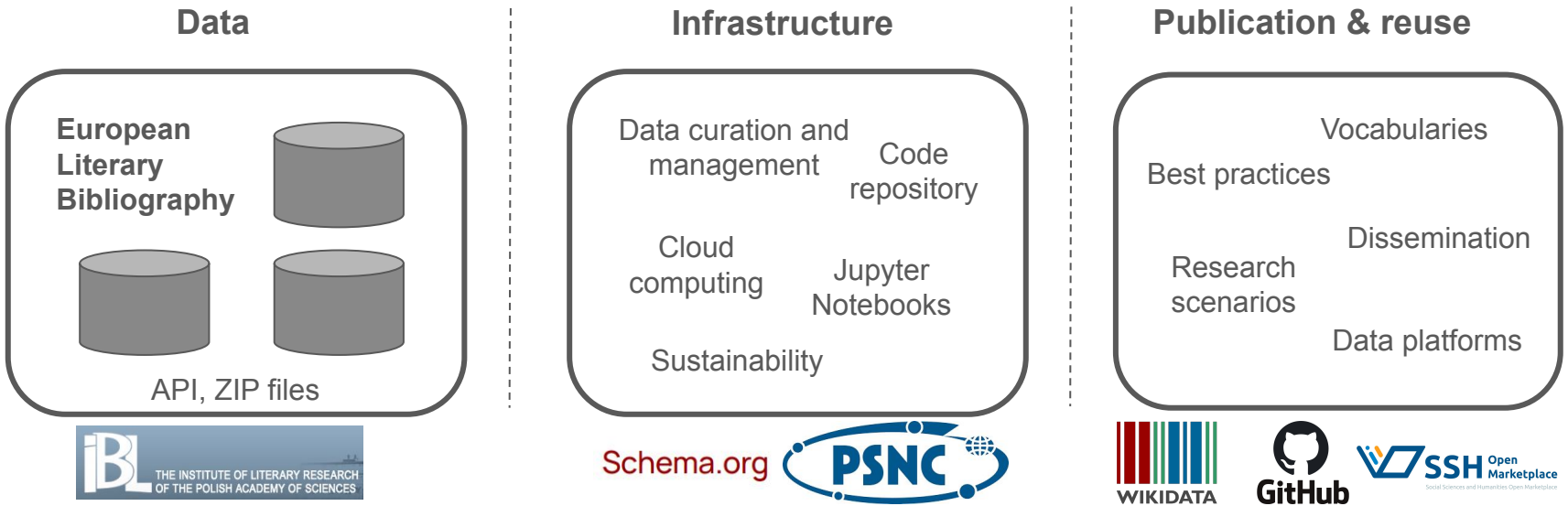
A reproducible approach

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.

ATRIUM jupyter THE INSTITUTE OF LITERARY RESEARCH OF THE POLISH ACADEMY OF SCIENCES PSNC

Transformación



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

Transformations: A DARIAH Journal

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.

Publication details

Published on June 16, 2025
Accepted on April 3, 2025
Submitted on November 12, 2024

Volume 1
Workflows

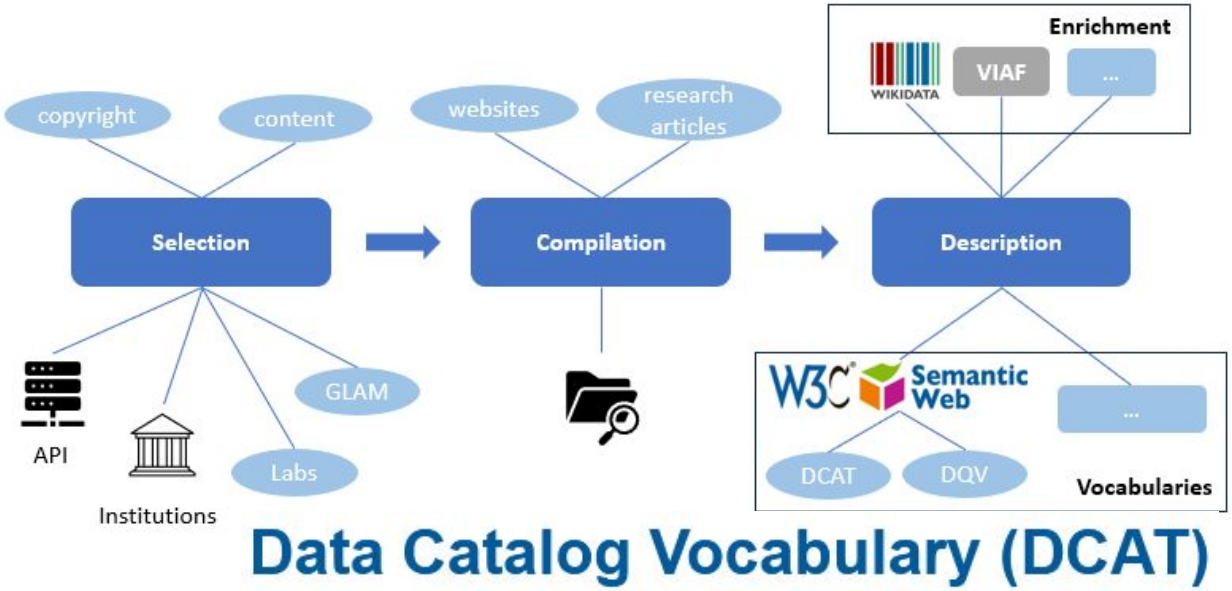
DOI
10.46298/transformations.14729

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

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

Transformación

Espacios de datos de patrimonio cultural



Workflow steps (10)

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

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

Transformación

Spanish-Civil-War-KGs

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

DOI 10.5281/zenodo.16747522

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

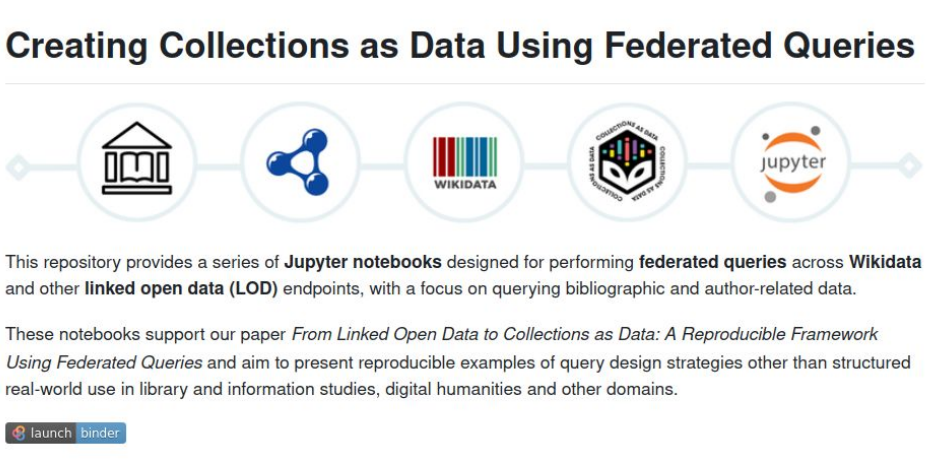
We provide 4 examples of use covering different topics and content:

- 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?

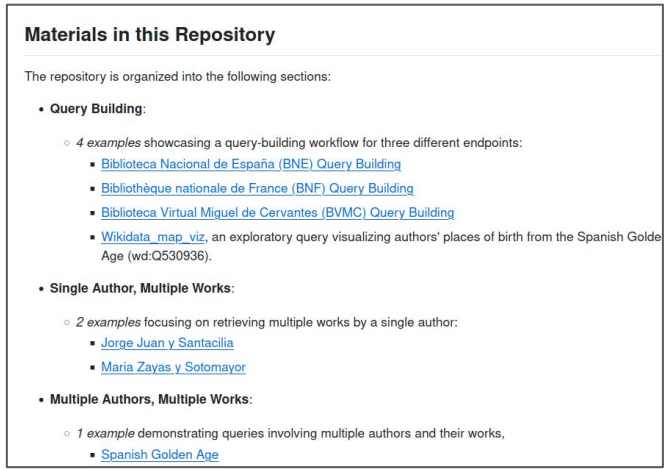


Creating Collections as Data Using Federated Queries

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 Golde 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]

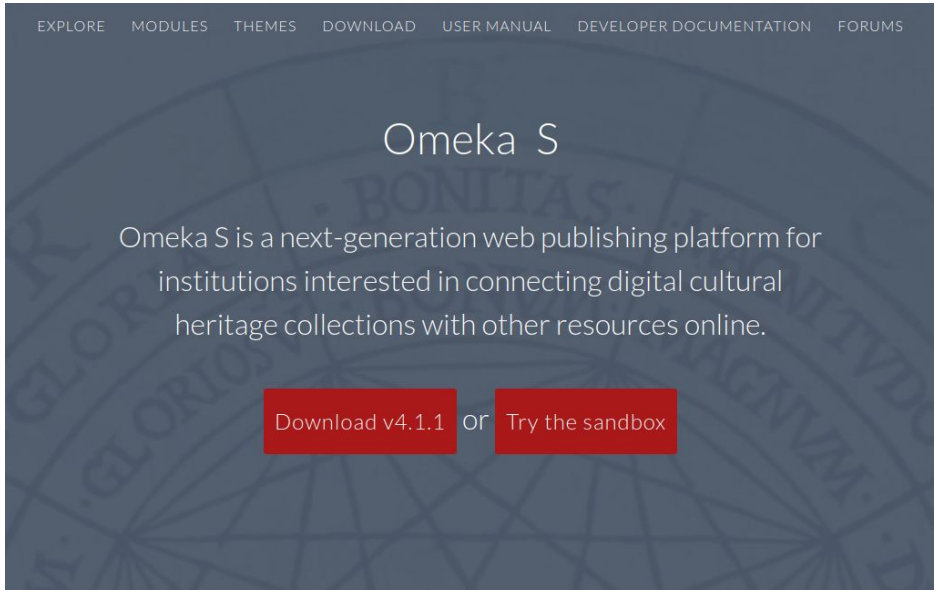
Transformación

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

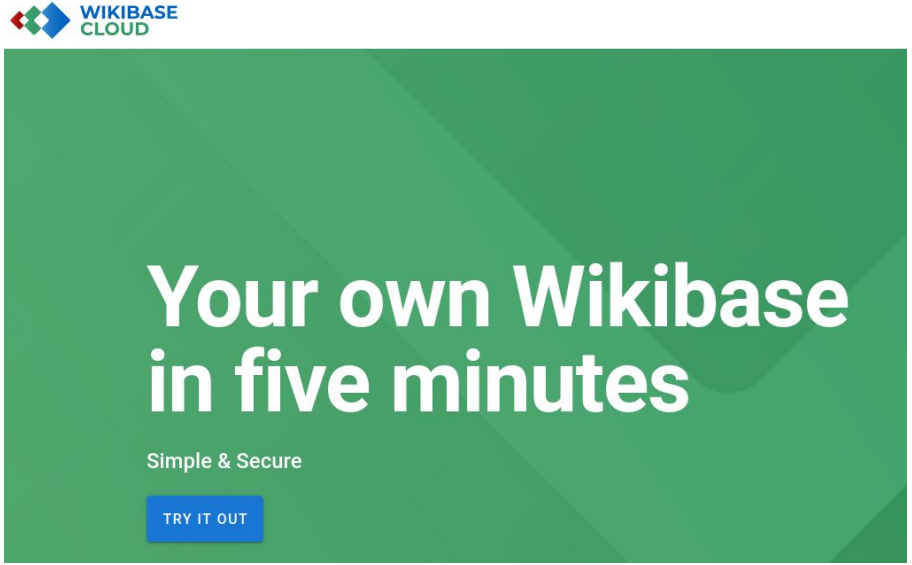
```
PREFIX bne-def: <https://datos.bne.es/def/>
PREFIX dcterms: <http://purl.org/dc/terms/>

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

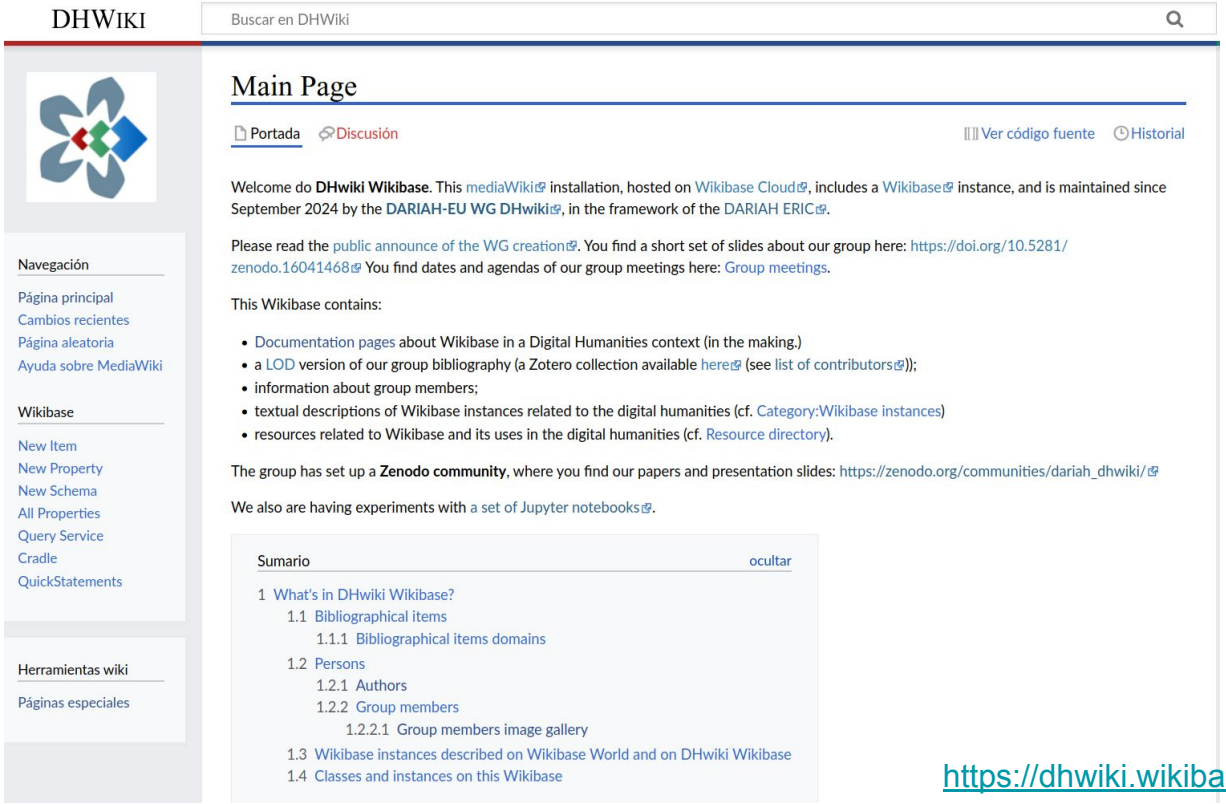


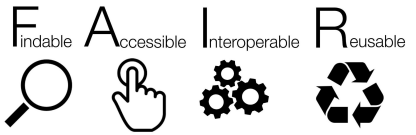
https://omeka.org/s/



https://www.wikibase.cloud/

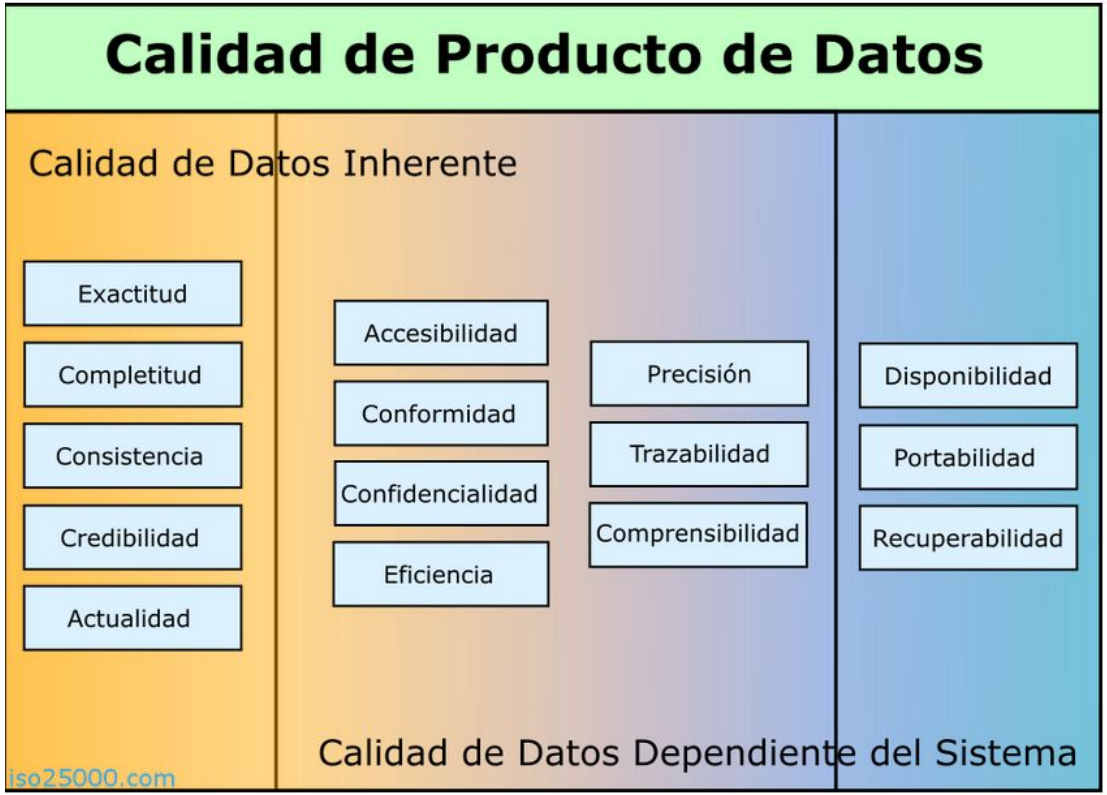
Transformación





https://dhwiki.wikibase.cloud/wiki/Main_Page

Calidad de datos



https://www.iso25000.com/index.php/normas-iso-25000/iso-25012

¿Qué medimos?

Calidad de datos

Ejemplos

SWJ SHEX https://shex.io/

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

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

jupyter

A reproducible framework to assess LOD

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

W3C SPARQL

Evaluating the quality of linked open data in digital libraries

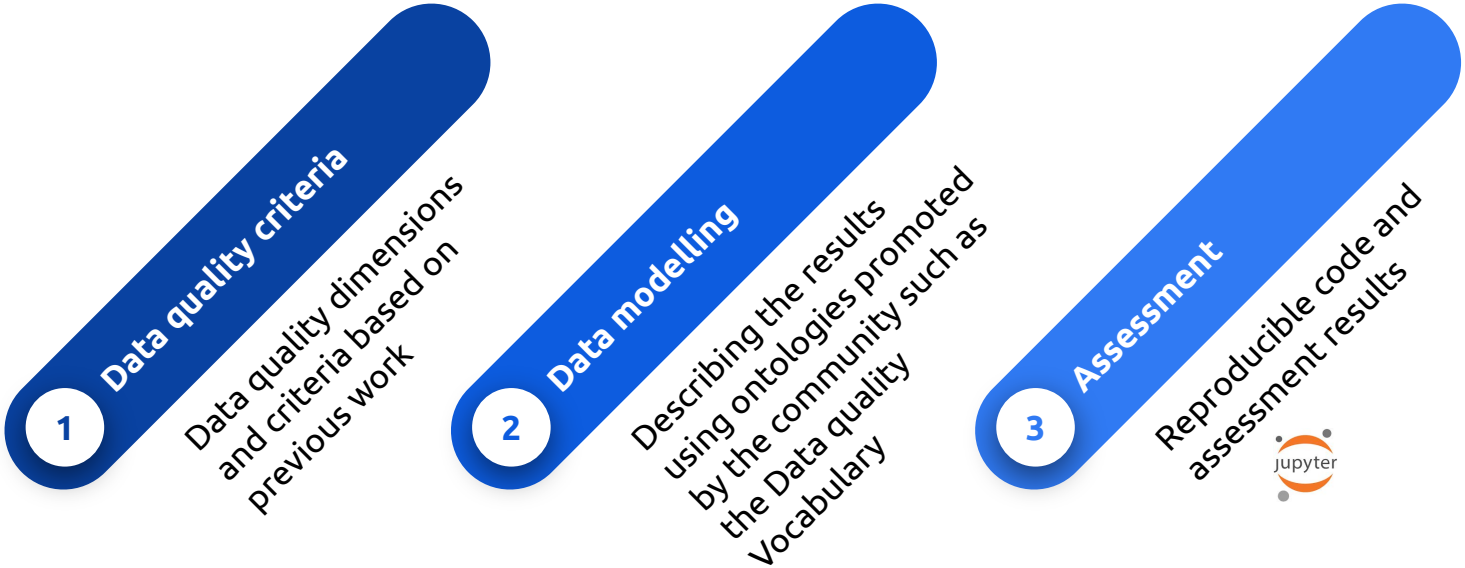
https://doi.org/10.1177/0165551520930951

JASIST
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

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



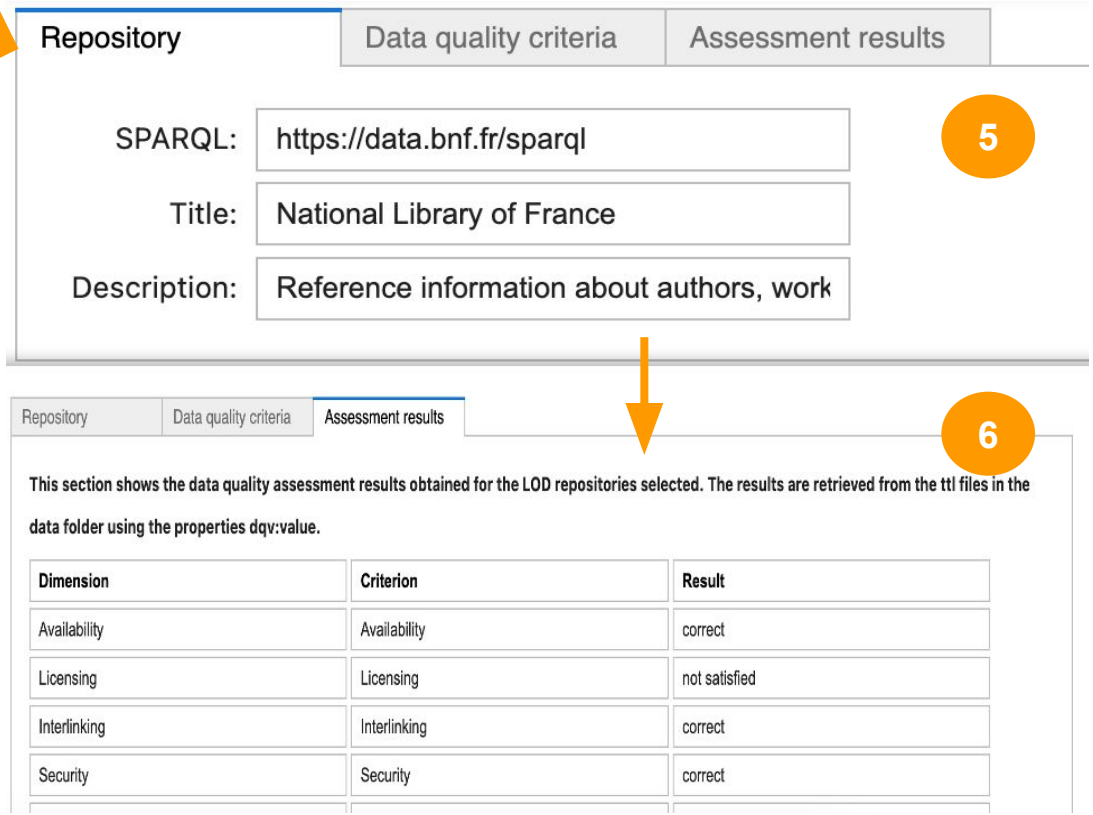
Assessment results

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

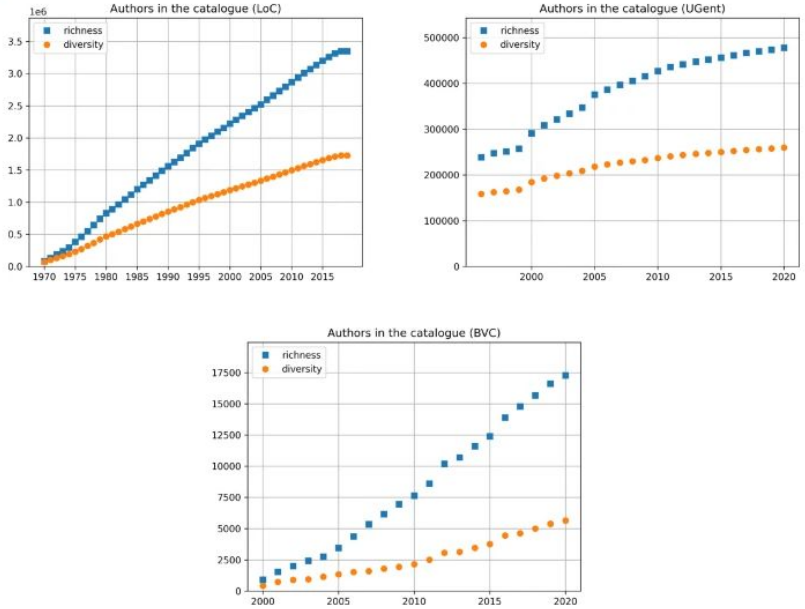


https://github.com/hibernator11/lod-quality-reproducible

Calidad

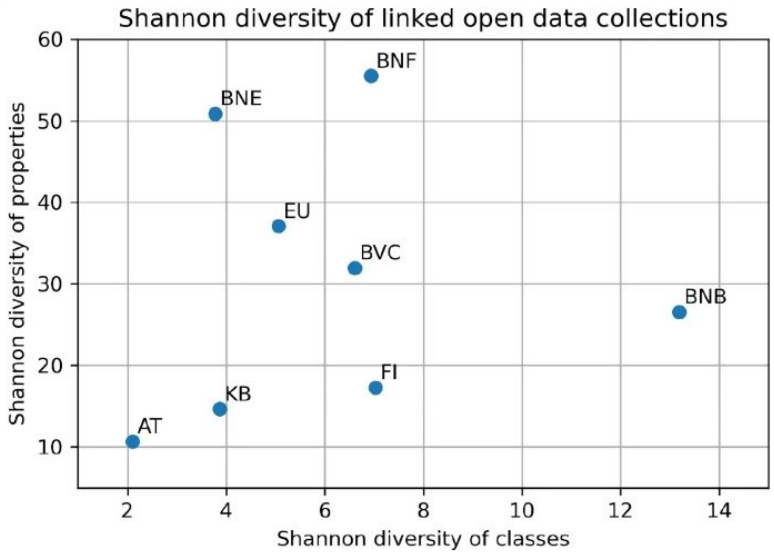
C. Carrasco, R., Candela, G. & Marco-Such, M. Measuring the diversity of data and metadata in digital libraries. *Int J Digit Libr* **26**, 5 (2025). https://doi.org/10.1007/s00799-025-00411-1

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

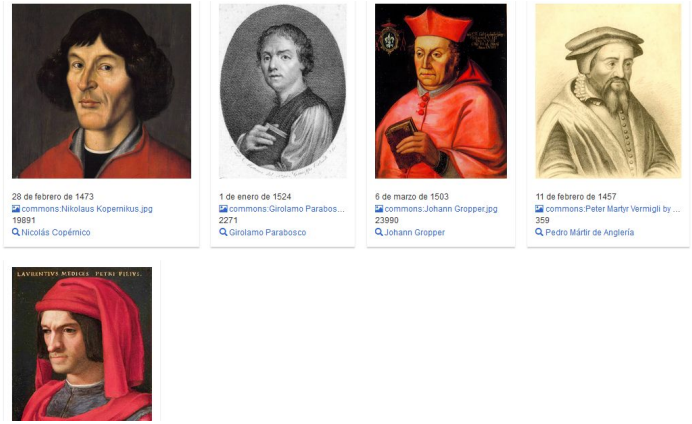
Visualización

Visualizaciones

- #defaultView:ImageGrid

https://w.wiki/8922

```
#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
```

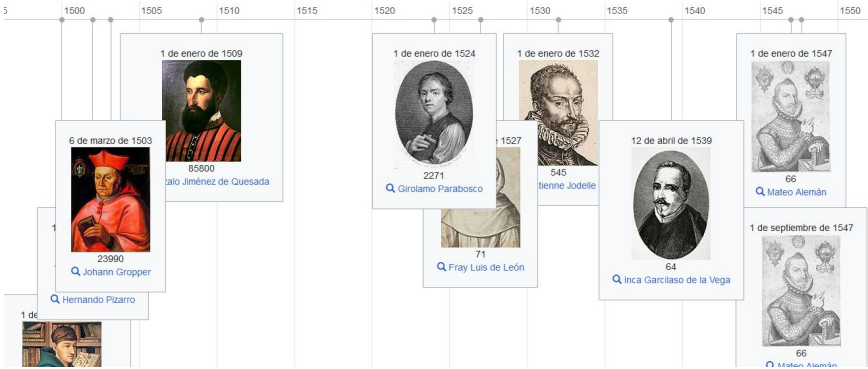


Visualización

Visualizaciones

- #defaultView:Timeline

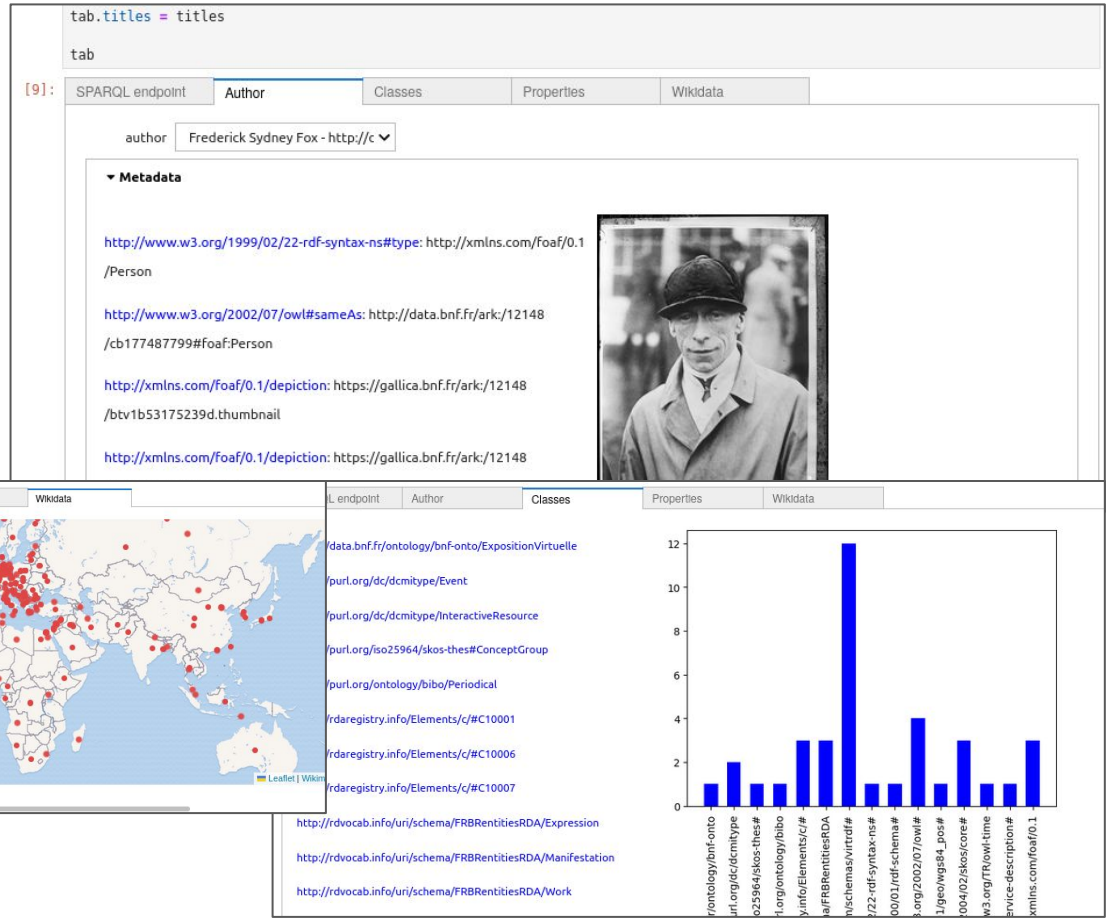
https://w.wiki/88$y




```
#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
```

Visualización

A **reproducible approach** to **browse** Linked Open Data in Cultural Heritage

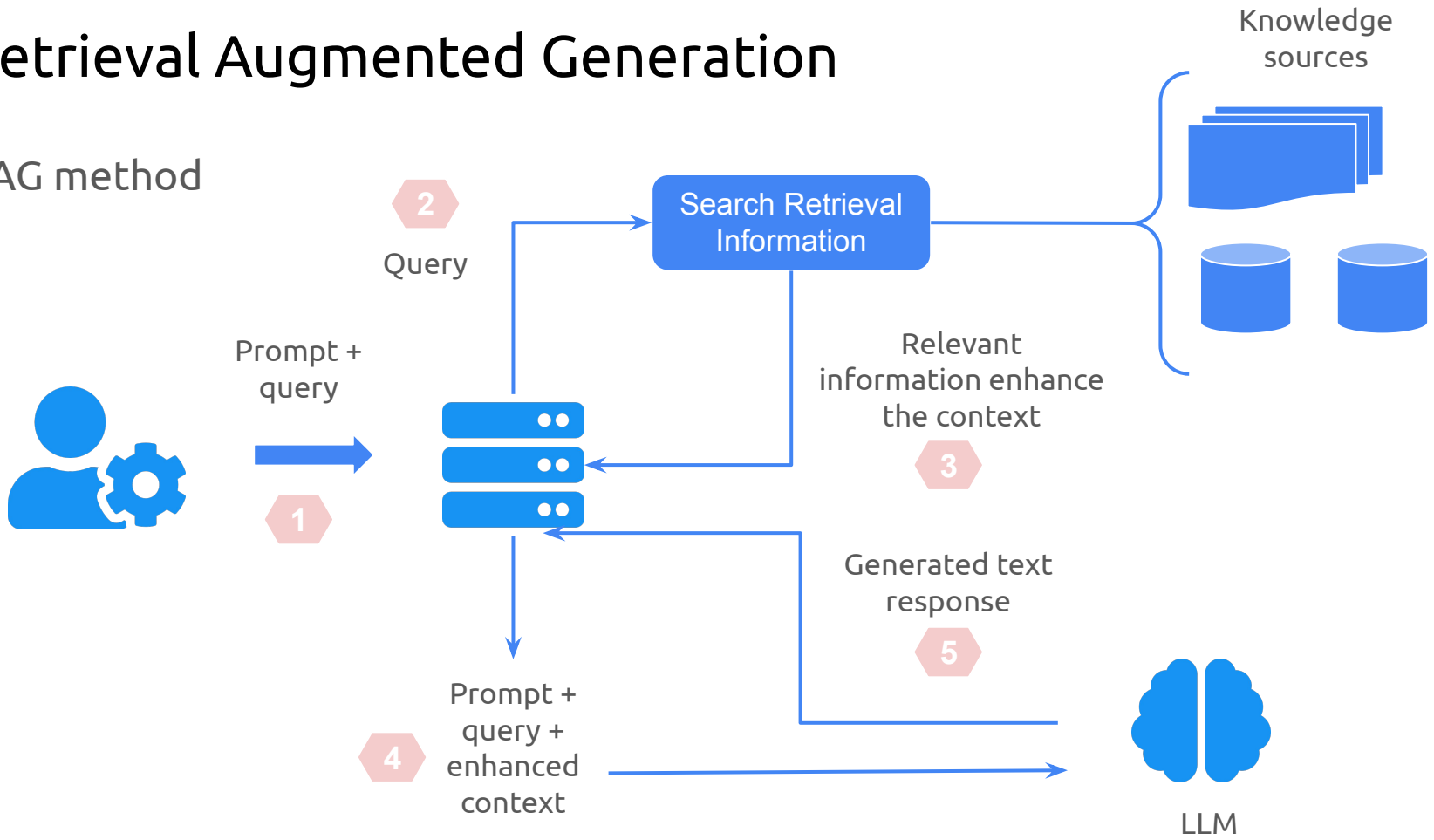


https://github.com/hibernator11/visual-ch-lod

https://doi.org/10.1145/3707647

Retrieval Augmented Generation

RAG method



Retrieval Augmented Generation