Visualización y reutilización de datos

Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Contenido

- * Ejemplos de transformación
- Introducción
- Reutilización
  - Visión por computador
  - Calidad de datos
- Visualización
  - Herramientas
  - Ejemplos

Transformación

Moving Image Archive

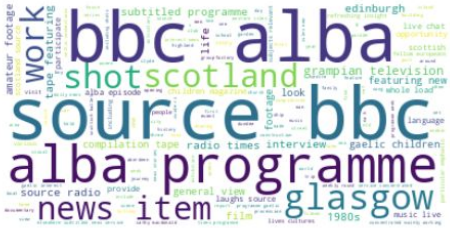
This dataset represents the descriptive metadata from the Moving Image Archive catalogue, which is Scotland's national collection of moving images.

- 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

Lee CSV y genera un fichero TTL con las tripletas



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/hibernator11/nls-fellowship-2022-23

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: t[illegible]mation 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.

Reutilización

Problema general…

¿Qué **datos** tenemos? ¿**Formato**? ¿Enriquecidos?

¿Es necesario realizar una **limpieza** de los datos?

¿Qué quiero **analizar**? ¿Cuál es el **objetivo**?

¿Quién es mi **audiencia**? ¿Estudiantes? ¿Investigadores?

¿Qué **infraestructura** tengo para desplegar mi código? ¿Hay **límites** de memoria o espacio?

¿Qué **licencia** tienen los datos?

https://www.glamlabs.io/

Visión por computador

Problemas comunes

Object detection is the task of detecting instances of objects of a specific class within an image or video.

Classification

Detection

Segmentation

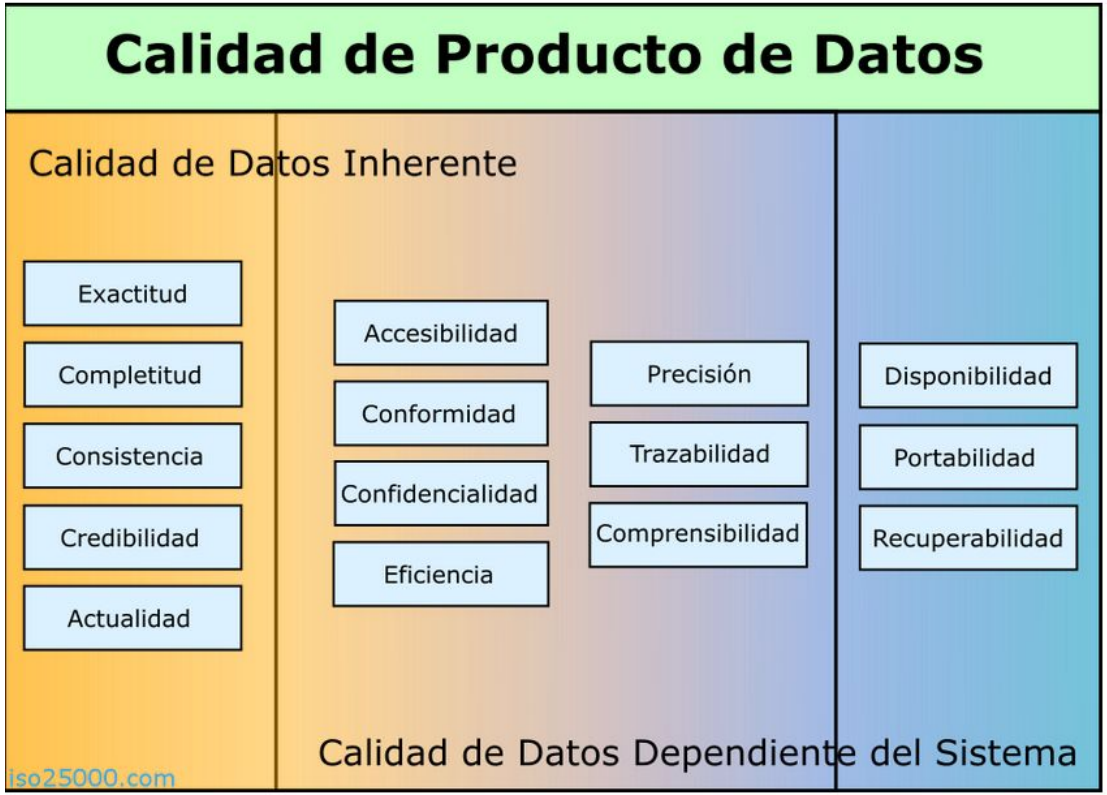
https://www.freecodecamp.org/news/how-to-detect-objects-in-images-using-yolov8/

Visión por computador

https://yolov8.com/

https://github.com/hibernator11/workshop-notebooks-dhnb2024

Calidad de datos



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

¿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

Calidad de datos

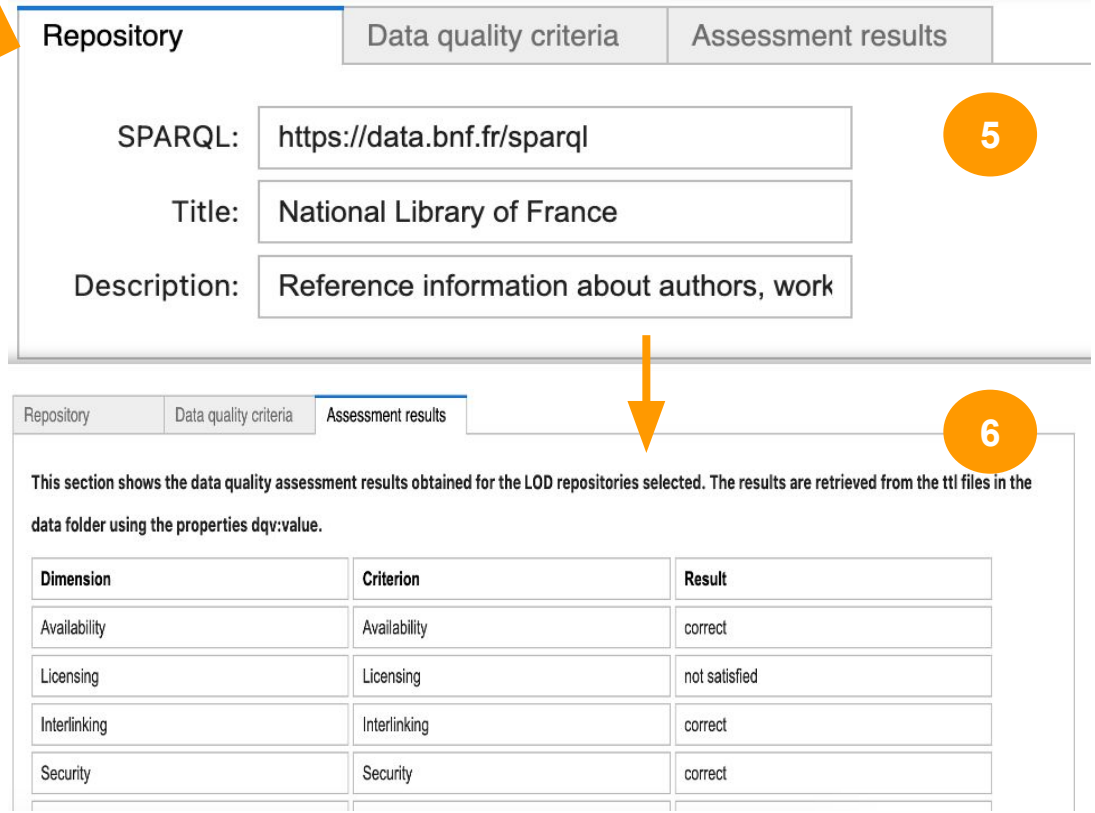
Calidad de datos

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

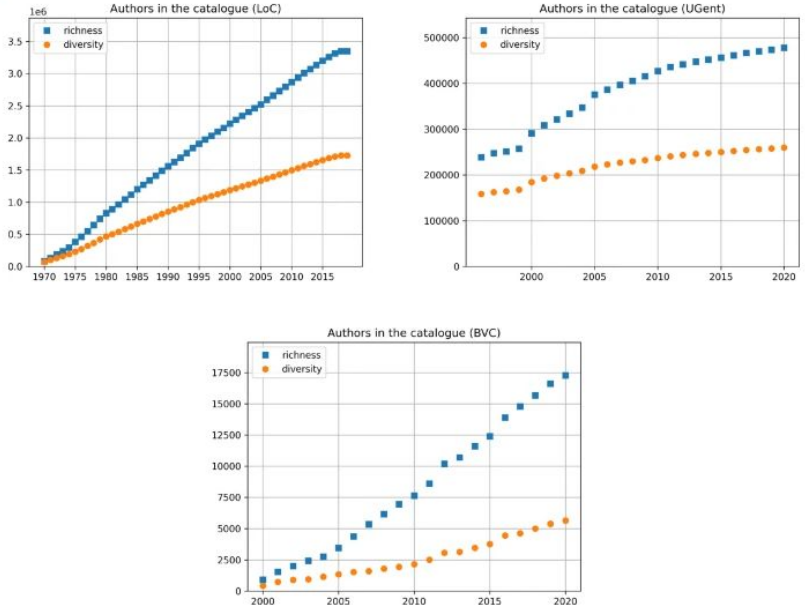


https://github.com/hibernator11/lod-quality-reproducible

Calidad de datos

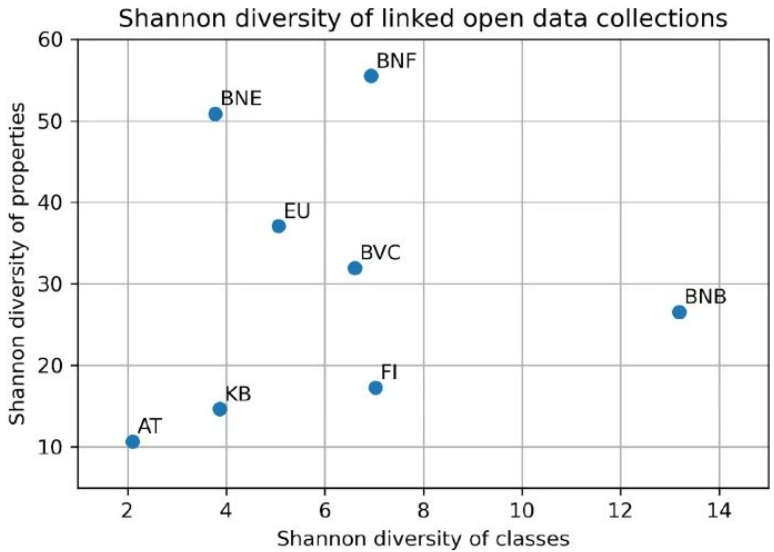
C. Carrasco, R., Candela, G. & Marco-Such, M. Measuring the diversity of data and metadata in digital libraries. *Int J Digit Libr* **26**, 5 (2025). https://doi.org/10.1007/s00799-025-00411-1

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

La generación de datos crece de forma exponencial

Los datos deben ser limpiados, filtrados, organizados, resumidos y analizados para facilitar la toma de decisiones

El enriquecimiento con repositorios externos puede facilitar la inclusión de información adicional

Visualización

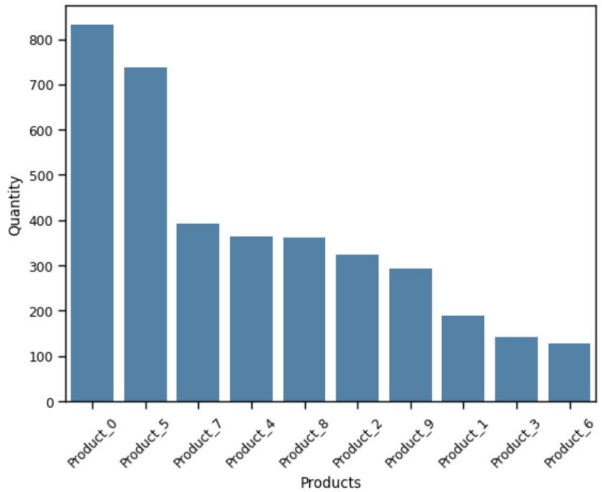
Las visualizaciones dependen de los objetivos perseguidos. Algunos de los más comunes son:

- Ranking
- Comparativas
- Correlaciones
- Distribución
- Evolución en el tiempo

Visualización

Ranking

- Trata de mostrar la relación de **orden** en un conjunto de datos

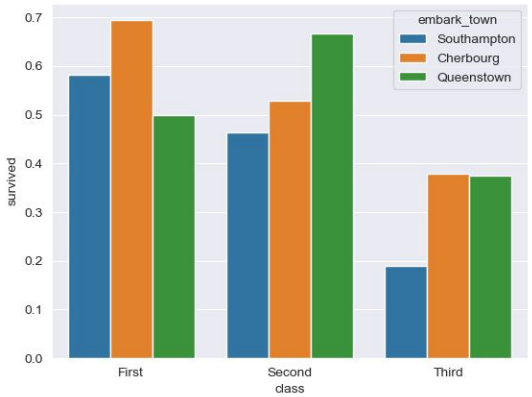


Bar Chart

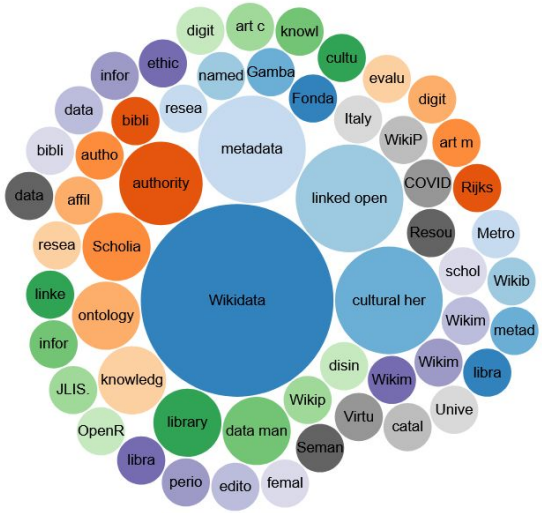
Visualización

Comparativa

- Enfrenta valores numéricos asociados a distintos elementos individuales o grupos



Grouped Bar Chart

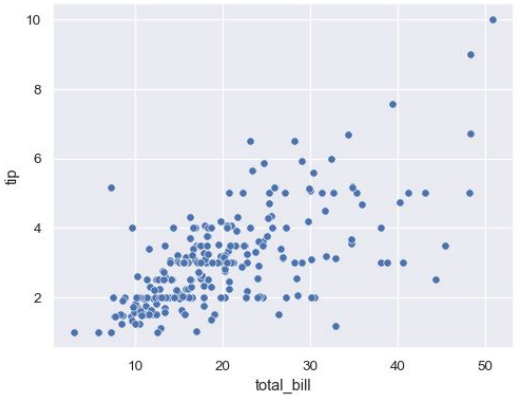


Bubble Chart

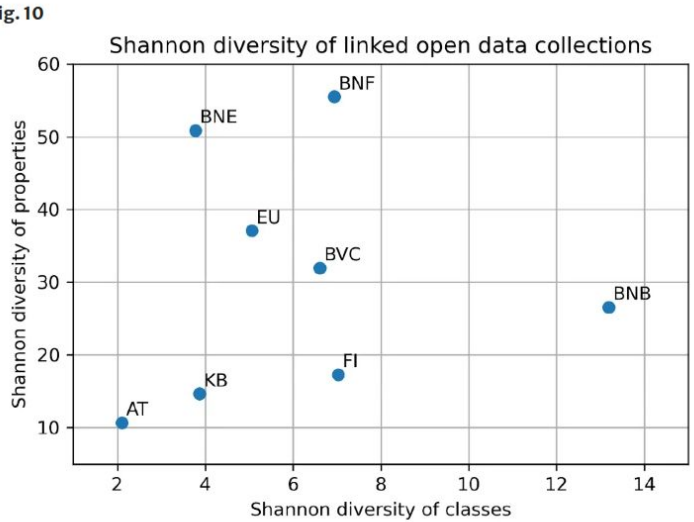
Visualización

Correlación

- Comprueba la relación entre pares de variables



Scatter plot



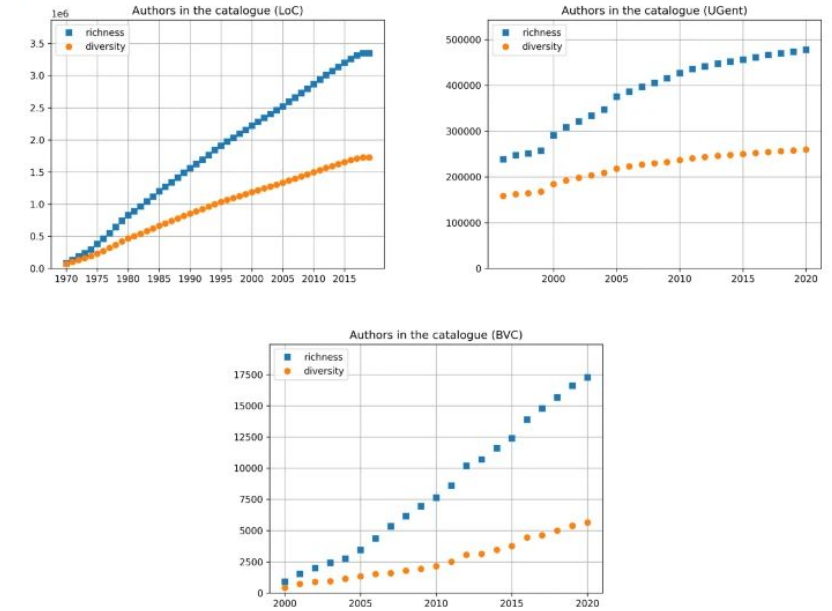
Shannon diversity of classes and properties in *Linked Open Data* published by libraries

Visualización

Evolución en el tiempo

- Reflejan el cambio de valor de una variable con respecto al tiempo

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

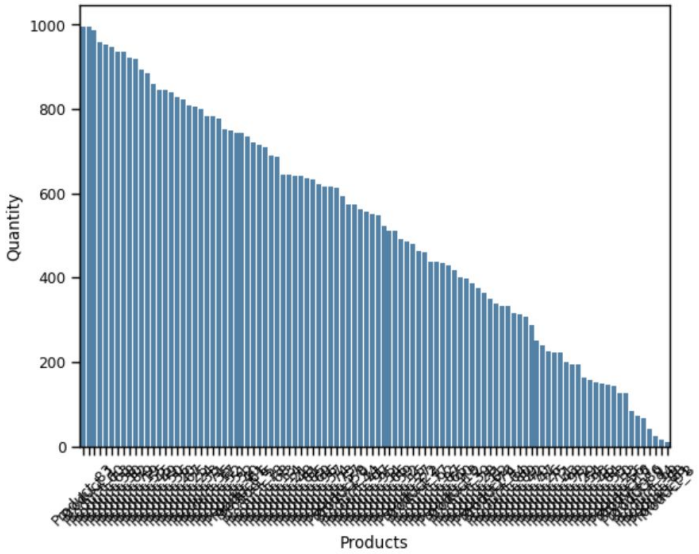
Visualización

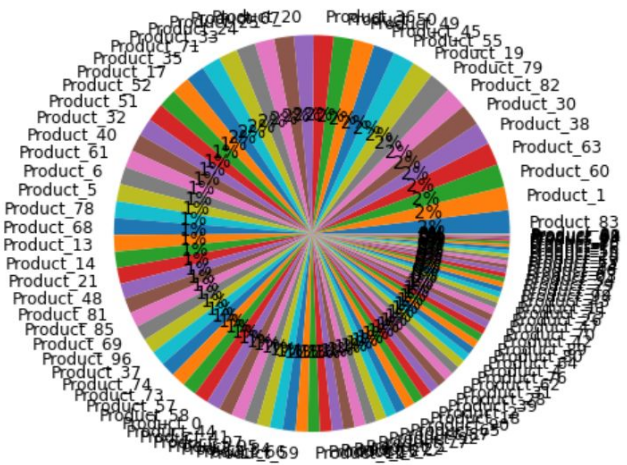
Selección de visualizaciones

- ¿Cómo seleccionar una visualización adecuada?
  - Finalidad:
    - ¿Análisis a lo largo del tiempo? ¿Distribución? ¿Comparación de valores?
  - Público objetivo:
    - ¿Experto en visualizaciones? ¿Experto en dominio?
  - Uso:
    - Exploratorio / Comunicativo
  - Características del dataset:
    - ¿Qué tipo de datos se van a utilizar? ¿Qué volumen de datos se está manejando? ¿Cuántos valores distintos hay? ¿Cuántas dimensiones entran en juego?

Visualización

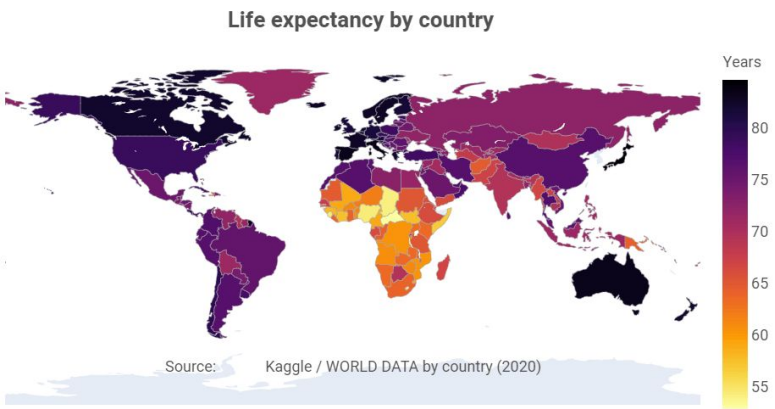
Problema **volumen** de datos

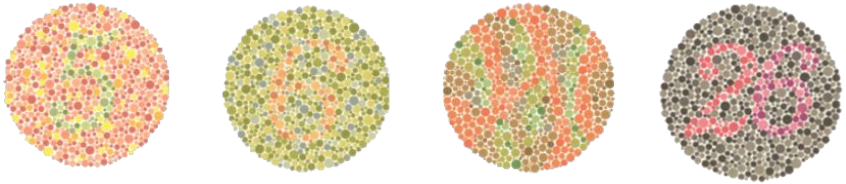




Visualización

Problema **uso del color, daltonismo, choques culturales**



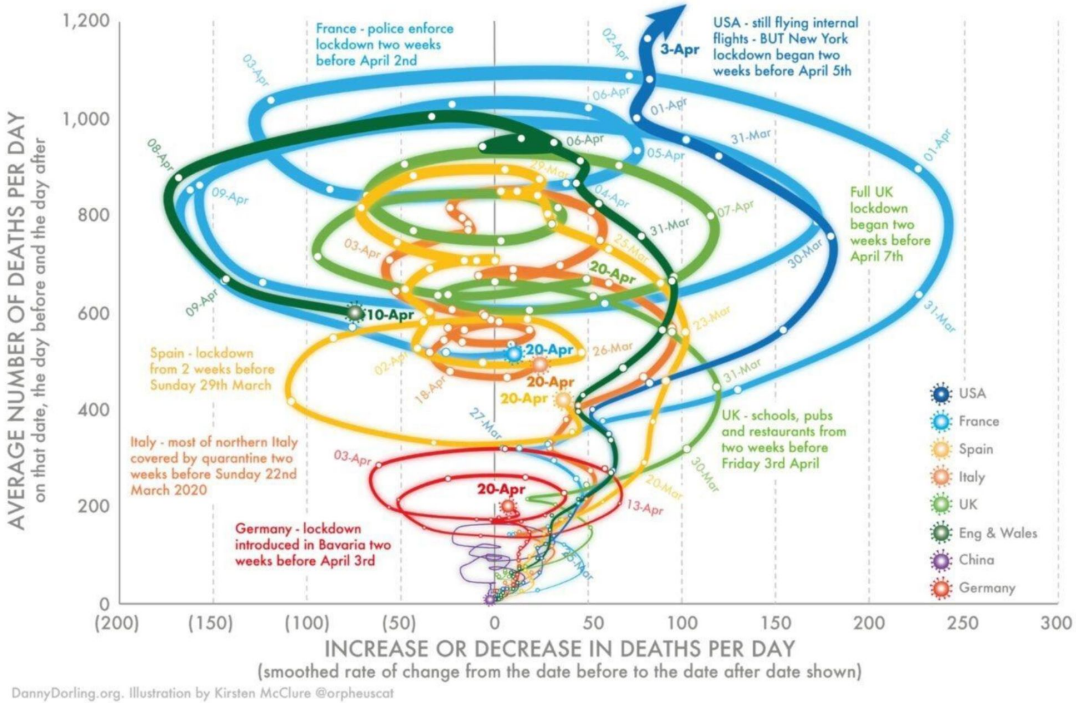


Fuente: Grupo Visioon. Test de daltonismo
https://visioon.es/blog/test-daltonismo-oftalmologia/

Fuente: Ayed, C. B., Halili, S., Tan, Y., & Grubb, A. M. (2023). Toward Internationalization and Accessibility of Color-based Goal Model Interpretation.

Visualización

Problemas **sobrecarga de información**



Visualización

Problemas generales

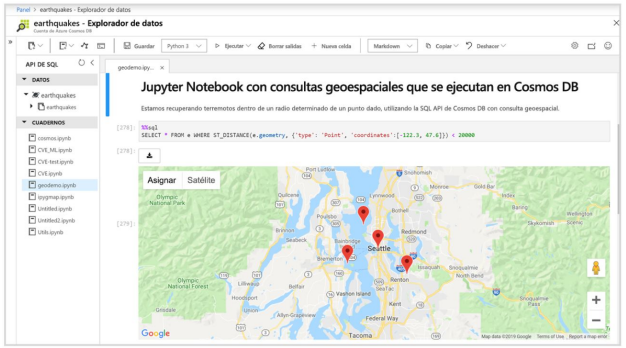
Via Reddit.
https://www.reddit.com/r/chile/comments/fpe6j5/megavisi%C3%B3n_being_very_megavisi%C3%B3n/

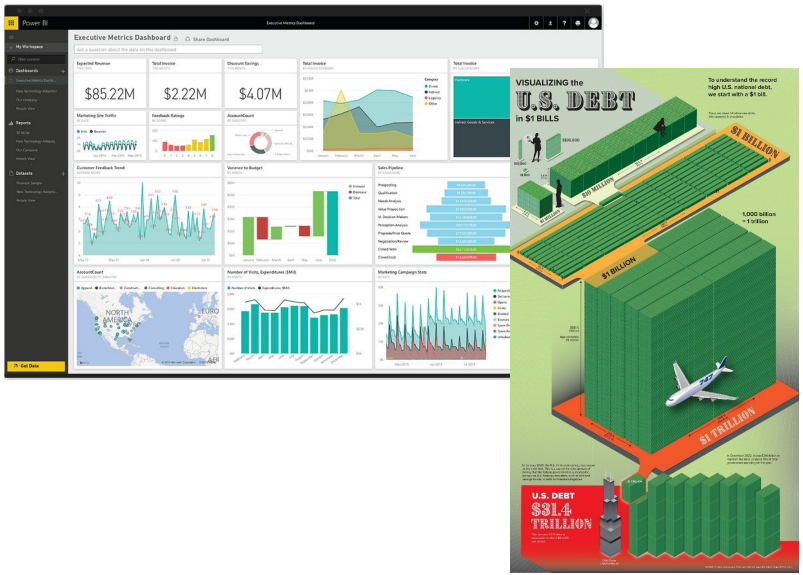
Visualización

Habitualmente es necesario utilizar múltiples visualizaciones para transmitir la información.

Podemos diferenciar las agrupaciones en:

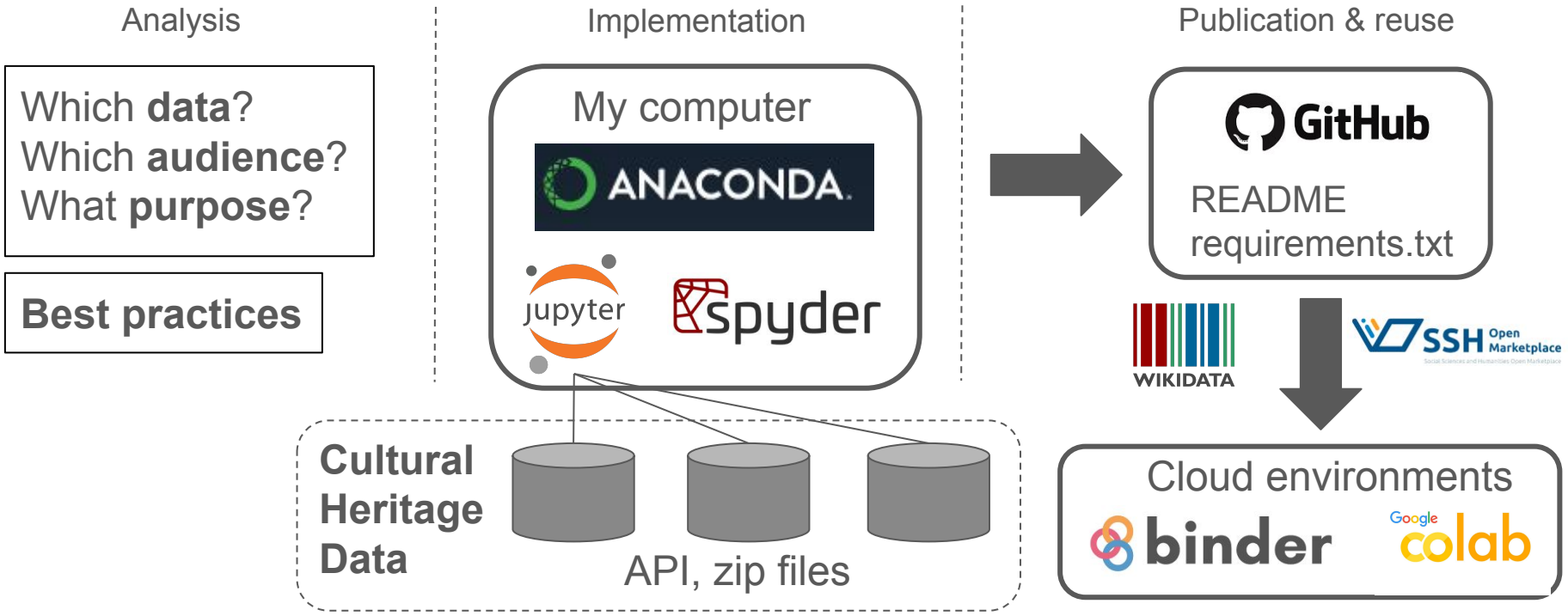
- **Notebooks**: incluyen narrativa y código
- **Cuadros de mando**: organizan las visualizaciones de forma analítica
- **Infografías**: comunican para concienciar sobre un tema particular





Visión por computador

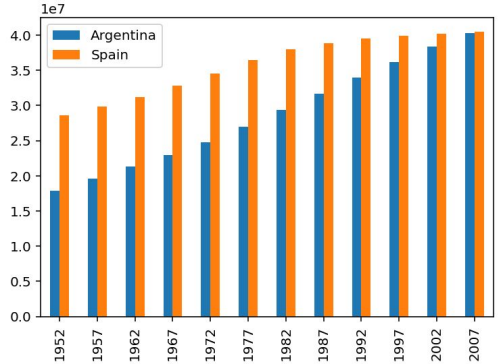
Herramientas - Jupyter Notebooks

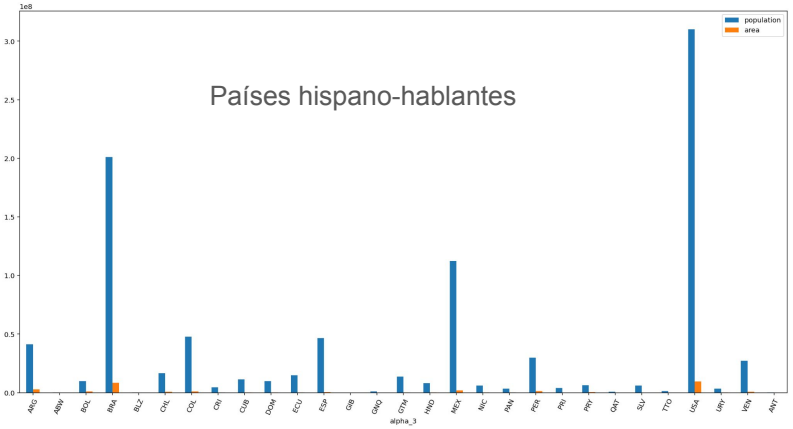


Visualización

Herramientas - Pandas

- Permite el análisis de datos tabulados (CSV, TSV)
- Importar, limpiar y gestionar datos
- Fusionar datos de forma eficiente
- Crear visualizaciones a partir de los datos





https://github.com/hibernator11/fi-ua

Visualización

Folium Herramientas

Python data, leaflet.js maps

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library. Manipulate your data in Python, then visualize it in a Leaflet map via Folium.

Concepts

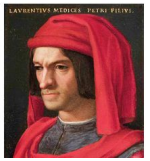
Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

The library has a number of built-in tilesets from OpenStreetMap, Mapbox, etc, and supports custom tilesets. Folium supports both Image, Video, GeoJSON and TopoJSON overlays and has a number of vector layers built-in.

Visualización

Herramientas - Wikidata

- Permite la creación de visualizaciones a partir de los datos de Wikidata
- Mapas, líneas de tiempo, cuadros de imágenes, etc.
- No es necesario instalar ningún software



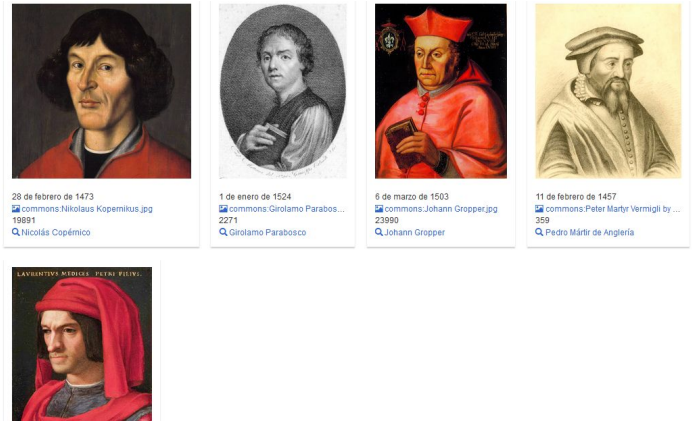
Visualización

Herramientas - Wikidata

- #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
```

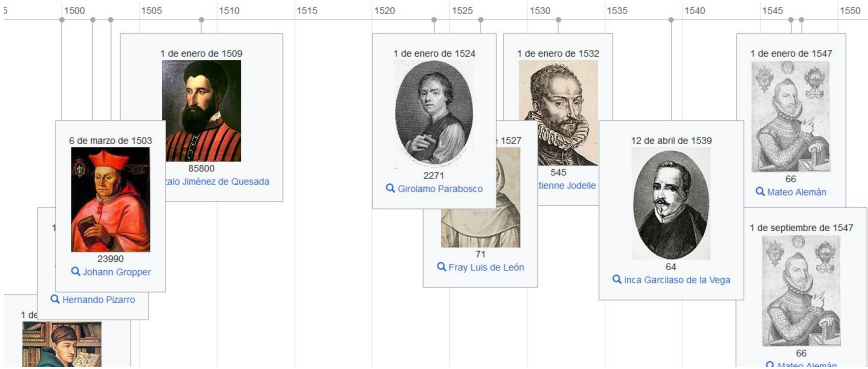


Visualización

Herramientas - Wikidata

- #defaultView:Timeline

https://w.wiki/88$y

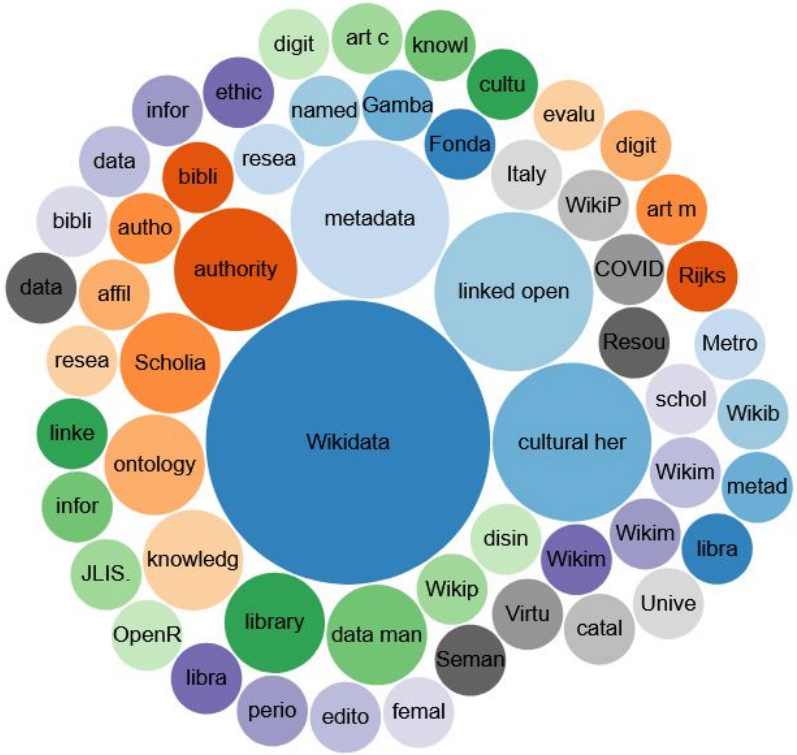



```
#defaultView:Timeline
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento .
  ?autor wdt:P18 ?imagen

  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
```


MUSEO DEL PRADO | LÍNEA DEL TIEMPO
Busca un artista, una obra, un evento...
Contexto histórico
Federico I Barbarroja
Jaime I de Aragón
Enrique III de Inglaterra
Eduardo III de Inglaterra
Geoffrey Chaucer
Enrique II de Inglaterra
1133 - 1189
Eduardo I de Inglaterra
1239 - 1307
Guerra de los Cien Años
1337 - 1453
Francisco I de Francia
1494 - 1
Felipe II de España
Enrique VIII de Inglaterra
Carlos I de Inglaterra
1600 - 1649
Carlos I de España
1500 - 1558
Revolución inglesa
1643 - 1651
Thomas Jefferson
George Washington
1732 - 179
Abraham Lincoln
Primera Guerra
Napoleón Bonaparte
1769 - 1821
Guerra de Secesión
1861 - 1865
Segunda Guerra...
1939 - 1945
Frontal con
Frontal de
Cacería de
Maestro de Lluça
Frontal de
Parte superior del...
Anónimo
Retablo de San Cristóbal
Anónimo
Fantasía moral
La Creación
El Descendimie
Weyden, Rogier
Tríptico del Jardín de las...
El Bosco
Carlos V en la batalla de...
Tiziano Vecellio di
El caballero de la mano en el...
El Greco
Las tres
Pedro Pablo
Bayeu y Subías,...
1734 - 1795
La maja
Goya y Lucientes,...
1746 - 1828
Las meninas
Velázquez, Diego R...
Juana I
MUSEO DEL PRADO
Piotr
Chicos en la playa
Sorolla y B
El 3 de mayo en Madrid o...
Goya y Lucientes, F...
Estud
cabez
Álvarez
https://www.museodelprado.es/coleccion/linea-del-tiempo
1100
1200
1300
1400
1500
1600
1700
1800
1900

Visualización

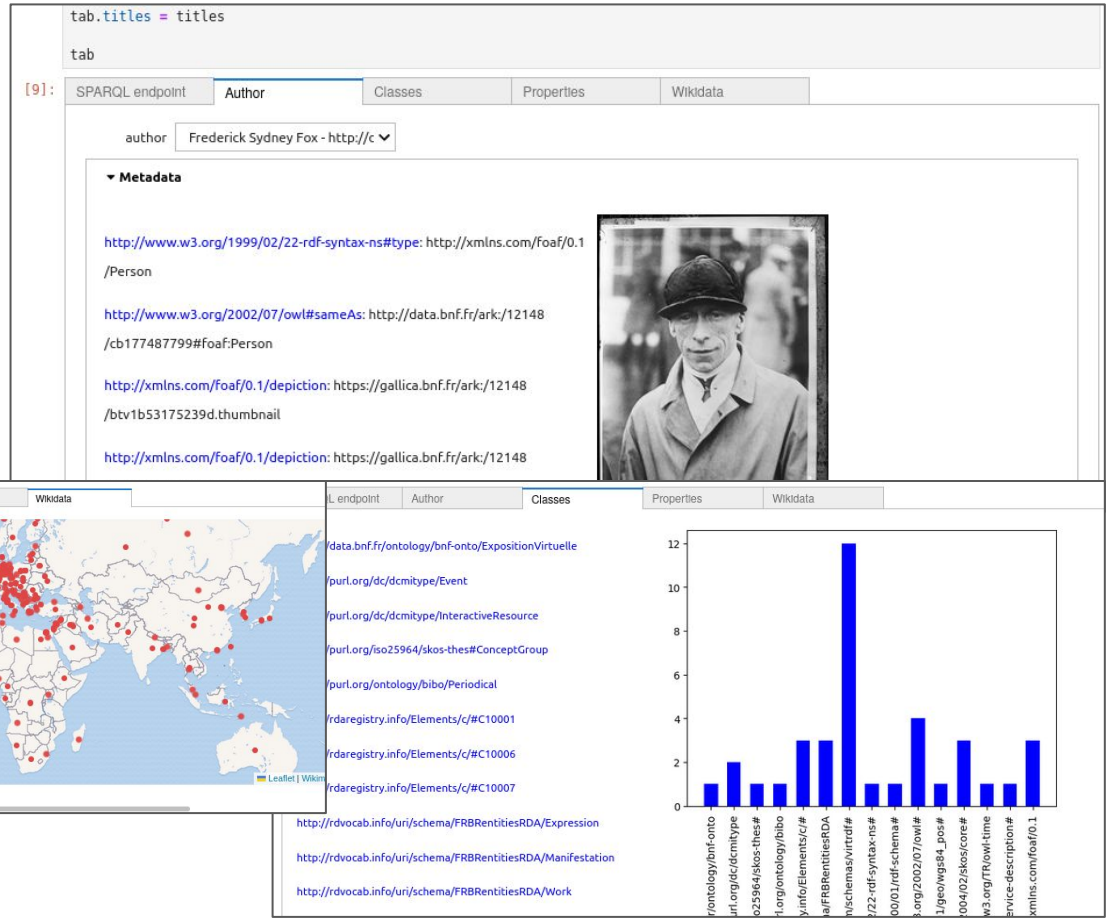
Herramientas - Wikidata



https://github.com/hibernator11/wikidata-review

Visualización

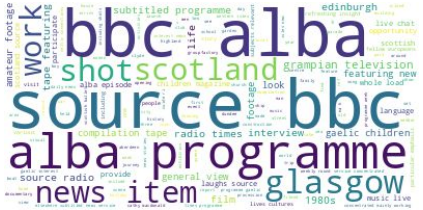
A **reproducible approach** to **browse** Linked Open Data in Cultural Heritage



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

https://doi.org/10.1145/3707647

Visualización



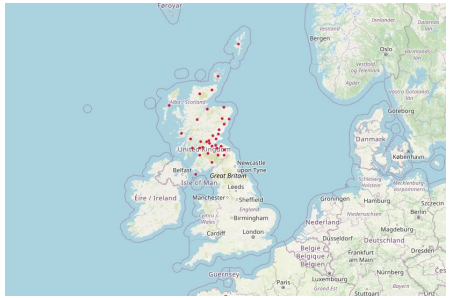
Moving Image Archive

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



jupyter

Folium

MARCXML MARC 21 XML Schema Official Web Site

Schema.org

GeoNames

WIKIDATA

WIKIDATA

Visualización

Proyecto presentado en la actividad "Compartir pantalla" con la Asociación de Humanidades Digitales Hispánicas.

launch binder

DOI 10.5281/zenodo.10123566

Introducción

Este proyecto forma parte de la actividad "Compartir pantalla" que introduce Wikidata en el ámbito bibliotecario como plataforma para facilitar el acceso, enriquecimiento y la visibilidad a través de la edición de recursos y el uso de SPARQL para recuperar información en forma de datos y recursos gráficos.

Además, se incluyen ejemplos de sentencias SPARQL proporcionados por Wikidata introduciendo los diferentes comandos e instrucciones, visualizaciones reales utilizadas en la web de Biblioteca Virtual Miguel de Cervantes y la International GLAM Labs Community, y otros ejemplos ilustrativos como por ejemplo sentencias federadas y creación de datasets.

https://github.com/hibernator11/hdh-compartir-pantalla-2023