Explorando *Wikidata* para la investigación en Ciencias Sociales

Exploring *Wikidata* for Social Science research

Wenceslao Arroyo-Machado

Cómo citar este artículo:

Arroyo-Machado, Wenceslao (2024). "Explorando *Wikidata* para la investigación en Ciencias Sociales [Exploring *Wikidata* for Social Science research]". *Infonomy*, 2(2) e24034.
https://doi.org/10.3145/infonomy.24.034

Wenceslao Arroyo-Machado
https://orcid.org/0000-0001-9437-8757
https://directorioexit.info/ficha4831
Arizona State University, *School of Public Affairs*
Center for Science, Technology and Environmental Policy Studies
Phoenix, AZ 85004, USA
warroyom@asu.edu

Resumen

Este texto subraya la relevancia de recursos abiertos como *Wikidata* para las Ciencias Sociales, proporcionando acceso a información que facilita la investigación en áreas sociales, económicas y culturales. Además de describir la estructura y funcionamiento de *Wikidata*, que utiliza un sistema de grafos abierto y multilingüe para modelar relaciones complejas y hacer la información accesible, este artículo ofrece indicaciones prácticas para empezar a utilizar y acceder a los datos. A través de estudios de caso, se discute sobre el potencial de *Wikidata* en diversos ámbitos de la investigación. Aunque *Wikidata* ofrece grandes posibilidades para el análisis flexible y detallado, enfrenta desafíos de integridad y sesgo de datos. Es por ello crucial abordar su uso con un enfoque crítico y contribuir activamente a la mejora de la plataforma.

Palabras clave

Wikidata; Ciencias Sociales; Datos abiertos; Acceso a datos; Estudios de caso; Integridad de datos; Estructura de grafos; Contribución colaborativa; Análisis detallado; Sesgo de datos; SPARQL; Ejemplos.

Abstract

This text underscores the importance of open resources like *Wikidata* for Social Sciences, providing access to data that facilitates research in social, economic,

and cultural areas. In addition to describing the structure and operation of *Wikidata*, which uses an open, multilingual graph system to model complex relationships and make information accessible, the document also provides practical guidance on how to start using and accessing the data. Through case studies, it illuminates the potential of *Wikidata* in various research areas. While *Wikidata* offers vast possibilities for flexible and detailed analysis, it faces challenges of data integrity and bias. It is crucial to approach its use with a critical mindset and actively contribute to enhancing the platform.

Keywords
Wikidata; Social Sciences; Open data; Data access; Case studies; Data integrity; Graph structure; Collaborative contribution; Detailed analysis; Data bias; SPARQL; Examples.

1. Introducción

A menudo suelo destacar el impacto que han supuesto los medios sociales para campos como la cienciometría, pero lo cierto es que la Web ofrece en sí misma un recurso de inestimable valor para las Ciencias Sociales (**Askitas**; **Zimmermann**, 2015). Dentro de este contexto, las bases de datos de acceso libre se han convertido en recursos indispensables, no solo por ofrecer una encomiable cantidad de datos sin barreras, sino por promover la cultura abierta. La accesibilidad a estas bases permite a investigadores de todo el mundo explorar y analizar temas sociales, económicos y culturales de manera detallada, facilitando así una comprensión más amplia y profunda de los fenómenos sociales. Es posible citar algunas de ellas, como *OpenAlex* para el acceso a datos bibliométricos (**Priem** *et al.*, 2022) u *OpenStreetMap* para datos geográficos (**Haklay**; **Weber**, 2008). Fuentes que no solo favorecen el desarrollo de la investigación individual, sino que también contribuyen al avance colectivo del conocimiento en las Ciencias Sociales.

A diferencia de las bases de datos relacionales tradicionales, *Wikidata* está diseñada específicamente para la manipulación y gestión de conocimiento abierto. Esto se logra mediante el almacenamiento de información interconectada de manera semántica a través de una estructura de grafos

En esta misma línea, hay una cuestión que suelo abordar a menudo y que es la diferencia entre *Wikidata* y *Wikipedia*, dos plataformas que si bien tienen muchos puntos en común, son radicalmente diferentes. Sin embargo, esta confusión, que más adelante abordaremos, tiene una importante repercusión, que es la obviación de una fuente de información para la investigación de gran relevancia. Ya sea porque se le asocia un valor o utilidad completamente distinto al suyo o porque directamente se desconoce, lo cierto es que este recurso sigue pasando muy desapercibido en un contexto en el que debería tener reservado un espacio destacado. No obstante, es reconfortante ver que una parte considerable de la comunidad investigadora

está aprovechando su potencial para aplicaciones diversas, como plataformas de análisis bibliométricos (**Nielsen** *et al.*, 2017) o algoritmos de identificación de género (**González-Salmón**; **Robinson-García**, 2023).

1.1. Los *open knowledge graphs*

Es hora de hablar sobre *Wikidata*, una base de datos documental y colaborativa de interés multidisciplinar. Antes de profundizar en sus aplicaciones, es crucial entender su estructura fundamental: un *open knowledge graph* multilingüe y colaborativo. A diferencia de las bases de datos relacionales tradicionales, *Wikidata* está diseñada específicamente para la manipulación y gestión de conocimiento abierto. Esto se logra mediante el almacenamiento de información interconectada de manera semántica a través de una estructura de grafos. Esta estructura no solo facilita la modelación de relaciones complejas entre datos, sino que también asegura que la información sea accesible y utilizable de manera eficiente tanto por máquinas como por humanos. El núcleo o la base de este sistema reside así en los grafos, compuestos por dos elementos fundamentales:

- **Entidades**: Cada entidad representa un elemento del mundo real, como personas, lugares, objetos o conceptos.
- **Relaciones**: Estas son conexiones semánticas que vinculan dos entidades, estableciendo un tipo de relación entre ellas.

A partir de ello se pueden construir frase o tripletas, que es la unión básica de dos entidades por medio de una relación. Por ejemplo, la película *Creatura* es dirigida por Elena Martín Gimeno (Figura 1).

Figura 1. Ejemplo de tripleta generada entre una película y su directora

1.2. *Wikidata* no es *Wikipedia*

Como ya remarcaba antes, *Wikidata* es completamente diferente de *Wikipedia*. Mientras que esta última se enfoca en la creación de artículos enciclopédicos, *Wikidata* está diseñada para recopilar y compartir datos estructurados sobre una variedad de entidades, independientemente de si están presentes o no en *Wikipedia*. Esto permite una gran flexibilidad en términos de qué información se puede incluir y cómo puede ser reutilizada. Además, mientras que en *Wikipedia* los usuarios contribuyen principalmente escribiendo y editando textos, en *Wikidata* añaden y actualizan datos en un formato estructurado. Esto involucra la creación de enlaces entre datos que reflejan relaciones reales en el mundo, como conexiones entre personas, lugares, eventos y conceptos. Como resultado, *Wikidata* se convierte en un instrumento con un considerable potencial en la Ciencia y en las Ciencias Sociales en particular. Los principales elementos a considerar en *Wikidata* aparecen recogidos en la Tabla 1.

Wikidata es completamente diferente de *Wikipedia*. Mientras que esta última se enfoca en la creación de artículos enciclopédicos, *Wikidata* está diseñada para recopilar y compartir datos estructurados sobre una variedad de entidades, independientemente de si están presentes o no en *Wikipedia*

Tabla 1. Principales elementos disponibles en *Wikidata*

Elemento	Descripción
Ítems *Entidades*	Son las entidades principales en *Wikidata*. Cada ítem posee una etiqueta que es el nombre común, una descripción breve, y un identificador único que comienza con una "Q" seguida de un número. Por ejemplo: *Creatura* (Q113729314) y Elena Martín (Q50842870)
Propiedades *Relación*	Utilizadas para describir datos específicos de los ítems y para establecer relaciones entre ellos. Cada propiedad también tiene un identificador único que comienza con una "P" seguida de un número. Por ejemplo: director (P57)
Declaraciones *Tripleta*	Son construcciones de datos que combinan ítems y propiedades para formular afirmaciones detalladas sobre las entidades. Por ejemplo: Q113729314 → P57 → Q50842870

1.3. Limitaciones de *Wikidata*

No obstante, no todo es positivo. Al depender *Wikidata* de contribuciones colaborativas, enfrenta desafíos significativos relacionados con la integridad y el sesgo de los datos. El hecho de que cualquier usuario pueda editar y añadir información democratiza el proceso de contribución, pero también aumenta la posibilidad de errores. Además, la variabilidad en la frecuencia y profundidad

de las contribuciones puede resultar en un conocimiento sesgado, con ciertas áreas geográficas, culturales o temáticas recibiendo menos atención y desarrollo en comparación con otras. Un aspecto que ya ha sido vislumbrado en la *Wikipedia*, que opera bajo la misma filosofía colaborativa (**Tripodi**, 2021). Este desequilibrio en la representación de datos puede afectar la fiabilidad de los análisis basados en *Wikidata*. Aunque existen mecanismos de verificación y corrección, la dimensión de la base de datos complica la tarea de mantener una precisión universal. Por lo tanto, es crucial abordar el uso de *Wikidata* con un enfoque crítico, verificando la información con fuentes adicionales y manteniendo una conciencia clara de estas limitaciones para garantizar su utilidad como recurso de conocimiento abierto y confiable.

2. Acceso a los datos de *Wikidata*

Una vez comprendida la naturaleza de *Wikidata*, toca entender su funcionamiento y distintas vías de acceso. Los datos de *Wikidata* están disponibles para ser consultados de múltiples maneras, lo que refleja su diseño abierto y flexible.

Tabla 2. Principales formas de acceso a datos de *Wikidata*

Método	Acceso	Descripción
Wikidata Search *https://www.wikidata.org*	Navegación manual — Dataset pequeño	Este es el buscador básico de *Wikidata*. Funciona como un punto de entrada inicial para usuarios que desean explorar y contribuir a la base de datos. Es una herramienta intuitiva que permite realizar búsquedas sencillas y también ofrece opciones para la colaboración directa en la edición y mejora de los ítems.
Wikidata Query Service (WDQS) *https://query.wikidata.org*	Recuperación automatizada — Dataset mediano	Esta es la aplicación principal para la consulta masiva de datos en *Wikidata*. Utiliza el lenguaje de consulta SPARQL, que es específico para bases de datos basadas en grafos.
Wikibase REST API *https://doc.wikimedia.org/Wikibase/master/js/rest-api*	Recuperación automatizada — Dataset mediano	La API REST de *Wikibase* permite la manipulación automatizada de los datos de *Wikidata*. Ofrece *endpoints* para la creación, consulta, actualización y eliminación de ítems en *Wikidata*, facilitando la integración con aplicaciones externas.
Volcados de base de datos *https://www.wikidata.org/wiki/Wikidata:Database_download/es*	Recuperación automatizada — Dataset grande	Los volcados de la base de datos de *Wikidata* ofrecen una copia completa de todos los datos en varios formatos, como JSON. Estos volcados son ideales para análisis de grandes volúmenes de datos.

En la Tabla 2 se muestran las principales, que no únicas, vías de acceso. Pues, dado que *Wikidata* es completamente abierta, tanto en lo referente a sus datos como a su estructura, existe la posibilidad de desarrollar nuevos servicios y plataformas que se apoyen en esta base de datos. Ejemplos de tales iniciativas incluyen *Reasonator* y *Scholia*:
https://reasonator.toolforge.org
https://scholia.toolforge.org
Reasonator proporciona una interfaz más amigable y visual para explorar datos de *Wikidata*, mientras que *Scholia* se especializa en visualizar información científica y académica, aprovechando los datos de *Wikidata* para generar perfiles de investigadores, publicaciones y mucho más.

3. Estrategias de consulta

Al abordar la tarea de explorar y extraer datos de *Wikidata*, el lenguaje de consulta SPARQL es el instrumento clave, ofreciendo un acceso profundo y flexible a la vasta base de datos. Familiarizarse con SPARQL es esencial y, afortunadamente, existen numerosos recursos que pueden facilitar este aprendizaje.

He aquí algunas recomendaciones prácticas para comenzar a trabajar con SPARQL en *Wikidata*:

1. Aprende con *Wikidata*: La propia *Wikidata* ofrece una gran cantidad de recursos para aprender a trabajar con sus datos en SPARQL[1], revisa sus recursos introductorios para comprender los aspectos básicos antes de realizar consultas.

2. Utiliza el asistente de búsqueda: *Wikidata* ofrece un asistente de búsqueda integrado que puede ayudarte a formular tus consultas SPARQL. Este asistente es especialmente útil para principiantes, ya que proporciona una interfaz gráfica para construir consultas sin necesidad de escribir el código directamente.

3. Aprende de ejemplos: Una de las mejores maneras de aprender SPARQL es observando y modificando ejemplos existentes. *Wikidata* y otras comunidades en línea tienen numerosos ejemplos que puedes estudiar y adaptar a tus necesidades. Experimentar con ejemplos te permite entender cómo se estructuran las consultas y cómo interactúan con la base de datos.

4. Usa la IA pero con conocimiento: Aprovecha las herramientas de inteligencia artificialque pueden sugerir formas de optimizar tus consultas o explorar los datos de manera más efectiva. Sin embargo, es crucial entender los fundamentos detrás de las consultas para poder ajustarlas adecuadamente y evitar depender completamente de la automatización o errar en puntos críticos.

[1] *https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help/es*

5. Comienza con consultas simples y divide la consulta en varias partes: Si estás comenzando SPARQL, empieza con consultas sencillas y aumenta su complejidad gradualmente. Además, dividir una consulta compleja en subconsultas más manejables puede hacer el proceso más comprensible y menos propenso a errores.

6. Recupera solo campos imprescindibles: Para mejorar el rendimiento y la claridad de tus consultas, limita los resultados a aquellos campos que realmente necesitas. Esto no solo acelera las consultas sino que también hace que los resultados sean más fáciles de analizar.

7. Revisa siempre los resultados: Aunque a primera vista parezca que los resultados son correctos, revisa cuidadosamente que no se haya introducido ruido o que falten datos.

4. Casos prácticos

A continuación incluyo a modo de ejemplo varias consultas realizadas en SPARQL con las que recuperar datos que podrían ser de utilidad para distintas investigaciones dentro de las Ciencias Sociales.

4.1. Estudios bibliométricos

Este caso se centra en la elaboración de una lista detallada de científicos, incluyendo sus nombres y las universidades donde han estudiado o trabajado, con los que poder analizar, por ejemplo, las redes académicas y profesionales (Tabla 3).

Tabla 3. Consulta en SPARQL para recuperar científicos y sus universidades

```
SELECT DISTINCT ?cientifico ?nombreCientifico ?universidad ?nombreUniversidad
WHERE {
  ?cientifico wdt:P31 wd:Q5;
              wdt:P106 wd:Q901;
              wdt:P108|wdt:P69 ?universidad.
  ?cientifico rdfs:label ?nombreCientifico.
  ?universidad rdfs:label ?nombreUniversidad.
  FILTER(LANG(?nombreCientifico) = "es").
  FILTER(LANG(?nombreUniversidad) = "es").
} LIMIT 100
```

4.2. Estudios de género

En este caso se recopilan datos demográficos sobre individuos nacidos en España en los últimos 20 años, identificando el nombre y el género de cada individuo (Tabla 4). Con ello se hace posible identificar para un país concreto los géneros asociados a cada nombre en un contexto reciente.

Tabla 4. Consulta en SPARQL para recuperar el nombre de personas nacidas en España y su género

```
SELECT ?person ?givenName ?genderLabel WHERE {
  ?person wdt:P31 wd:Q5;
        wdt:P21 ?gender;
        wdt:P27 wd:Q29;
        wdt:P569 ?birthdate;
        wdt:P735 ?givenNameItem.
  ?givenNameItem rdfs:label ?givenName.
  ?gender rdfs:label ?genderLabel.
  FILTER(YEAR(?birthdate) >= 2000).
  FILTER(LANG(?givenName) = "es").
  FILTER(LANG(?genderLabel) = "es").
} LIMIT 100
```

4.3. Análisis de trayectorias

En esta ocasión se recuperan las trayectorias profesionales de individuos que estudiaron en la Universidad de Granada, extrayendo datos específicos sobre sus campos de ocupación post-graduación (Tabla 5).

Tabla 5. Consulta en SPARQL para recuperar el nombre de de individuos que estudiaron en la Universidad de Granada, con su ocupación

```
SELECT ?alumnus ?alumnusName ?occupation ?occupationName WHERE {
  ?alumnus wdt:P69 wd:Q1232180;
        wdt:P106 ?occupation.
  ?alumnus rdfs:label ?alumnusName.
  ?occupation rdfs:label ?occupationName.
  FILTER(LANG(?alumnusName) = "es").
  FILTER(LANG(?occupationName) = "es").
} LIMIT 100
```

4.4. Análisis de redes sociales y de influencia

Por último, se identifican artistas que han sido influenciados por David Bowie (Tabla 6). Al explorar las conexiones entre Bowie y otros artistas a través de sus influencias declaradas, se puede analizar su legado y permeabilidad en generaciones y géneros artísticos.

Tabla 6. Consulta en SPARQL para recuperar artistas influenciados por David Bowie

```
SELECT ?artist ?artistLabel WHERE {
  ?artist wdt:P737 wd:Q5383;
  rdfs:label ?artistLabel.
  FILTER(LANG(?artistLabel) = "en").
}
```

5. Contribuciones a *Wikidata*

No olvides que es fundamental no solo usar los datos sino contribuir. Son muchos los puntos en los que puedes ayudar, por ejemplo traduciendo. Además de la traducción, puedes mejorar la calidad de los datos existentes corrigiendo errores, actualizando información obsoleta o añadiendo nuevas referencias fiables. Crear nuevos elementos o ítems también es muy importante, especialmente para ampliar la cobertura de temas y regiones menos representadas. Contribuir con imágenes y otros archivos multimedia en *Wikimedia Commons* y vincularlos a *Wikidata* enriquece visualmente la base de datos. Son muchas las formas de contribuir y las oportunidades que vas a tener de hacerlo, por lo que no desaproveches ninguna.

6. Conclusión

A modo de cierre, solo queda subrayar una vez más el elevado potencial de *Wikidata*, que puede ser de utilidad no solo como fuente de datos para la investigación sino como recurso complementario. Siempre, eso sí, revisando la consistencia de los datos y verificando que no existan sesgos o problemas en los mismos que puedan afectar a los resultados. Finalmente, te recomiendo consultar el vídeo complementario subido a *YouTube* (**Arroyo-Machado**, 2024): *https://www.youtube.com/watch?v=ShIdh68BNs8*

7. Referencias

Arroyo-Machado, Wenceslao (director) (2024). *Explorando Wikidata para la investigación en Ciencias Sociales, con Wenceslao Arroyo-Machado*, marzo 25. *https://www.youtube.com/watch?v=ShIdh68BNs8*

Askitas, Nikolaos; **Zimmermann, Klaus F.** (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, *36*(1), 2-12. *https://doi.org/10.1108/IJM-02-2015-0029*

González-Salmón, Elvira; **Robinson-García, Nicolás** (2023). WikiGenDex: Un nuevo algoritmo de identificación de género basado en fuentes abiertas. *Infonomy*, *2*(1). *https://doi.org/10.3145/infonomy.24.010*

Haklay, Mordechay; **Weber, Patrick** (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12-18. *https://doi.org/10.1109/MPRV.2008.80*

Nielsen, Finn-Årup; **Mietchen, Daniel**; **Willighagen, Egon** (2017). Scholia, Scientometrics and Wikidata. In: E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna*;* O. Hartig (eds.). *The Semantic Web: ESWC 2017 Satellite Events* (pp. 237-259). Springer International Publishing.

Priem, Jason; **Piwowar, Heather**; **Orr, Richard** (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. *https://doi.org/10.48550/ARXIV.2205.01833*

Tripodi, Francesca (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 14614448211023772. *https://doi.org/10.1177/14614448211023772*