

Gestión de datos de investigación sobre violencia de género y su integración a repositorios, bajo principios de ciencia abierta

Trabajo de Fin de Máster Máster Universitario en Ingeniería Geomática y Geoinformación



Autor:

Martínez Montes Carlos

Tutor:

Anquela Julián, Ana Belén

Cotutor/a externo:

Hernández Zetina, Sandra

Introducción

La finalidad del proyecto es crear y publicar un conjunto de datos abiertos sobre violencia de género, organizado de acuerdo con los principios FAIR y las directrices de la Carta Internacional de Datos Abiertos. Utiliza la información recogida durante la investigación del grupo de tecnologías geoespaciales de la Universidad Politécnica de Valencia, en el marco de las subvenciones AICO2022 (expediente CIAICO/2021/292), subvencionado por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana. Para lograr este objetivo, se organiza y estructura los datos eficientemente antes de almacenarlos en un repositorio de datos abiertos, facilitando su reutilización y análisis.

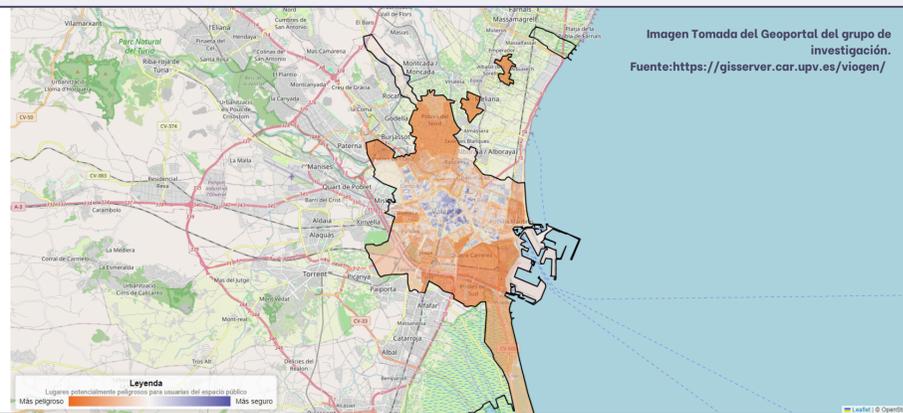


Imagen Tomada del Geoport del grupo de investigación.
Fuente: <https://gisserver.car.upv.es/viogen/>

Objetivo

El objetivo principal de este proyecto es organizar y estructurar los datos de manera eficiente antes de su almacenamiento definitivo en un repositorio de datos abiertos. Esto permitirá que los datos sean una pieza central y una referencia fundamental para la investigación para la que fueron creados, siguiendo los principios FAIR.

Metodología

En el desarrollo de este proyecto, se siguieron una serie de pasos metódicos para seleccionar, clasificar, y documentar los datos de manera que sean accesibles y utilizables para diversos propósitos, incluyendo la publicación en revistas científicas. A continuación, se muestra el organigrama seguido:



Resultados

Los resultados de este proyecto se pueden dividir en estos puntos:

- Recolección y Organización de Datos
- Desarrollo de Software
- Publicación en un repositorio de datos
- Facilitar la Investigación y la Colaboración

Tablas generadas

Datos de DATA_ES_VLC.csv

NAME	DESCRIPTION	DATA TYPE
VLC-ID	Unique identifier for each record	Integer
VLC-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
VLC-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
VLC-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
VLC-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
VLC-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
VLC-HP	Average rent price	Float
VLC-AMI	Average monthly income	Float
VLC-TI	Represents traffic intensity and distance to the nearest street or road. The first value indicates the traffic intensity, and the second value indicates the distance to the nearest street or road	Text (requires data processing for numerical analysis)
VLC-SP	Sale price	Float
VLC-%-UNEMP	Percentage of unemployed individuals	Float
VLC-%-LOW-ED	Percentage of individuals with low educational attainment	Float
VLC-%-ADULT-POP	Percentage of adult population	Float
VLC-GR-AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
VLC-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)
VLC-RE-CENTERS	Represents the distance to the nearest reintegration centres	Float (may contain null values)

Datos de DATA_US_SFO.csv

NAME	DESCRIPTION	DATA TYPE
SFO-ID	Unique identifier for each record	Integer
SFO-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
SFO-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
SFO-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
SFO-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
SFO-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
SFO-HP	Average rent price	Float
SFO-TI	Represents traffic intensity and distance to the nearest street or road. The first value indicates the traffic intensity, and the second value indicates the distance to the nearest street or road	Text (requires data processing for numerical analysis)
SFO-SP	Sale price	Float
SFO-%-UNEMP	Percentage of unemployed individuals	Float
SFO-%-ADULT-POP	Percentage of adult population	Float
SFO-GR-AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
SFO-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)

Datos de TWT_ES_VLC.csv

NAME	DESCRIPTION	DATA TYPE
VLC-USER	Unique identifier for each user	Integer
VLC-TIME	Timestamp of the tweet, including date and time	Datetime
VLC-TWEETS	Text content of the tweet	Text
VLC-REPLYS	Number of replies to the tweet	Float (may contain null values)
VAL-RETWEETS	Number of retweets	Float (may contain null values)
VAL-LIKES	Number of likes on the tweet	Float
VAL-LAT	Latitude associated with the tweet's location	Float
VAL-LON	Longitude associated with the tweet's location	Float
VLC-CLASIF	Classification of the tweet regarding gender-based violence (1: Related, 0: Not related)	Integer
VLC-PROB	Probability that the tweet is related to gender-based violence (calculated by the NLP model)	Float
VLC-VEC	Feature vector generated from processing the tweet's text	Float

Datos de MAP_ES_VLC.csv

NAME	DESCRIPTION	DATA TYPE
VLC-ID	Unique identifier for each record	Integer
VLC-COD-PART	Participant code	Integer
VLC-INCID	Type of reported incident (e.g., Harassment, Intimidation, Rape)	Text
VLC-OBS-INS	Additional observations about insecurity	Text
VLC-DATA	Date and time of the report	Text
VLC-LON	Geographic longitude of the reported location	Float
VLC-LAT	Geographic latitude of the reported location	Float



Conclusión:

El proyecto subió datos a un repositorio abierto y desarrolló herramientas útiles, fomentando la investigación sobre violencia de género, en línea con el objetivo 5.1 de los ODS. Adoptó ciencia abierta y principios FAIR para asegurar datos públicos y reutilizables. La documentación y el rol de curador de datos garantizan integridad y accesibilidad, estableciendo un modelo para futuras investigaciones en gestión de datos abiertos.

LITERATURA RELACIONADA

- Abadal, E., Abad-García, F., Anglada, L., Boté-Vericad, J.-J., Esteve, A., González-Teruel, A., . . . Santos-Hermosa, G. (2023). Ciencia abierta en España 2023: informe de situación y análisis de la percepción.
- Anquela Julián, A. B., Balaguer-Puig, M., Gallego Salguero, A., Hernández-Zetina, S., & Vicente Chiva, J. (2023-11-30). Analysis of quantitative variables obtained from open data in gender violence prevention studies. Case Study: Valencia (Spain).
- Matarese, V., & Shashok, K. (2019, 04 03). Transparent Attribution of Contributions to Research: Aligning Guidelines to Real-Life Practices. (E. consultan, Ed.) Retrieved 07 06, 2024, from <https://doi.org/10.3390/publications7020024>