UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Geodésica, Cartográfica y Topográfica

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

Trabajo Fin de Máster

Máster Universitario en Ingeniería Geomática y Geoinformación

AUTOR/A: Martínez Montes, Carlos

Tutor/a: Anquela Julián, Ana Belén

Cotutor/a externo: Hernández Zetina, Sandra Lucía

CURSO ACADÉMICO: 2023/2024

“El presente documento ha sido realizado completamente por el firmante; no ha sido entregado como otro trabajo académico previo y todo material tomado de otras fuentes ha sido convenientemente entrecomillado y citado su origen en el texto, así como referenciado en la bibliografía”

MARTINEZ MONTES CARLOS - 23888254V

Firmado digitalmente por MARTINEZ MONTES CARLOS - 23888254V
Fecha: 2024.07.16 14:04:39 +02'00'

Resumen

Actualmente, la tendencia de datos abiertos se ha integrado a diversos campos y la investigación científica no ha quedado ajena a estos conceptos. El proceso de investigación y sus diversas etapas implica considerar no sólo las formas de conseguir los fondos o financiamiento para su correcta ejecución, sino que implican otras fases que involucren el ciclo de los datos que intervienen (creación, procesamiento, análisis) antes, durante y después de la investigación, para garantizar no sólo su reproducibilidad, sino su reutilización para futuras investigaciones. En este trabajo se presenta el esquema considerado para la organización de los conjuntos de datos involucrados en el Proyecto sobre violencia de género denominado Criteria Taronja, proyecto aprobado en el marco de las Subvenciones a grupos de investigación consolidados AICO 2022 con número de expediente CIAICO/ 2021 292 y subvencionado por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana. Se han integrado los principios FAIR para lograr la localización, accesibilidad, interoperabilidad y reutilización tanto de los datos del proyecto, y se describe el proceso metodológico utilizado para la recopilación, generación y procesamiento de los datos y software necesario para su interpretación.

Resum

Actualment, la tendència de dades obertes s'ha integrat en diversos camps i la investigació científica no ha quedat al marge d'aquests conceptes. El procés d'investigació i les seves diverses etapes impliquen considerar no només les formes d'aconseguir els fons o el finançament per a la seva correcta execució, sinó que impliquen altres fases que involucren el cicle de les dades que intervenen (creació, processament, anàlisi) abans, durant i després de la investigació, per garantir no només la seva reproductibilitat, sinó també la seva reutilització per a futures investigacions. En aquest treball es presenta l'esquema considerat per a l'organització dels conjunts de dades involucrats en el Projecte sobre violència de gènere denominat Criteria Taronja, projecte aprovat en el marc de les Subvencions a grups d'investigació consolidats AICO 2022 amb número d'expedient CIAICO/2021/292 i subvencionat per la Conselleria d'Innovació, Universitats, Ciència i Societat Digital de la Generalitat Valenciana. S'han integrat els principis FAIR per aconseguir la localització, accessibilitat, interoperabilitat i reutilització tant de les dades del projecte, i es descriu el procés metodològic utilitzat per a la recopilació, generació i processament de les dades i el programari necessari per a la seva interpretació.

Abstract

Currently, the open data trend has been integrated into various fields and scientific research has not been left out of these concepts. The research process and its various stages involves considering not only the ways of obtaining funds or financing for its correct execution, but also other phases that involve the cycle of the data involved (creation, processing, analysis) before, during and after the research, to guarantee not only its reproducibility, but also its reuse for future research. This paper presents the scheme considered for the organisation of the datasets involved in the project on gender violence called Criteria Taronja, a project approved within the framework of the Grants to consolidated research groups AICO 2022 with file number CIAICO/ 2021 292 and subsidised by the Department of Innovation, Universities, Science and Digital Society of the Generalitat Valenciana. The FAIR principles are integrated to achieve the localisation, accessibility, interoperability and reuse of the project data, and the methodological process used for the collection, generation and processing of the data and the software necessary for its interpretation is described.

Índice de ilustraciones

Índice de Tablas

Índice

1. Introducción

1.1. Finalidad

El objetivo 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. Este *dataset* tiene como objetivo principal servir como base para la recopilación, difusión, y reutilización de datos, que apoye la investigación, la formulación de políticas y la acción colectiva destinada a eliminar la violencia contra las mujeres y niñas, en línea con el Objetivo de Desarrollo Sostenible 5.1 de poner fin a todas las formas de discriminación contra todas las mujeres y las niñas en todo el mundo. Al hacerlo, el proyecto busca mejorar la calidad de los datos disponibles, facilitar la interoperabilidad entre sistemas, y promover una respuesta global más eficaz y coordinada a este grave problema social.

1.2. Descripción de antecedentes

Este proyecto utiliza la información de los datos recogidos durante la investigación realizada por el grupo de investigación de tecnologías geoespaciales de la Universidad Politécnica de Valencia, en el marco de las subvenciones a grupos de investigación consolidados-AICO2022, con número de expediente CIAICO/2021/292, y subvencionado por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana.

Se toma como referente la base de datos empleada en el trabajo final de grado "Determinación de áreas libres de violencia de género extraoficial en la ciudad de Valencia", realizado por Raquel Castillejo del Pozo como proyecto final para la obtención del grado en Geomática y topografía, para crear una nueva base de datos independiente y específica para este estudio.

Es fundamental destacar que en los últimos años se han impulsado numerosas políticas para incentivar la publicación de datos abiertos tanto a nivel internacional como local, convirtiéndolo en una tendencia para todos los proyectos de investigación. Estas políticas no solo promueven la transparencia y la accesibilidad, sino que también fomentan la colaboración interdisciplinaria y el intercambio de conocimiento. La implementación de datos abiertos permite a los investigadores validar resultados, replicar estudios y generar nuevas hipótesis, contribuyendo significativamente al avance de la ciencia y la tecnología.

A nivel internacional, la adopción de principios como los FAIR (*Findable, Accessible, Interoperable, Reusable*) y las directrices de la Carta Internacional de Datos Abiertos están transformando la gestión de los datos de investigación. Estas normativas mejoran la accesibilidad y la calidad de los datos, facilitando la colaboración y el intercambio de información entre investigadores a nivel global. La tendencia hacia la ciencia abierta es evidente y se está consolidando como un estándar en los proyectos de investigación, promoviendo una mayor transparencia, reproducibilidad y el impacto de los resultados científicos.

La Universidad Politécnica de Valencia (UPV) ha desempeñado un papel destacado en la implementación de políticas y prácticas de ciencia abierta. A través de iniciativas como el

repositorio de datos abiertos de la UPV y su participación en proyectos europeos de ciencia abierta, la universidad ha demostrado un firme compromiso con la accesibilidad y la reutilización de los datos de investigación. Estos esfuerzos forman parte de una estrategia más amplia para mejorar la calidad y el impacto de la investigación, asegurando que los datos generados por la universidad estén disponibles para la comunidad global de investigadores y contribuyan al avance del conocimiento científico.

El informe "Ciencia abierta en España 2023" (Abadal, y otros, 2023) destaca cómo la ciencia abierta está ganando terreno como una práctica estándar en la investigación. Según el informe, la adopción de políticas de ciencia abierta en instituciones académicas españolas está promoviendo una mayor transparencia y colaboración en la investigación.

La UPV, siguiendo estas tendencias, ha integrado principios de ciencia abierta en sus políticas institucionales, lo que no solo mejora la visibilidad y el impacto de la investigación realizada, sino que también facilita el cumplimiento de las normativas internacionales en la gestión de datos de investigación. También, es crucial mencionar el estado actual de los Objetivos de Desarrollo Sostenible (ODS). Al examinar la evaluación del progreso de los 17 objetivos (ilustración 1), se observa que el objetivo 5, que se centra en la igualdad de género, es uno de los que menos metas ha alcanzado.

Aunque se ha logrado un progreso razonable, es de suma importancia la creación de nuevas políticas y metas para impulsar el estudio y la implementación de medidas efectivas contra la violencia de género. Este proyecto contribuye a este esfuerzo al proporcionar datos y herramientas esenciales que pueden ser utilizados para mejorar las políticas y las intervenciones dirigidas a erradicar la violencia contra mujeres y niñas. La integración de los datos en repositorios abiertos fomenta una mayor colaboración y el intercambio de conocimientos, alineándose con las metas globales de los ODS y promoviendo una sociedad más justa y equitativa.

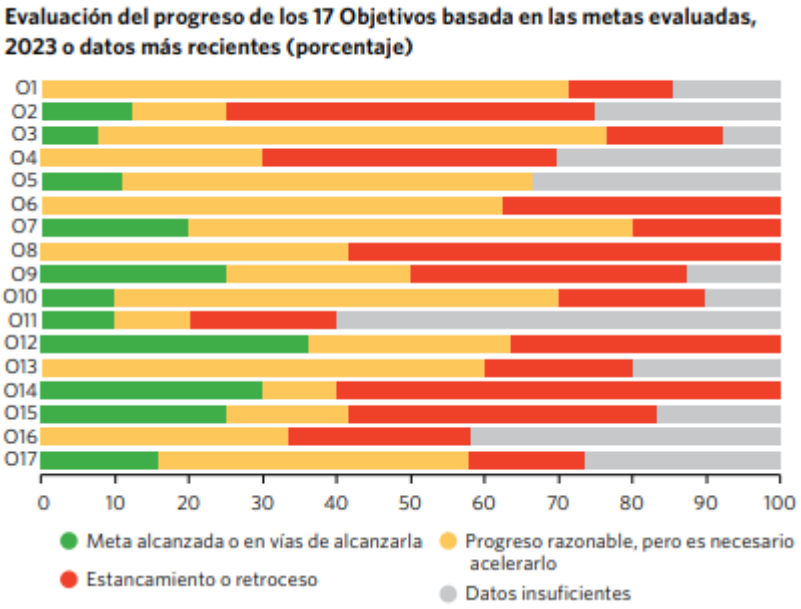

Ilustración 1: "Evaluación del progreso de los 17 Objetivos". Fuente: https://mdgs.un.org/

1.3. Localización

El proyecto de investigación se ha realizado en las ciudades de:

- Valencia: Ciudad piloto utilizada por el Grupo de Investigación de Tecnologías Geoespaciales de la Universidad Politécnica de Valencia, donde se han aplicado y validado metodologías innovadoras en la recolección y análisis de datos.

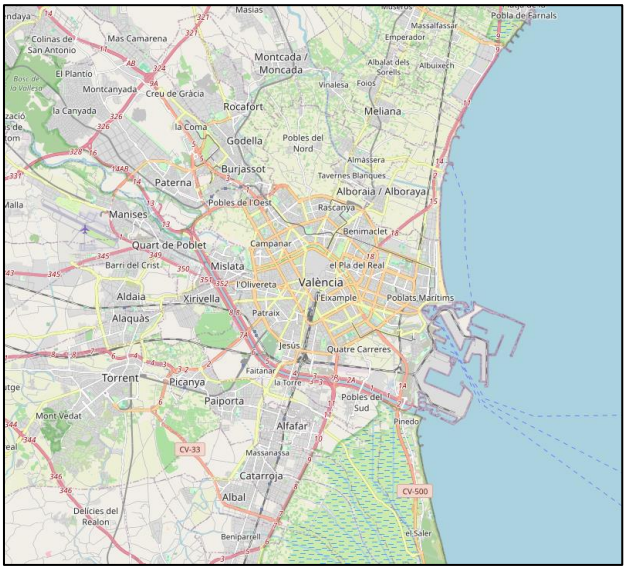

Ilustración 2: "Imagen de Valencia". Fuente: https://www.openstreetmap.org

- Dublín: Elegida por la abundancia y calidad de sus datos abiertos, que ofrecen una rica fuente de información para análisis urbanos y geoespaciales.

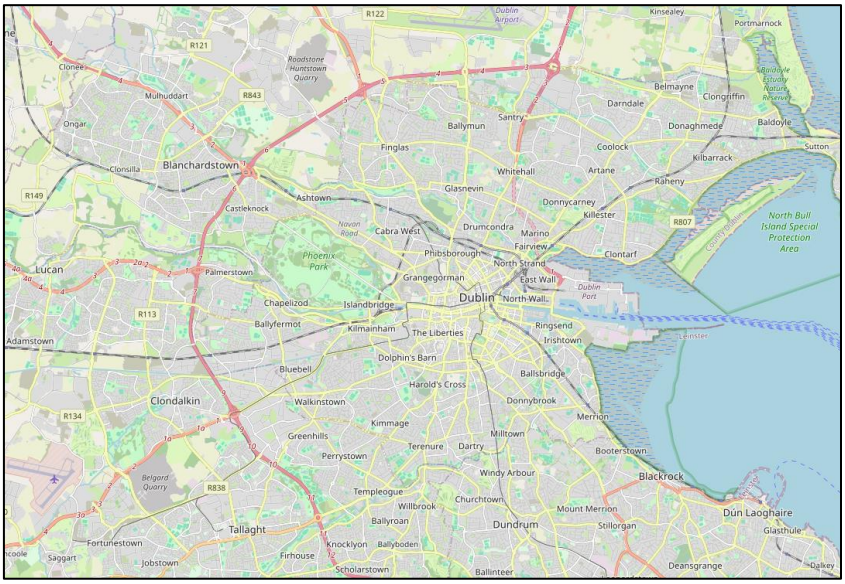

Ilustración 3: "Imagen de Dublín". Fuente: https://www.openstreetmap.org

- Toluca: Incluida en el estudio por sus iniciativas locales en la apertura de datos y participación ciudadana, proporcionando una perspectiva valiosa para el análisis comparativo.

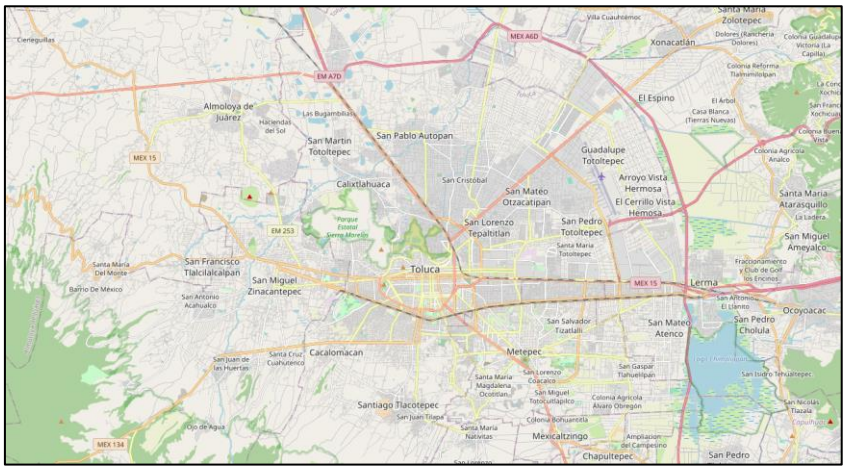

Ilustración 4: "Imagen de Toluca". Fuente: https://www.openstreetmap.org

- San Francisco: Seleccionada debido a su robusta infraestructura de datos abiertos y su historial de proyectos exitosos de ciencia ciudadana, lo que la convierte en un caso de estudio relevante para prácticas de datos urbanos.

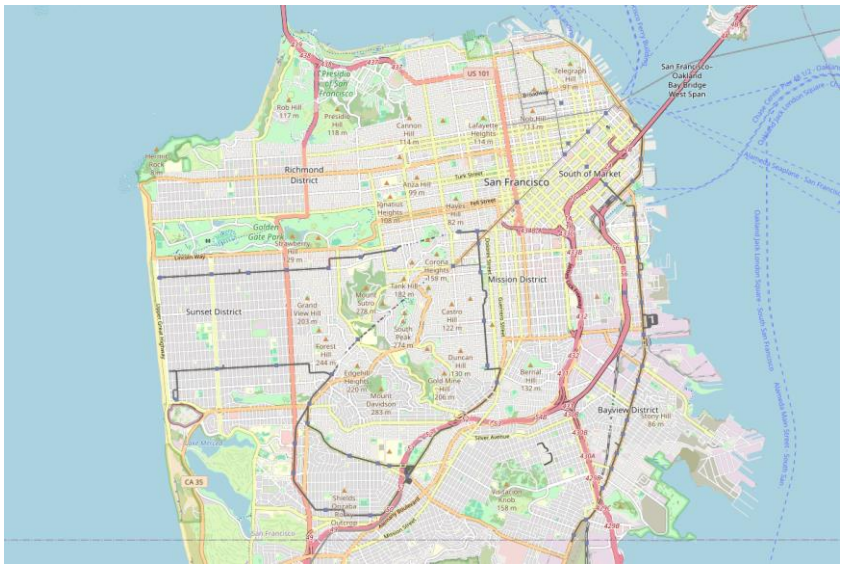

Ilustración 5: "Imagen de San Francisco". Fuente: https://www.openstreetmap.org

1.4. Conceptos previos

Para comprender completamente el alcance y la importancia de este proyecto, es esencial familiarizarse con algunos conceptos clave que forman la base de la ciencia abierta. Este apartado proporciona una breve introducción a estos conceptos, que son esenciales para comprender los objetivos del proyecto en el contexto de la investigación científica actual.

1.4.1. Principios FAIR (*Findable, Accessible, Interoperable, Reusable*)

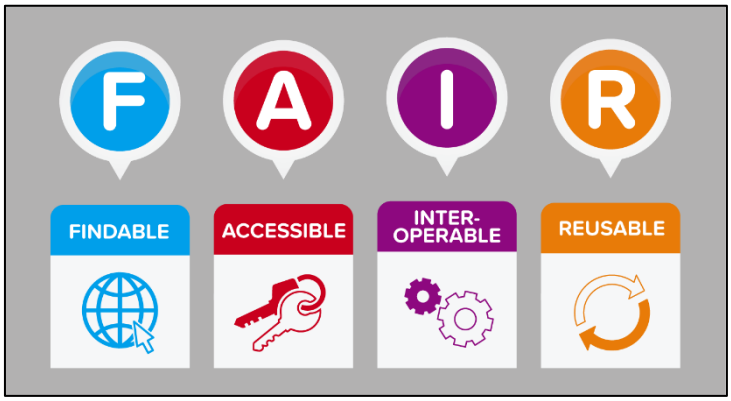

Ilustración 6: "Principios FAIR". Fuente: https://www.idecor.gob.ar/

En la sociedad actual se vive rodeados de datos, estamos inmersos en su cultura y se observa tanto su crecimiento continuo como la propia capacidad para generarlos, almacenarlos y procesarlos. El aumento de las aplicaciones tecnológicas capaces de crearlos y utilizarlos motivó que la comunidad científica volcada ya con la Investigación Abierta (Open Science) se orientase al acceso público de los datos obtenidos por experimentación, especialmente los obtenidos con fondos públicos.

De este modo surgió la necesidad de definir unas buenas prácticas para la publicación de datos científicos que fuesen claramente especificadas y ampliamente compartidas y aplicadas. El 15 de marzo de 2016 se publicó en la revista Scientific Data de Nature el artículo "**Principios FAIR para el manejo y administración de datos científicos** (Wilkinson, Dumontier, Aalbersberg, & Appleton, 2016)**.**

Los principios FAIR están diseñados para mejorar cómo se gestiona y accede a los datos científicos y digitales. Estos principios aseguran que los datos sean fáciles de encontrar, accesibles para todos, integrables con otros datos y reutilizables en distintos contextos. Aquí se explica cada uno:

- **Localizables (*Findable*):**
    - Definición: Los datos deben ser localizables por los usuarios tanto humanos como máquinas. Esto se logra mediante la asignación de identificadores únicos y persistentes y la creación de metadatos descriptivos.
    - Aspectos Clave:
        - **i.** Identificadores Persistentes (PID): Asignar identificadores únicos como DOI (*Digital Object Identifier*) para asegurar que los datos sean fácilmente localizables y referenciables.
        - **ii.** Metadatos Descriptivos: Crear metadatos detallados que describan el contenido, la fuente, y el contexto de los datos, facilitando su búsqueda y descubrimiento.
        - iii. Indexación: Asegurar que los datos y los metadatos estén indexados en sistemas de búsqueda relevantes.
- **Accesibles (*Accessible*)**
    - Definición: Los datos deben ser accesibles para el mayor número posible de usuarios mediante el uso de protocolos abiertos. Deben existir condiciones claras que permitan el acceso a los datos, incluso si estos requieren permisos.
    - Aspectos Clave:
        - **i.** Protocolos Abiertos: Utilizar protocolos estándar y abiertos (e.g., HTTP, FTP) para la recuperación de datos.
        - **ii.** Permisos y Licencias: Definir licencias claras que especifiquen cómo se pueden usar y redistribuir los datos, como Creative Commons.
        - **iii.** Acceso a Metadatos: Asegurar que los metadatos sean accesibles incluso cuando los datos en sí están restringidos.
- **Interoperables (Interoperable)**
    - Definición: Los datos deben ser compatibles con otros datos y sistemas mediante el uso de estándares y lenguajes comunes. Esto facilita su integración y análisis en conjunto con otras fuentes de datos.
    - Aspectos Clave:
        - **i.** Estándares de Datos: Utilizar estándares para la representación y estructura de los datos (e.g., XML, JSON, RDF).
        - **ii.** Vocabularios Controlados: Adoptar vocabularios y ontologías comunes para describir los datos y asegurar la coherencia semántica.
        - **iii.** Formatos Compatibles: Asegurar que los datos se puedan integrar fácilmente con otros sistemas y plataformas.
- **Reutilizables (Reusable)**
    - Definición: Los datos deben estar bien descritos y documentados, y deben compartirse bajo condiciones que permitan su reutilización en diferentes contextos, asegurando su aplicabilidad a largo plazo.
    - Aspectos Clave:
        - **i.** Documentación Completa: Proveer documentación exhaustiva que explique cómo se generaron, validaron, y pueden utilizarse los datos.
        - **ii.** Calidad y Conformidad: Asegurar que los datos cumplan con estándares de calidad y que se mantengan actualizados y verificables.
        - **iii.** Licencias Claras: Utilizar licencias que permitan la reutilización explícita de los datos, describiendo cualquier restricción.

1.4.2. Carta Internacional de Datos Abiertos

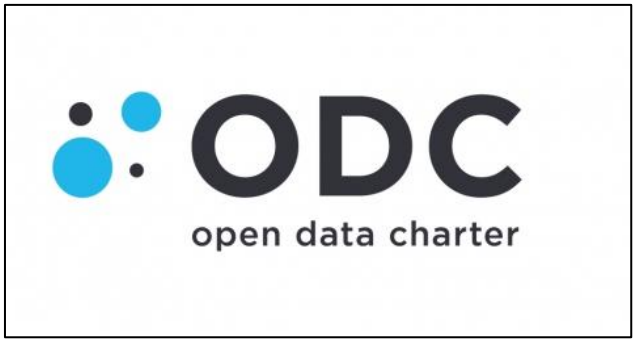

Ilustración 7: "Open Data Charter". Fuente: https://datos.gob.es

La Carta Internacional de Datos Abiertos establece un marco de principios para la publicación y gestión de datos abiertos. Su comprensión es esencial para la implementación efectiva de políticas y prácticas de datos abiertos en proyectos de investigación y gestión de datos.

La Carta se basa en ocho principios fundamentales que guían la apertura y gestión de datos. A continuación, se describen cada uno de ellos:

- **Apertura por Defecto**
  - Definición: Los datos deben ser abiertos y accesibles de manera predeterminada, a menos que existan razones legítimas para no hacerlo, como preocupaciones de privacidad, seguridad o confidencialidad.
  - Aplicación: Publicar datos proactivamente en formatos accesibles y reutilizables.
- **Oportunos y Exhaustivos**
  - Definición: Los datos deben publicarse tan pronto como estén disponibles y ser completos, reflejando la mayor cantidad posible de información relevante.
  - Aplicación: Proveer datos actualizados y completos en intervalos regulares.
- **Accesibles y Utilizables**
  - Definición: Los datos deben estar disponibles en formatos convenientes y accesibles que permitan su uso fácil, acompañados de la documentación necesaria.
  - Aplicación: Usar formatos estándar y legibles por máquina (e.g., CSV, JSON), y proporcionar documentación clara.
- **Comparables e Interoperables**
  - **Definición:** Los datos deben ser publicados en formatos que permitan la comparación y la integración con otros conjuntos de datos.
  - **Aplicación:** Utilizar estándares y esquemas comunes para facilitar la interoperabilidad.
- **Para Mejorar la Gobernanza y la Participación Ciudadana**
  - **Definición:** Los datos deben ser utilizados para mejorar la toma de decisiones, la eficiencia de los servicios, y fomentar la participación ciudadana.
  - **Aplicación:** Involucrar a los ciudadanos en la creación y uso de datos abiertos, y utilizar los datos para la rendición de cuentas.

- **Para el Desarrollo Inclusivo e Innovación**
    - **Definición:** Los datos abiertos deben promover la inclusión social y la innovación, facilitando la creación de nuevos productos, servicios, y aplicaciones.
    - **Aplicación:** Fomentar el uso de datos abiertos por parte de empresas, ONGs, y desarrolladores para crear soluciones innovadoras.
- **Gobernados por Normas Claras y Abiertas**
    - **Definición:** La gestión de datos abiertos debe estar regida por normas claras y políticas transparentes que promuevan su uso ético y responsable.
    - **Aplicación:** Desarrollar políticas y guías para la publicación y el uso de datos, garantizando la protección de datos sensibles.
- **Sostenibles**
    - **Definición:** Los programas de datos abiertos deben ser sostenibles, asegurando que los datos estén disponibles y mantenidos a largo plazo.
    - **Aplicación:** Garantizar financiación adecuada y estrategias para la conservación y actualización continua de datos.

1.4.3. Las licencias Creative Commons

Ilustración 8: "Licencias Creative commons". Fuente: https://blogs.iadb.org

Las licencias Creative Commons son un conjunto de licencias públicas de derechos de autor que permiten a los creadores otorgar permisos claros sobre cómo se puede usar, compartir y redistribuir su trabajo. Estas licencias proporcionan un marco flexible para la distribución de obras protegidas por derechos de autor.

- **Tipos de Licencias Creative Commons**

Las licencias Creative Commons combinan diferentes atributos para formar una licencia específica que se adapta a las necesidades del creador. A continuación, se describen los seis tipos principales de licencias Creative Commons:

- **CC BY (Reconocimiento)**
    - Descripción: Permite a otros distribuir, remezclar, adaptar y construir a partir de la obra, incluso con fines comerciales, siempre y cuando den crédito al creador original.
    - Uso Ideal: Para maximizar la difusión y el uso del trabajo sin restricciones importantes.

- **CC BY-SA (Reconocimiento – Compartir Igual)**
    - Descripción: Permite la distribución, remezcla, adaptación y construcción sobre la obra, incluso comercialmente, siempre que se otorgue crédito al creador y las nuevas creaciones se licencien bajo términos idénticos.
    - Uso Ideal: Para asegurar que las versiones derivadas sigan compartiendo bajo los mismos términos.
- **CC BY-ND (Reconocimiento – Sin Obras Derivadas)**
    - Descripción: Permite la redistribución comercial y no comercial, siempre que la obra se comparta sin cambios y en su totalidad, dando crédito al creador.
    - Uso Ideal: Cuando se desea permitir el uso sin modificaciones del trabajo original.
- **CC BY-NC (Reconocimiento – No Comercial)**
    - Descripción: Permite a otros distribuir, remezclar, adaptar y construir sobre la obra, pero no comercialmente, y siempre dando crédito al creador original.
    - **Uso Ideal:** Para permitir el uso libre en contextos no comerciales.
- **CC BY-NC-SA (Reconocimiento – No Comercial – Compartir Igual)**
    - Descripción: Permite la distribución, remezcla, adaptación y construcción sobre la obra no comercialmente, siempre y cuando se otorgue crédito al creador original y las nuevas creaciones se licencien bajo términos idénticos.
    - Uso Ideal: Para asegurar que las obras derivadas se compartan bajo los mismos términos y no se utilicen comercialmente.
- **CC BY-NC-ND (Reconocimiento – No Comercial – Sin Obras Derivadas)**
    - Descripción: Permite la redistribución, siempre y cuando se otorgue crédito al creador original, pero no se pueden realizar cambios en la obra ni usarla con fines comerciales.
    - Uso Ideal: Para permitir la distribución sin modificaciones y en contextos no comerciales.

- **Elementos de las Licencias Creative Commons**

Las licencias Creative Commons se basan en la combinación de cuatro elementos, que determinan el grado de libertad y restricciones impuestas a la obra:

- BY (Reconocimiento): Requiere que se dé crédito al autor original.
- NC (No Comercial): Restringe el uso de la obra para fines comerciales.
- ND (Sin Obras Derivadas): Prohíbe la creación de obras derivadas basadas en la original.
- SA (Compartir Igual): Requiere que las nuevas obras se distribuyan bajo una licencia idéntica a la original.

1.4.4. Licencia MIT

Ilustración 9: MIT License. Fuente: https://www.licen.cc/es/licencias/mit/

La Licencia MIT es una licencia de software libre que permite a los desarrolladores otorgar permisos claros sobre cómo se puede usar, modificar y redistribuir su código. Esta licencia proporciona un marco flexible y permisivo para la distribución de software protegido por derechos de autor.

- **Descripción de la Licencia MIT**
  La Licencia MIT permite a cualquier persona obtener una copia del software y los archivos de documentación asociados, y tratarlos sin restricciones, incluyendo los derechos de uso, copia, modificación, fusión, publicación, distribución, sublicencia y/o venta de copias del software, y permitir a las personas a las que se proporciona el software hacer lo mismo, siempre y cuando se incluya el aviso de copyright y la nota de la licencia en todas las copias o partes sustanciales del software.
- **Características Clave de la Licencia MIT**
  - Permisiva: La Licencia MIT es muy permisiva, permitiendo a otros usar, modificar y distribuir el software con muy pocas restricciones.
  - Compatibilidad: Es compatible con muchas otras licencias, incluyendo licencias de software propietario y otras licencias de código abierto.
  - Simplicidad: Su texto es breve y claro, lo que reduce la barrera de entrada para desarrolladores y organizaciones que desean utilizar tu código.
- **Uso Ideal de la Licencia MIT**
  La Licencia MIT es ideal para desarrolladores que desean maximizar la adopción y reutilización de su software sin imponer restricciones significativas. Es particularmente útil para proyectos que buscan:
  - Difusión Amplia: Facilitar que el software sea utilizado por la mayor cantidad de personas posible, incluidos los desarrolladores comerciales.
  - Contribuciones: Fomentar contribuciones y mejoras al código por parte de la comunidad, gracias a su permisividad.
  - Integración: Permitir que el software se integre fácilmente con otros proyectos, sean estos de código abierto o propietario.

1.4.5. Taxonomía CRediT: (Contributor Roles)

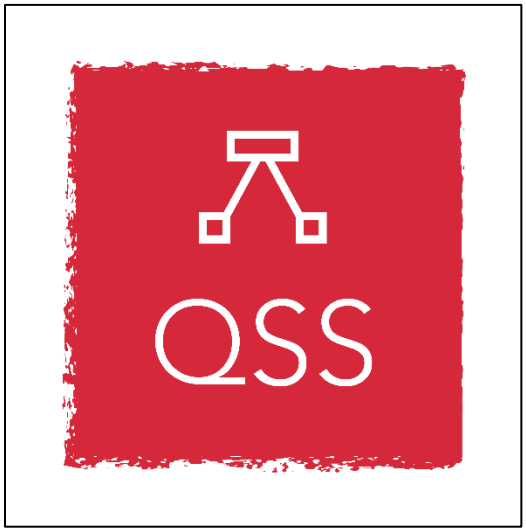

Ilustración 10: "Taxonomia CrediT". Fuente: https://direct.mit.edu/

CrediT (Contributor Roles Taxonomy) es una iniciativa desarrollada por CASRAI (Consortia Advancing Standards in Research Administration) para la diferenciación y reconocimiento de la contribución de cada firma en un artículo científico. (ULPGC, 2022)

Se trata pues de una clasificación ordenada desde la ciencia de la taxonomía de los distintos roles que pueden intervenir en un proyecto de investigación de forma que permita incluir y diferenciar la contribución de cada miembro en la firma posterior de cualquier publicación o artículo científico generado a partir de la misma.

CrediT define una taxonomía de alto nivel, con 14 roles, que aglutinan las distintas responsabilidades que desempeñan cada uno de ellos y su contribución a la producción académica científica de un documento de investigación

1. **Conceptualización**: Definir las ideas, los objetivos y los objetivos generales de la investigación.
2. **Curación de contenidos y datos:** actividades para generar metadatos, depurar datos y preservar datos de investigación para su reutilización posterior. Esto incluye el código de software cuando sea necesario para interpretar los propios datos.
3. **Análisis formal de los datos:** se refiere al uso de técnicas formales como estadísticas, matemáticas, computacionales u otras para analizar o sintetizar los datos del estudio.
4. **Adquisición de los fondos**: obtener financiamiento para el proyecto que da como resultado esta publicación.
5. **Investigación**: Proceso de investigación, incluida la realización de experimentos o la recolección de datos/evidencia.
6. **Metodología**: creación o desarrollo de la metodología; creación de modelos,
7. **Administración del proyecto:** La gestión y coordinación de la planificación y ejecución de la actividad de investigación son responsabilidades de la administración del proyecto.
8. **Recursos materiales**: Materiales para el estudio, reactivos, materiales, pacientes, muestras de laboratorio, animales, instrumentación, recursos informáticos u otras herramientas de análisis.

9. **Software**: Responsabilidad por supervisar y liderazgo en la planificación y llevar a cabo tareas de investigación, además de brindar tutoría externa al equipo central.
10. **Supervisión**: Responsabilidad de supervisión y liderazgo en la planificación y ejecución de actividades de investigación, incluyendo la tutoría externa al equipo central Responsabilidad de supervisión y liderazgo en la planificación y llevar a cabo de actividades de investigación, además de la asistencia externa al equipo central.
11. **Validación**: La evaluación de la reproducibilidad o reproducibilidad general de los resultados, experimentos y otros resultados de la investigación, tanto como parte de la actividad como por separado.
12. **Visualización**: Preparación, elaboración y/o presentación del trabajo en línea, especialmente la visualización/presentación de datos.
13. **Redacción-borrador original**: reparación, creación y/o presentación del trabajo publicado, en particular la elaboración del borrador inicial (excluyendo la traducción sustantiva)
14. **Redacción-revisión y edición**: Preparación, elaboración y/o presentación del trabajo que se ha publicado por los integrantes del grupo de investigación original, específicamente revisión crítica, comentario o revisión - incluyendo las etapas anteriores o posteriores a la publicación.

La curación de datos, uno de los roles definidos por CrediT, es de suma importancia para cualquier proyecto de investigación, ya que implica la generación de metadatos, la depuración y la preservación de los datos de investigación para su reutilización posterior. La figura del curador de datos es esencial para asegurar la integridad, accesibilidad y correcta interpretación de la información recopilada. Este rol no solo contribuye a la calidad y fiabilidad de los datos, sino que también facilita su trazabilidad y reutilización, aspectos clave que permiten que este proyecto sea un ejemplo en la gestión de datos y pueda guiar a otras investigaciones en la adopción de buenas prácticas en la administración de la información científica.

2. Objetivos

2.1. Objetivo general

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.

2.2. Objetivo específico

Para lograr este objetivo general, se plantean los siguientes objetivos específicos:

1. **Estandarizar los datos:**
    - Desarrollar un sistema de estandarización de los datos que garantice la consistencia y calidad de la información, facilitando su reutilización y análisis posterior. Permitirá definir una base sólida para el uso de estos datos en la elaboración de artículos científicos.

2. **Crear una documentación exhaustiva:**
    - Documentar de forma detallada del proceso de organización y subida de los datos a la base de datos abierta elegida, asegurando que cualquier usuario pueda entender y utilizar los datos correctamente. La documentación servirá como guía tanto para el uso del *dataset* como para la preparación del "*data paper*".
3. **Preparar el *dataset* como referencia para un data paper:**
    - Organizar y describir el *dataset* de manera que pueda ser utilizado como referencia principal para la redacción de un futuro *data paper*. El *data paper* explicará detalladamente el método de generación y organización de los datos proporcionando una guía clara y precisa sobre el proceso utilizado contribuyendo a la difusión del *dataset* y a la visibilidad de la base de datos.
4. **Promover la accesibilidad y colaboración:**
    - Diseñar el repositorio de modo que sea accesible para la comunidad científica, fomentando la colaboración y el intercambio de datos entre investigadores. Esto no solo beneficiará la creación de nuevos conocimientos, sino que también apoyará la elaboración de un *data paper* basado en estos datos, potenciando su impacto en la comunidad científica.

3. Datos

Para el estudio, se han elegido las ciudades de Valencia, Dublín y San Francisco. Estas ciudades fueron seleccionadas por su diversidad y representatividad en términos de tamaño, ubicación geográfica y características urbanas.

La recopilación de datos fue realizada por el grupo de investigación de Tecnologías Geoespaciales de la Universidad Politécnica de Valencia como se menciona en el artículo (Anquela Julián, Balaguer-Puig, Gallego Salguero, Hernández-Zetina, & Vicente Chiva, 2023-11-30).

Además, se dispone de datos de tweets obtenidos mediante *web scraping* y analizados utilizando algoritmos de procesamiento de lenguaje natural (NLP) y redes neuronales para identificar y clasificar tweets que discuten la violencia de género en la ciudad de Valencia.

Finalmente se incluyen también otros puntos recopilados durante varios mapathons conducidos por el campus de la Universidad Politécnica de Valencia para un proyecto científico que busca identificar ubicaciones potencialmente inseguras.

3.1. Valencia

Valencia es la tercera ciudad de España en número de habitantes, con aproximadamente 800,000 personas y un área de 135 km^2. Situada en la costa mediterránea, en el centro del Golfo de Valencia, y a orillas del antiguo cauce del río Turia (hoy convertido en un gran parque urbano de 136 hectáreas), es una ciudad plana, con espacios abiertos, un clima mediterráneo templado y una alta calidad de vida reconocida internacionalmente.

Se diseñó un modelo de datos espaciales con variables relacionadas con el estatus socioeconómico de la población y variables contextuales del espacio urbano de la ciudad de Valencia, relacionadas con la percepción de riesgo de las mujeres. Estos datos se obtuvieron de los portales de datos abiertos de diferentes instituciones oficiales: Instituto Nacional de Estadística (INE), Generalitat Valenciana (GVA), Ayuntamiento de Valencia (VLC), Infraestructura de Datos Espaciales de Valencia (IDEV) y OpenStreetMap (OSM).

Las variables consideradas son las siguientes:

- Variables socioeconómicas:
    - Desempleados (en porcentaje) (UNP): Personas que han buscado trabajo activamente en las últimas 4 semanas y están disponibles para trabajar. Información obtenida del INE por distrito.
    - Bajo nivel educativo (en porcentaje) (LLE): Población de 25 años o más que no ha completado la escuela secundaria. Información proporcionada por el INE y agrupada por distrito.
    - Ingreso medio del hogar (en euros) (AMI): Ingreso mensual promedio de todos los miembros de la unidad familiar. Información descargada del INE por secciones censales.
    - Densidad de población (en habitantes/km^2) (DEN): Número de habitantes entre 15 y 64 años por km^2. Información obtenida del INE por distrito.
- Servicios:
    - Servicios positivos: Hospitales (HOS), centros de salud (CDS) y estaciones de policía (POL). Información obtenida del portal de datos abiertos del VLC.
    - Servicios negativos: Locales de ocio nocturno (PUB), centros de acogida (ACG) y zonas verdes urbanas (ZVD). Información obtenida del IDEV, VLC y OpenStreetMap.

- Variables inmobiliarias:
    - Precios de venta de viviendas en euros/m^2 (VTA): Obtenidos del INE por distritos.
    - Precios de alquiler de viviendas en euros/m^2 (ALQ): Datos obtenidos del sitio web del Observatorio de la Vivienda de la UPV por distrito.
    - Densidad de tráfico (TFC): Valor promedio del número de vehículos por hora en una vía. Datos obtenidos del portal VLC y procesados con un script en Python.

3.2. Dublín

Dublín, la capital de Irlanda, tiene una población de aproximadamente 1.2 millones de habitantes y un área de 318 km^2. La ciudad se caracteriza por su rica historia, su vibrante cultura y su creciente economía tecnológica.

Las variables consideradas para Dublín son similares a las de Valencia, con algunas diferencias debido a la disponibilidad de datos:

- **Variables socioeconómicas**:
  - **Desempleados (UNP)**: Información obtenida del Central Statistics Office (CSO) por distrito.
  - **Bajo nivel educativo (LLE)**: Información proporcionada por el CSO y agrupada por distrito.
  - **Ingreso medio del hogar (AMI)**: Información descargada del CSO por secciones censales.
  - **Densidad de población (DEN)**: Información obtenida del CSO por distrito.
- **Servicios**:
  - **Servicios positivos**: Hospitales, centros de salud y estaciones de policía. Información obtenida de OpenStreetMap.
  - **Servicios negativos**: Locales de ocio nocturno y zonas verdes urbanas. Información obtenida de Dublín City Council y OpenStreetMap.
- **Variables inmobiliarias**:
  - **Precios de alquiler de viviendas (ALQ)**: Datos obtenidos del CSO por distrito.
- **Densidad de tráfico (TFC)**: Datos obtenidos de TomTom y procesados de manera similar a los de Valencia.

Para Dublín, la información se ha rasterizado con un tamaño de celda de 50 m^2 debido a la mayor superficie de la ciudad.

3.3. San Francisco

San Francisco, ubicada en California, EE. UU., es conocida por su diversidad cultural, su economía innovadora y sus icónicos paisajes urbanos. La ciudad tiene una población de aproximadamente 870,000 habitantes y un área de 121 km^2.

Las variables consideradas para San Francisco son las siguientes:

- **Variables socioeconómicas**:
  - **Desempleados (UNP)**: Información obtenida del U.S. Census Bureau por distrito.
  - **Densidad de población (DEN)**: Información obtenida del U.S. Census Bureau por distrito.
- **Servicios**:
  - **Servicios positivos**: Hospitales, centros de salud y estaciones de policía. Información obtenida de DataSF y OpenStreetMap.
  - **Servicios negativos**: Locales de ocio nocturno y zonas verdes urbanas. Información obtenida de DataSF y OpenStreetMap.

- **Variables inmobiliarias**:
    - **Precios de venta de viviendas (VTA)**: Información obtenida del U.S. Census Bureau por distritos.
    - **Precios de alquiler de viviendas (ALQ)**: Datos obtenidos de DataSF por distrito.
- **Densidad de tráfico (TFC)**: Datos obtenidos de TomTom y procesados de manera similar a los de Valencia.

Para San Francisco, la información se ha rasterizado con un tamaño de celda de 50 m^2 debido a la mayor superficie de la ciudad.

Para todas las ciudades, se ha realizado un preprocesamiento de datos para unificar los diferentes formatos y unidades espaciales. Se ha creado una base de datos espacial implementando las capas de datos en un GIS. Posteriormente, se realizó un análisis estadístico para estudiar la relación entre las variables, utilizando el coeficiente de correlación de Pearson y la matriz de correlación.

4. Metodología

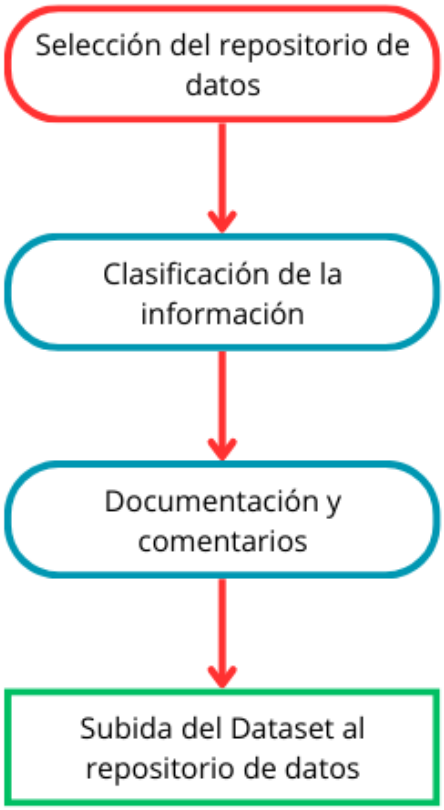

Ilustración 11: Organigrama metodología empleada

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 describen las etapas principales de este proceso:

4.1. Selección de Repositorios de Datos

Con el objetivo principal de organizar la información para subirla a un repositorio de datos abiertos, que será la pieza central para futuras publicaciones de artículos científicos y *data papers*, fue necesario investigar los repositorios más utilizados y los requisitos de cada uno para la subida de información. Entre los diversos repositorios destacaron el de Dryad, Figshare, Harvard Dataverse, ICPSR, Mendeley Data, Roper Center for Public Opinion Reesearch y Zenodo.

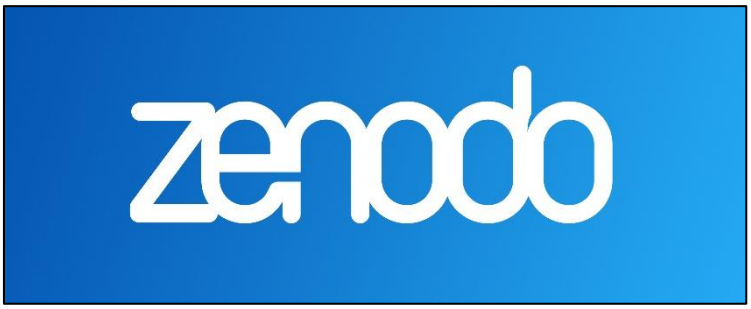

Ilustración 12: "Imagen Zenodo". Fuente: https://about.zenodo.org

Haciendo una comparación entre las diferentes páginas se ha seleccionado Zenodo ya que destaca del resto por diversas razones:

- Ofrece acceso y depósito de datos gratuitos, lo cual es una ventaja significativa frente a plataformas como Dryad, que cobra tarifas por la subida de los datos y FigShare, que tiene respaldo comercial, lo que puede ser un inconveniente para algunos usuarios.
- Proporciona una interfaz simple con características robustas como el control de versiones y la asignación de Identificadores de Objetos Digitales (DOIs) a todas las cargas, asegurando que los datos sean fácilmente citables.
- Soporta una amplia gama de tipos y formatos de datos sin imponer restricciones significativas.

Por otro lado, Zenodo está alojado por CERN, aprovechando la experiencia de la institución en la gestión de datos y la preservación a largo plazo, lo que añade un nivel de confianza y fiabilidad. La plataforma también apoya los principios FAIR (*Findable, Accessible, Interoperable, Reusable*), que se alinea perfectamente con los objetivos a alcanzar en el proyecto.

Finalmente, en términos de interoperabilidad, las APIs abiertas de Zenodo facilitan la integración con otros sistemas, y su integración con GitHub simplifica la preservación del software de investigación. Este punto también es sumamente importante para el proyecto porque para el procesado y representación de los datos se ha utilizado programas de creación propia.

4.2. Clasificación de la Información

Para asegurar claridad y reproducibilidad, se añadió documentación detallada a todos los datos y scripts generados. Se incluyeron comentarios en el código explicando cada sección y se adjuntaron archivos de texto que describen los datos, su estructura y origen. Esto ayuda a otros investigadores a entender y replicar el trabajo.

La clasificación se divide en tres categorías principales: datos en bruto, información tratada y software.

Los datos en bruto representan los datos obtenidos a través de los "mapatones" y los tweets recopilados por *Web Scrapping*. Clasificando estos datos se facilita a otros investigadores el acceso a los datos originales que permita realizar nuevas interpretaciones o validar los resultados publicados, incrementando así la transparencia y reproducibilidad de la investigación.

La información tratada comprende los datos que han requerido un proceso de análisis o transformación a partir de los datos en bruto. Esta categoría incluye desde resultados estadísticos y gráficos hasta modelos matemáticos y tablas resumen. Proporcionar acceso a la información tratada es vital para facilitar la comprensión de los datos obtenidos y permitir a otros investigadores replicar los análisis realizados.

El software incluye todo tipo de código, scripts y programas desarrollados para la recolección, procesamiento o análisis de los datos. La clasificación y preservación del software es el punto de partida para que otros investigadores puedan utilizar y adaptar estas herramientas en sus propios estudios. Zenodo facilita la integración con GitHub, permitiendo la captura automática de snapshots del repositorio de software y la asignación de DOIs para cada versión, asegurando que el software sea citable y accesible de manera efectiva. Es recomendable proporcionar una documentación detallada del código, incluyendo instrucciones de uso y licencias, para maximizar su utilidad y reutilización.

Con la clasificación de la información en estas categorías se mejora la organización y el acceso a los datos cumpliendo con los principios FAIR (*Findable, Accessible, Interoperable, and Reusable*). Este enfoque facilita la colaboración científica, la validación de resultados y el avance del conocimiento al hacer que los datos y herramientas sean fácilmente accesibles y reutilizables por la comunidad científica. Además, estos datos servirán como pieza central para futuros artículos y *data papers*, proporcionando una base sólida y verificable para nuevas investigaciones y publicaciones.

En resumen, la metodología de clasificación adoptada por Zenodo, respaldada por su integración con GitHub y el uso de metadatos detallados, garantiza que los datos en bruto, la información tratada y el software sean gestionados de manera eficiente y accesible, promoviendo así la transparencia y la reproducibilidad en la investigación científica.

4.3. Documentación y Comentarios

Para asegurar la claridad y la reproducibilidad, se añadió documentación exhaustiva a todos los datos y scripts generados. Esta documentación no solo incluye comentarios detallados en el código, sino también un documento Word que explica la estructura de los diferentes ficheros de datos.

4.3.1. Comentarios en el Código

En el caso de los scripts de datos, se incluyeron comentarios detallados en el código fuente. Estos comentarios explican la función de cada sección del script, las transformaciones realizadas y cualquier supuesto relevante. Esto facilita a otros investigadores la comprensión del flujo de trabajo y la replicación del análisis. Los comentarios en el código abordan aspectos como las funciones y métodos, detallando el propósito y el uso de cada uno, las transformaciones de datos que explican los cambios aplicados como filtrado, agregación o normalización, y cualquier supuesto realizado durante el análisis.

4.3.2. Documentación en Word

Para los ficheros de datos, se adjuntó un documento Word que describe en detalle la naturaleza de los datos, su estructura y su procedencia. Este documento proporciona una explicación clara de cada conjunto de datos, incluyendo detalles como el formato, las variables incluidas y su origen.

El documento Word incluye una introducción que explica el propósito de la documentación y cómo debe ser utilizada por los investigadores. Luego, se detalla la estructura general de los ficheros de datos, describiendo el formato y el esquema de las tablas. Además, se proporciona una explicación detallada de cada fichero individual.

Por ejemplo, el *dataset* de Valencia (DATA_ES_VLC.csv) se describe como un conjunto de datos procesados en una malla de 25 metros cuadrados. Se explica la estructura de las columnas, como VLC-ID, que es un identificador único para cada registro, VLC-GEOM, que contiene la geometría del área en formato GeoJSON, y VLC-D/N, que indica si el registro corresponde a datos tomados durante el día o la noche.

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-RP	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)

Tabla 1: "Datos de DATA_ES_VLC.csv"

Asimismo, el *dataset* de Dublín (DATA_IR_DUB.csv) se describe con un enfoque similar, indicando que los datos se han procesado en una malla de 50 metros cuadrados. La estructura es la siguiente:

NAME	DESCRIPTION	DATA TYPE
DUB-ID	Unique identifier for each record	Integer
DUB-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
DUB-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
DUB-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
DUB-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
DUB-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
DUB-AMI	Average monthly income	Float
DUB-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)
DUB-SP	Sale price	Float
DUB-%-UNEMP	Percentage of unemployed	Float
VLX-%-LOW-ED	Percentage of individuals with low educational attainment	Float
DUB-%-ADULT-POP	Percentage of adult population	Float
DUB-GR.AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
DUB-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)

Tabla 2: "Datos de DATA_IR_DUB.csv"

El *dataset* de San Francisco (DATA_US_SFO.csv) sigue el mismo esquema de descripción, indicando que se ha procesado en una malla de 50 metros cuadrados. El esquema es el siguiente:

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-RP	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)

Tabla 3: "Datos de DATA_US_SFO.csv"

Para el *dataset* de tweets recopilados de Valencia (TWT_ES_VLC.csv), se proporciona una descripción detallada de las columnas que incluyen VLC-USER, VLC-TIME, VLC-TWEETS, VLC-REPLYS, VAL-RETWEETS, VAL-LIKES, VAL-LAT, VAL-LON, VLC-CLASIF, VLC-PROB y VLC-VEC. Se explica que estos datos han sido analizados utilizando técnicas de procesamiento de lenguaje natural (NLP) para clasificar y entender mejor los patrones de violencia de género en los tweets. La estructura es la siguiente:

NAME	DESCRIPTION	DATA TYPE
VLC-USER	Unique identifier for each user	Text
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

Tabla 4: "TWT_ES_VLC.csv"

El conjunto de datos de puntos registrados durante los mapatones de Valencia (MAP_ES_VLC.csv) contiene las ubicaciones donde los usuarios han percibido inseguridad o han sido víctimas de acoso. Se ha añadido una descripción para identificar las diversas columnas y el tipo de datos que se registran. Las columnas son las siguientes: VLC-ID, VLC-COD-PART, VLC-INCID, VLC-OBS-INS, VLC-DATA, VLC-LON y VLC-LAT. Como se puede observar a continuación:

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, Insecurity, Both).	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

Tabla 5: "MAP_ES_VLC.csv"

Esta documentación garantiza que cualquier usuario, ya sea parte del equipo original o un investigador externo, pueda comprender, utilizar y citar correctamente los datos. La combinación de comentarios detallados en el código y una documentación exhaustiva en Word proporciona una guía completa para el uso y la replicación de los datos y análisis, facilitando la transparencia y la reproducibilidad de la investigación. Además, se ha llevado a cabo un proceso de anonimización de los datos para cumplir tanto con los acuerdos de privacidad aceptados durante la participación en el mapatón como con la privacidad de los usuarios que han escrito los tweets.

4.4. Subida del *dataset* al repositorio de datos

Requiere un estudio previo identificando la información que será requerida por la aplicación para completar el proceso; Zenodo ofrece varias configuraciones para adaptarse a las necesidades del cliente.

Una vez identificada la información requerida, se puede abordar el proceso de subir los datos al servidor, revisando que toda la documentación esté correctamente presentada y se cumplan los estándares de la plataforma. Los pasos a realizar son los siguientes:

- **Cumplimentar la información básica**:
  Al cumplimentar esta información básica, se debe proporcionar un título claro y conciso para el proyecto que refleje su contenido y propósito. Además, se debe incluir una descripción detallada explicando el contexto, la metodología y los objetivos del proyecto. Es fundamental incluir a todos los autores que han contribuido al trabajo, con sus respectivas afiliaciones y correos electrónicos. Finalmente, es muy importante añadir palabras clave relevantes que faciliten la búsqueda y categorización del proyecto.

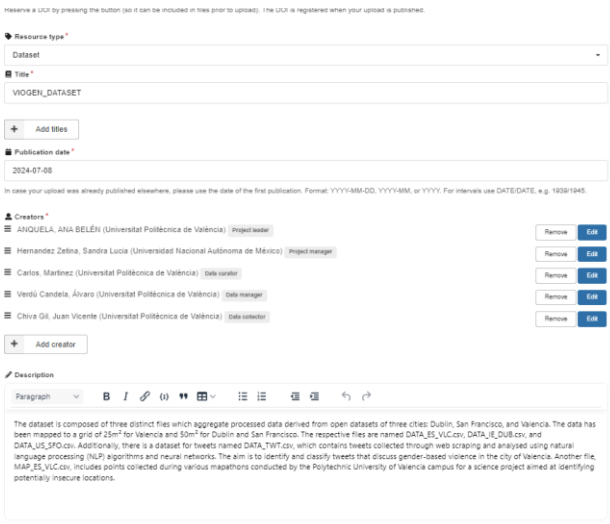

Ilustración 13: "Información básica del dataset". Fuente: https://zenodo.org/uploads/

- **Seleccionar opciones de licencia y accesibilidad**:
  En este paso se debe indicar la licencia por la que se ha optado. En nuestro caso, la de Creative Commons Attribution 4.0 International (CC BY 4.0) que permite la redistribución y reutilización de la obra con la condición de que se otorgue el crédito apropiado al creador original. La elección es particularmente relevante en el contexto de los datos abiertos y la investigación científica, ya que maximiza la accesibilidad y el uso de los datos sin restricciones significativas.

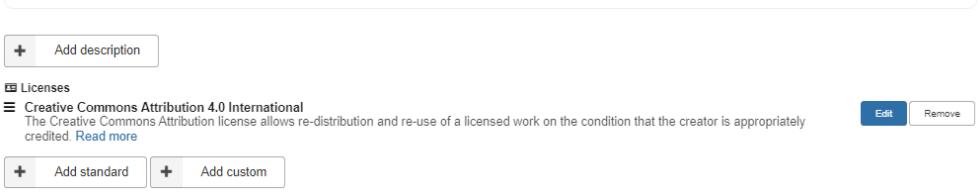

Ilustración 14: "Licencia Creative commons". Fuente: https://zenodo.org/uploads/

- **Información recomendada:**
  Se proporciona aquí información adicional recomendada que maximice la utilidad y el impacto de los datos subidos a Zenodo; entre otras, las palabras clave y el idioma del *dataset*.

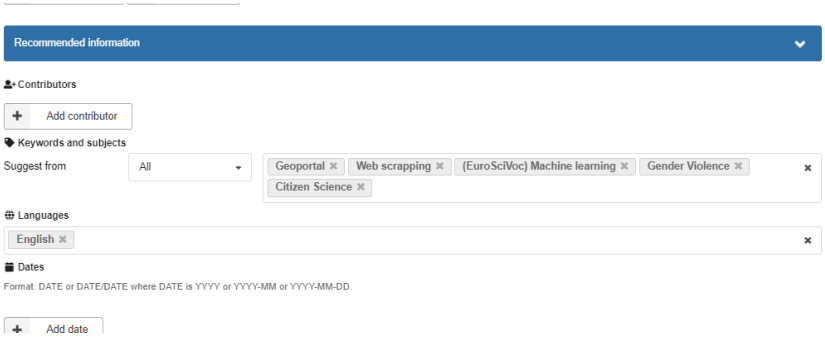

Ilustración 15: "Información recomendada". Fuente: https://zenodo.org/uploads/

- **Financiación:**
  Se incluye en este apartado información detallada sobre la financiación del proyecto para asegurar la transparencia y reconocimiento de las fuentes de apoyo económico. La financiación ha sido otorgada por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital bajo el programa AICO2022, con el número de expediente CIAICO/2021/292.

Ilustración 16: "Financiación". Fuente: https://zenodo.org/uploads/

- **Software:**
  Se ha subido software relevante para el procesamiento de datos al repositorio de datos de GitHub bajo la Licencia MIT. Esta licencia permite usar, copiar, modificar, fusionar, publicar, distribuir, sublicenciar y/o vender copias del software a cualquier persona, siempre que incluya el aviso de copyright y la nota de la licencia en todas las copias o partes sustanciales del software. Esta elección de licencia maximiza la adopción y reutilización del software, fomentando contribuciones de la comunidad y asegurando que el código sea fácilmente integrable con otros proyectos, tanto de código abierto como propietario.

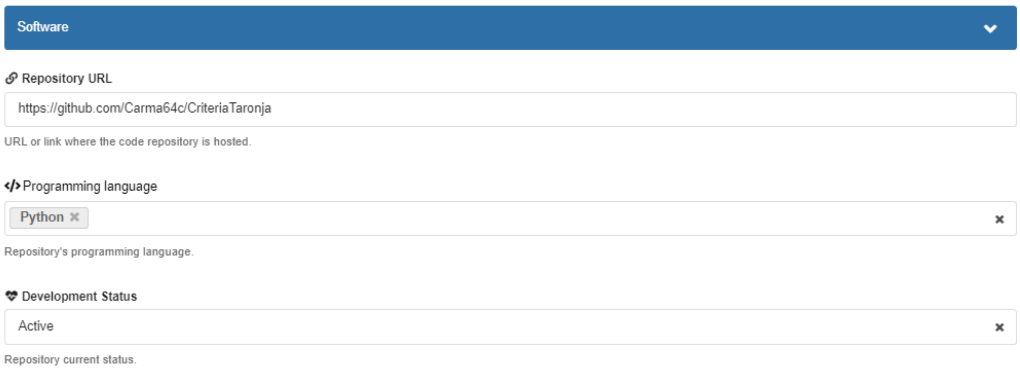

Ilustración 17: "Software". Fuente: https://zenodo.org/uploads/

Cumplimentados todos los apartados de la aplicación estamos en disposición de publicar el *dataset* en el repositorio de Zenodo incluyendo su descripción detallada, metadatos y documentación adicional que acompaña los datos.

La elección de la licencia de Creative Commons Attribution 4.0 International (CC BY 4.0) para los datos y la Licencia MIT para el software asegura que tanto los datos como el software sean accesibles, reutilizables y citable por otros investigadores. Además, se han proporcionado palabras clave relevantes y se ha registrado la información de financiamiento, lo que facilita la búsqueda y el reconocimiento adecuado del trabajo. Con todos estos elementos en su lugar, el *dataset* está listo para ser utilizado por la comunidad científica para avanzar en la investigación sobre la violencia de género y otros temas relacionados.

5. Resultados

El proyecto ha logrado varios resultados significativos, destacando tanto en la recopilación de datos como en la creación de herramientas y recursos para la investigación sobre violencia de género. A continuación, se detallan los principales resultados obtenidos:

- **Recolección y Organización de Datos**
  Se han recopilado y organizado diversos conjuntos de datos provenientes de múltiples fuentes, incluyendo iniciativas de ciencia ciudadana y datos abiertos proporcionados por instituciones gubernamentales. Los datos se han estructurado y procesado según los principios FAIR (*Findable, Accessible, Interoperable, and Reusable*), asegurando su calidad y utilidad para la investigación. Los conjuntos de datos incluyen:
    - DATA_ES_VLC.csv: Datos de Valencia procesados en una malla de 25 metros cuadrados.
    - DATA_IR_DUB.csv: Datos de Dublín procesados en una malla de 50 metros cuadrados.
    - DATA_US_SFO.csv: Datos de San Francisco procesados en una malla de 50 metros cuadrados.
    - TWT_ES_VLC.csv: *Dataset* de tweets recopilados de Valencia, analizados utilizando técnicas de procesamiento de lenguaje natural (NLP).
    - MAP_ES_VLC.csv: *Dataset* de puntos recopilados en el campus de la universitat politècnica de valència (upv) en diversos mapatones centrados en cartografiar zonas inseguras para las mujeres.
- **Desarrollo de Software**
  Se ha desarrollado y subido software relevante para el procesamiento y análisis de datos al repositorio de GitHub bajo la Licencia MIT. Este software incluye scripts y programas que permiten el procesamiento y análisis de datos de manera eficiente. La documentación exhaustiva y los comentarios detallados en el código facilitan su uso y reutilización por parte de otros investigadores.
- **Publicación en Zenodo**
  El dataset, junto con el software y la documentación, se ha publicado en el repositorio de datos de Zenodo. La publicación incluye:
    - Descripción detallada del *dataset* y los metadatos correspondientes.
    - Documentación exhaustiva que explica la estructura de los ficheros de datos, su procedencia y cómo utilizarlos.
    - Licencia: Los datos están bajo la licencia Creative Commons Attribution 4.0 International (CC BY 4.0), y el software bajo la Licencia MIT, asegurando su accesibilidad y reutilización.
    - Palabras clave y temas: Incluyen etiquetas relevantes como "Geoportal", "Web scrapping", "Machine learning", "Gender Violence", y "Citizen Science".
    - Información de financiación: Detalles sobre el apoyo financiero recibido de la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital bajo el programa AICO2022.

- **Facilitar la Investigación y la Colaboración**
  Con estos resultados, el proyecto proporciona una base sólida y verificable para futuras investigaciones y publicaciones científicas. Los datos y herramientas desarrollados no solo facilitan la investigación sobre la violencia de género, sino que también promueven la colaboración entre investigadores de diferentes disciplinas y geografías. La accesibilidad y reutilización de estos recursos, garantizada por las licencias seleccionadas, permiten que otros investigadores puedan construir sobre este trabajo, contribuyendo a una comprensión más profunda y a soluciones más efectivas para combatir la violencia de género.

Con estos datos, se han generado diversos resultados como un mapa de lugares potencialmente peligrosos para cada ciudad. Ver ejemplo (ilustración 18).

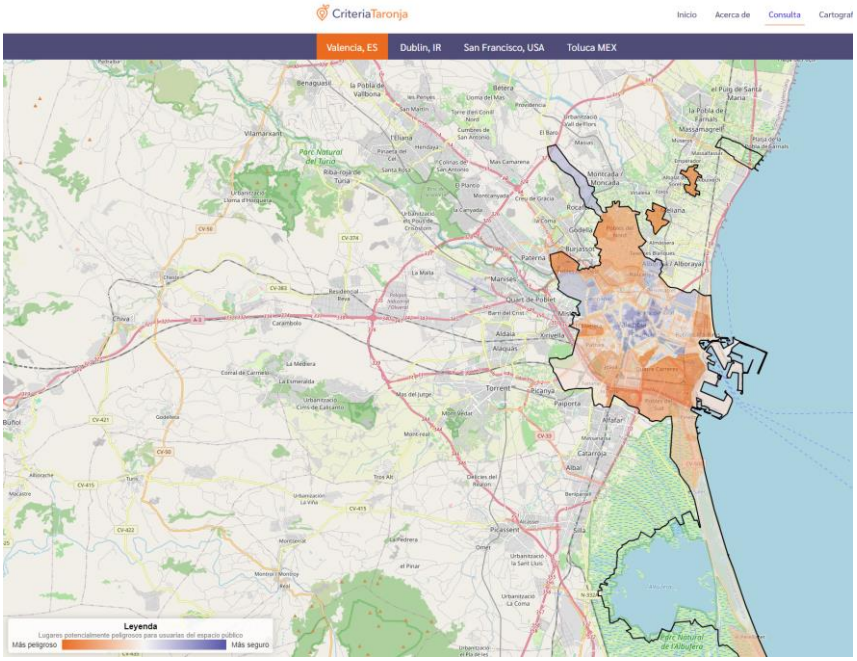

Ilustración 18: "Mapa de lugares potencialmente peligrosos". Fuente: https://gisserver.car.upv.es/viogen/

Además, se ha desarrollado una herramienta que permite modificar los supuestos utilizando los datos almacenados en los diversos *datasets*. Ver ejemplo (ilustración 19).

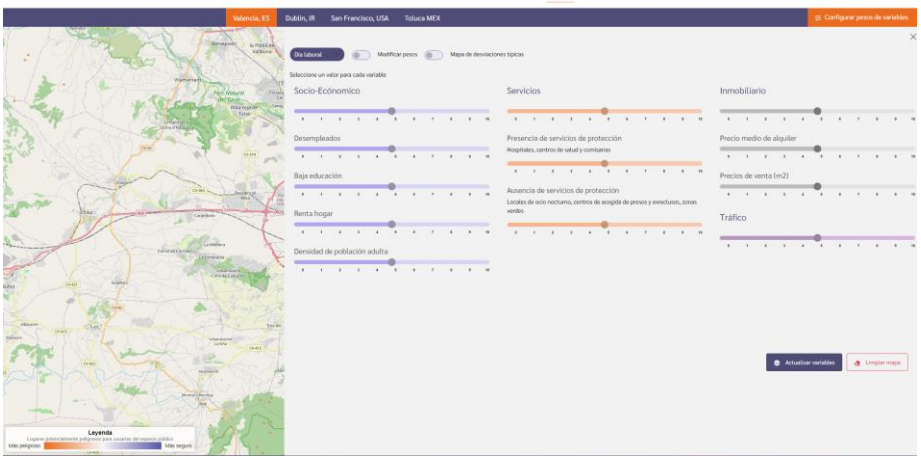

Ilustración 19: "Herramienta para modificar supuestos". Fuente: https://gisserver.car.upv.es/viogen/

6. Presupuesto

Se consideran los medios materiales y humanos para la organización del *dataset* en un periodo de 3 meses (abril-junio) trabajando en el proyecto a tiempo parcial. El servicio de programación ha dedicado personal equivalente de 63 días/hombre dentro del periodo. Se consideran tanto los costes directos del proyecto como los indirectos que se denomina Presupuesto de Ejecución Material (PEM). Sumando el beneficio industrial (6%) se obtiene el Presupuesto de Ejecución por Contrata y aplicando a este último los impuestos según ley (IVA 21%), el Presupuesto Total del proyecto.

Codigo	Capítulo	Unidad	Descripción	Ud Medida	Precio Unitario (€/Ud)	Cantidad	Importe
1	Costes Directos	DATA_ACUR_1	Remuneración Data Curator	Mes	2.083	3	6.249
		DATA_EQUIP_1	Ordenador incluyendo accesorios	Mes	84	3	252
		DATA_MATOFI_1	Material de oficina	Ud	95	1	95
2	Costes Indirectos	LOCAL_01	Alquiler Local Oficina	Mes	580	3	1.740
		SERV_01	Servicios (Agua, Electricidad, Telefonía)	Mes	115	3	345

Presupuesto Ejecución Material	8.681
Beneficio Industrial (6%)	521
Presupuesto Ejecución por contrata	9.202
I.V.A. (21%)	1.932
Presupuesto Total Proyecto	11.134

Tabla 6: "Presupuesto"

7. Conclusiones

El proyecto ha sido exitoso en varios aspectos clave, proporcionando un impacto significativo en la investigación sobre la violencia de género y estableciendo una base sólida para futuras investigaciones. A continuación, se presentan las principales conclusiones del proyecto:

- **Establecimiento de un Repositorio de Datos Abiertos**
  El proyecto ha logrado establecer un repositorio de datos abiertos y accesibles sobre la violencia de género, estructurado conforme a los principios FAIR y las directrices de la Carta Internacional de Datos Abiertos. Este repositorio no solo mejora la accesibilidad y la calidad de los datos disponibles, sino que también facilita su reutilización y análisis por parte de otros investigadores.
- **Estandarización y Documentación Exhaustiva**
  La estandarización de los datos y la creación de documentación exhaustiva aseguran que cualquier usuario, ya sea parte del equipo original o un investigador externo, pueda entender, utilizar y citar correctamente los datos. Esta documentación incluye detalles sobre la estructura de los ficheros de datos, su procedencia, y los procesos de recopilación y análisis, lo que aumenta la transparencia y reproducibilidad de la investigación.

- **Desarrollo de Herramientas de Software**
  El desarrollo de software relevante para el procesamiento y análisis de datos, publicado bajo la Licencia MIT, proporciona a la comunidad investigadora herramientas efectivas y reutilizables. Estas herramientas, junto con la documentación detallada y los comentarios en el código, facilitan la adopción y contribución al proyecto por parte de otros desarrolladores.
- **Publicación y Difusión en Zenodo**
  La publicación del *dataset* en Zenodo, con la inclusión de metadatos detallados, palabras clave relevantes y la información de financiación, asegura que los datos sean fácilmente localizables y utilizables. La elección de la licencia de Creative Commons Attribution 4.0 International (CC BY 4.0) para los datos y la Licencia MIT para el software maximiza la accesibilidad y reutilización, fomentando la colaboración y el intercambio de conocimiento.
- **Facilitar la Investigación y la Colaboración**
  El proyecto ha proporcionado una plataforma integral que no solo facilita la investigación sobre la violencia de género, sino que también promueve la colaboración entre investigadores de diferentes disciplinas y regiones. Los datos y herramientas desarrollados permiten realizar nuevas investigaciones y formular políticas más informadas y efectivas para combatir la violencia de género.
- **Impacto a Largo Plazo**
  El impacto a largo plazo de este proyecto incluye la mejora continua en la calidad y disponibilidad de datos sobre la violencia de género, el fomento de una mayor colaboración científica y el avance en la formulación de políticas y acciones basadas en evidencia. Al proporcionar recursos accesibles y bien documentados, el proyecto contribuye a una comprensión más profunda y soluciones más efectivas para este problema social crítico.

El proyecto ha cumplido con éxito sus objetivos de establecer un repositorio de datos abierto, desarrollar herramientas de software útiles, y fomentar la colaboración y la investigación en el ámbito de la violencia de género.

Pero, además, a la hora de evaluar la contribución general del trabajo a los objetivos definidos es necesario contraponer conceptos como fondo/forma o continente/contenido.

Mientras que el contenido del trabajo supone una contribución específica al objetivo 5.1 de los OSD, de poner fin a todas las formas de discriminación contra todas las mujeres y las niñas en todo el mundo, no es menos relevante que a la hora de plantear el formato (continente) para su comunicación a la comunidad científica se ha apostado por la ciencia abierta, con gran penetración en las últimas décadas, recogiendo los principios FAIR en la definición de datos y observando normativa establecida.

Los procedimientos empleados que se documentan en el trabajo y el conjunto de datos generados públicos, trazables y reutilizables hacen de este proyecto un ejemplo de estudio y tratamiento de datos públicos que puede guiar a otras investigaciones en la gestión de los datos y las pautas a seguir. Además, la figura del curador de datos es esencial para asegurar la integridad, accesibilidad y correcta interpretación de la información recopilada.

8. Bibliografía

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

Naciones Unidas. (2023). *Informe de los Objetivos de Desarrollo Sostenible 2023: Edición especial*. Naciones Unidas. doi: ISSN:2521-690

Raffaghelli, J. M. (2020). Supporting the development of critical data literacies in higher education: building blocks for fair data cultures in society. *17*(58). Retrieved from https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-020-00235-w

TEXAS, U. O. (2021, 07 20). *DIVISION OF DIGITAL STRATEGY AND INNOVATION*. Retrieved 06 16, 2024, from https://digitalstrategy.unt.edu/clear/teaching-resources/copyright-guide/creative-commons.html

ULPGC. (2022, 02 03). *https://biblioteca.ulpgc.es*. Retrieved 06 07, 2024, from https://biblioteca.ulpgc.es/blogs/acceso-abierto/2022/02/03/credit-reconocimiento-de-la-contribucion-de-cada-firma-en-un-documento-academico

Wikimedia. (2024, 01 14). *Open Content - A Practical Guide to Using Creative Commons Licences/The Creative Commons licencing scheme*. Retrieved 06 14, 2024, from https://meta.wikimedia.org/wiki/Open_Content_-_A_Practical_Guide_to_Using_Creative_Commons_Licences/The_Creative_Commons_licencing_scheme

Wilkinson, M., Dumontier, M., Aalbersberg, I., & Appleton, G. (2016). The FAIR Guiding Principles for scientific data management and stewardship.

Zenodo. (2021, 10 27). *Zenodo user guide*. Retrieved 06 25, 2024, from https://zenodo.org/records/5603317

9. Cartografía

N

Leyenda

Puntos clasificados por el tipo de registro

- Acoso
- Inseguridad
- Ambos

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

Titulo del mapa:
Mapa de puntos de inseguridad/acoso registrados

Autor: Carlos Martinez Montes
Fecha: 02/07/2024

Escala: 1/30000

0 600 1.200 1.800 m

Logotipos:

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA GEODÉSICA CARTOGRÁFICA Y TOPOGRÁFICA