DATA MANAGEMENT PLAN

1. Data Summary

¿Cuál es el propósito de la recolección/generación de datos y su relación con los objetivos del proyecto?

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Justificar la creencia de que en las grandes ciudades con un gran volumen de tráfico existe una menor calidad del aire que se respira. Para ello se realiza una comparativa de calidad de aire y volumen de tráfico en las ciudades de Madrid y Barcelona.

¿Qué tipo y formatos de datos se generarán/recolectarán?

What types and formats of data will the project generate/collect?

Se generarán datos en archivos de extensión *.csv* (*comma separated values*) compuesto de valores, generalmente en *comma flotante*. Además, habrá coordenadas geográficas para las estaciones de medición. Y un breve texto descriptivo de las mismas.

¿Se re-usarán datos ya existentes? ¿Cómo?

Will you re-use any existing data and how?

No se reusará los datos existentes, ya que son íntegramente generados en este proyecto.

¿Cuál es el origen de los datos?

What is the origin of the data?

Los datos son generados mediante sensores instalados en las ciudades de Madrid y Barcelona.

¿Cuál es el tamaño esperado de datos?

What is the expected size of the data?

A continuación, se realiza una estimación del tamaño que ocupará el conjunto de datos por medición:

- **Mediciones de la calidad del aire en Madrid**

Nombre	Descripción	Bytes	Tipo
PROVINCIA	Código de la provincia	4	Int
MUNICIPIO	Código del municipio	4	Int
ESTACIÓN	Código de la estación	4	Int
MAGNITUD	Magnitud de medida (SO_2, CO, NO, NO_2, PM 2.5, PM 10, Nox, O_3, TOL, BEN, EBE, MXY, PXY, OXY, TCH, CH_4, NM HC)	3	String
PUNTO_MUESTREO	Punto de recogida de la muestra: Recoge 3 argumentos separados por '_': Código del punto, de la magnitud y de la técnica de medida	13	String
ANO	Año medida	4	Int
MES	Mes medida	4	Int
D0i	Contaminación del día i	1	Int
V0i	Validación de la contaminación. Solo validos los registros con 'V'	1	Char

En Madrid se tomarán datos de la calidad del aire cada día y se estima que tendrá un tamaño aproximado de 38 Bytes. Entonces, si se multiplica el tamaño de una toma de datos por las 25 estaciones de medición que se estiman instalar, resultan unos 950 Bytes, aproximadamente, que tendrán que almacenarse en el servidor diariamente.

- **Mediciones del tráfico en Madrid**

Nombre	Descripción	Bytes	Tipo
Id	Identificador único del punto de medida	4	Int
Fecha	Fecha con formato yyyy-mm-dd hh:mm:ss	8	Date
Tipo_elem	Nombre del tipo de punto de medida: Urbano o M30	7	String
Intensidad	Intensidad del punto de medida en el periodo de 15 minutos (vehículos/hora). Un valor negativo implica la ausencia de datos.	4	Float
Ocupación	Tiempo de ocupación del punto de medida en el periodo de 15 minutos (%). Un valor negativo implica la ausencia de datos.	4	Int
Carga	Carga de vehículos en el periodo de 15 minutos. Parámetro que tienen en cuenta intensidad, ocupación y capacidad de la vía y establece el grado de uso de la vía de 0 a 100. Un valor negativo implica la ausencia de datos.	1	Int
Vmed	Velocidad media de los vehículos en el periodo de 15 minutos (Km/h). Sólo para puntos de medida interurbanos M30. Un valor negativo implica la ausencia de datos.	4	Int
Error	Indicación de si ha habido al menos una muestra errónea o sustituida en el periodo de 15 minutos. N: no ha habido errores ni sustituciones. E: los parámetros de calidad de algunas de las muestras integradas no son óptimos. S: Alguna de las muestras recibidas era totalmente errónea y no se ha integrado.	1	Char
Periodo_integración	Número de muestras recibidas y consideradas para el periodo de integración.	4	Int

En este último caso, se supone que se instalarán unos 4200 sensores que realicen la toma de medida de los datos del tráfico de Madrid. Cada toma se realizará con una frecuencia de 15 minutos y con un peso aproximado de 37 Bytes. Por lo tanto, se espera que el servidor tendrá que recibir una cantidad de 14 MB diarios.

- **Mediciones de la calidad del aire en Barcelona**

Nombre	Descripción	Bytes	Tipo
Nom_cabina	Nombre de la estación que ha tomado la medida	60	String
Qualitat_aire	Calidad del aire (Buena, Regular o Pobre)	7	String
Codi_dtes	Identificador de la estación	4	Int
Zqa	Código de la zona	4	Int
Codi_eoi	Código europeo de la cabina que ha tomado la medida	4	Int
Longitud	Longitud geográfica	4	Float
Latitud	Latitud geográfica	4	Float
Hora_o3	Hora de la medición de O3 (cada hora)	7	String
Qualitat_o3	Calidad del índice O3	7	String
Valor_o3	Valor de la medida O3	4	Float
Hora_no2	Hora de la medición de NO2 (cada hora)	7	String
Qualitat_no2	Calidad del índice NO2 (Buena, Regular o Pobre)	7	String
Valor_no2	Valor de la medida de NO2	4	Float
Hora_pm10	Hora de la medición de PM 10 (cada hora)	7	String
Qualitat_pm10	Calidad de las partículas en suspensión (Buena, Regular o Pobre)	7	String
Valor_pm10	Valor de la medida de PM 10	4	Float
Generat	Fecha y hora de cuándo se ha generado la medida	8	Datetime
DataTime	Timestamp de la hora de creación del fichero	8	Datetime

Cada toma de datos de la calidad del aire de Barcelona se realizará cada hora y se estima que tendrá un tamaño de 157 Bytes. Esto resulta que se almacenarán, aproximadamente, 4 Kb diarios por cada estación.

Por lo tanto, con el supuesto de colocar 10 estaciones habrá que almacenar un total de 37 Kb diarios.

- **Mediciones del tráfico en Barcelona**

Nombre	Descripción	Bytes	Tipo
IdTram	Identificador del tramo de medición	4	Int
Data	Fecha de la medición	8	Datetime
EstatActual	Estado actual (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)	1	Int
EstatPrevist	Estado previsto en 15 minutos (0=nada, 1=muy fluido, 2=fluido, 3=denso, 4=muy denso, 5=congestionado, 6=atasco)	1	Int

Cada medida del tráfico de Barcelona se tomará cada 15 minutos y se estima que cada una tendrá un tamaño de 14 Bytes. Por lo tanto, si además se tiene en cuenta que se estima una cantidad de 500 tramos de medida, resulta que tendrán que poder almacenarse 0.64 MB diarios, aproximadamente.

Por lo tanto, en total se espera que el servidor debe soportar aproximadamente 15MB diarios, lo que supondrá unos 11 GB en los 2 años de duración del proyecto.

¿A quién le resultará útil ('data utility')?

To whom might it be useful ('data utility')?

- En el ámbito de la política, para contemplar nuevas leyes contra la contaminación de la zona y así colaborar contra el cambio climático.
- Para estudios científicos sobre el aire.
- Para empresas que quieran producir nuevos servicios/productos: coches menos contaminantes.

2. Fair data

- ***Makin data findable, including provisions for metadata***

¿Se pueden encontrar los datos producidos y/o usados en el proyecto con metadatos? ¿Se pueden identificar y localizar mediante los estándares de identificación (¿p.e. identificadores persistentes y únicos como DOI?

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Se han escrito metadatos que describen el proyecto, así como los datos producidos. Además, al subir la documentación del proyecto junto con los datos utilizados al repositorio de UCrea, éste generará automáticamente un identificador digital, persistente y único que puede encontrarse en los metadatos del proyecto.

¿Cuáles son los estándares de nomenclatura que se siguen?

What naming conventions do you follow?

Se va a usar el estándar de metadatos *Dublin Core* extendido para la generación de los metadatos. A continuación, se listan los cambios realizados respecto a la versión original de *Dublin Core:*

En los metadatos del proyecto se han realizado lo siguientes cambios:

1. En la etiqueta *attribute* se ha añadido un atributo *xml, name,* que añade información extra de lo que contiene esta etiqueta. Por ejemplo:
   - <dc:attribute name="Campo Barcelona">O3: Ozono</dc:attribute>
   - <dc:attribute name="Campo Madrid">CONTAMINACION_AIRE: magnitud del parámetro de contaminación</dc:attribute>

   Estos atributos indican medidas que se han tomado, en el contenido de la etiqueta, y a dónde pertenecen, en el contenido del atributo.
2. En la etiqueta *source* se ha añadido un atributo *xml, name,* indicando a dónde pertenece el recurso, es decir, a Madrid o Barcelona.

En cuanto a los metadatos de los datos:

1. Tanto en *source* como en *identifier* se ha añadido el atributo *name* para indicar a qué pertenece el recurso o identificador dentro del campo.
    - <dc:source name="Tráfico">http://opendata-ajuntament.barcelona.cat/data/es/dataset/trams</dc:source>
2. Además, se ha creado el atributo *legend* para incluir la leyenda de los datos.

¿Se proporcionan palabras clave que optimizarán las posibilidades de reutilización?

Will search keywords be provided that optimize possibilities for re-use?

Sí. Se listan a continuación:

- Calidad aire Madrid
- Tráfico carretera Madrid
- Contaminación Madrid
- Medio ambiente
- Calidad aire Barcelona
- Tráfico carretera Barcelona
- Contaminación Barcelona
- Salud

¿Se ofrece un control de versiones claro?

Do you provide clear version numbers?

Todo el proceso que hayan sufrido los datos, desde su toma hasta su estudio está descrito en el repositorio de GitHub. Por lo tanto, el control de versiones vendrá dado por la herramienta Git, que proporciona dicho control.

¿Qué metadatos van a generarse? En el caso de que no exista ningún estándar de metadatos en la disciplina, por favor, describe qué tipos de metadatos son creados y cómo.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Como se ha dicho anteriormente, se crearán metadatos de los conjuntos de datos que se listan a continuación, siguiendo el estándar de *Dublin Core* extendido:

- Metadatos sobre el proyecto
- Metadatos sobre los datos de la calidad del aire de Madrid
- Metadatos sobre los datos de la calidad del aire de Barcelona
- Metadatos sobre los datos del tráfico de Madrid
- Metadatos sobre los datos del tráfico de Barcelona

- ***Makin data openly accessible***

¿Qué datos producidos y/o usados en el proyecto serán publicados abiertamente? Si ciertos conjuntos de datos no se pueden compartir (o son repartidos bajo ciertas restricciones), explicar por qué, claramente separando las razones legales y contractuales provenientes de restricciones voluntarias.

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Será posible acceder a públicamente a todo el proyecto, tanto a los datos, como a la documentación. Además, en este caso los conjuntos de datos no tienen restricciones legales, por lo tanto, serán completamente accesibles.

¿Cómo se harán los datos accesibles (p.e. mediante repositorio)?

How will the data be made accessible (e.g. by deposition in a repository)?

Como el proyecto será llevado a cabo junto con la Universidad de Cantabria, los datos y el proyecto serán subidos al repositorio UCrea.

¿Qué métodos o programas informáticos son necesarios para acceder a los datos?

What methods or software tools are needed to access the data?

Se puede acceder a través un navegador a la página web de UCrea.

Por otro lado, como los archivos en los que se van a almacenar serán de formato *.csv* no hará falta ningún software para poder verlos, cualquier bloc de notas será suficiente. Sin embargo, se recomienda utilizar hojas de cálculo tipo *Excel*, *LibreOffice Calc* o *software* tipo *Pandas* de *Python, R*.

Si hay restricciones en el uso, ¿cómo se dará acceso?

If there are restrictions on use, how will access be provided?

Los datos sobre la calidad de aire y el volumen de tráfico serán accesibles públicamente y fuera de restricciones legales. Por lo que serán totalmente accesibles.

- ***Making data interoperable***

¿Son los datos producidos en el proyecto interoperables, esto es, se permite el intercambio y reciclado de datos entre investigadores, instituciones, organizaciones, países etc. (i.e. agregar formatos estándar, tanto como es posible complementar con *software* accesible (*open)*, y en particular facilitar combinar con diferentes conjuntos de datos de diferentes orígenes)?

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and facilitating re-combinations with different datasets from different origins)?

Sí, ya que los datos serán guardados en formato *.csv* y, además, accesibles públicamente por internet.

Además, se proporcionan metadatos de los datos para así facilitar lo máximo posible la combinación con otros datos de distintos orígenes. Aunque el formato de metadatos, *Dublin Core* extendido, no es estándar, las modificaciones que se ha hecho son mínimas y se describen para facilitar su comprensión.

¿Qué vocabularios de datos y metadatos, estándares o metodologías se seguirán para hacer los datos interoperables?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Metadatos: Dublin Core extendido.

Metodologías: Se incluye código escrito en *Python,* con el paquete *Pandas*.

- ***Increase data re-use (through claryfying licenses)***

¿Cuál será la licencia de los datos para permitir el mayor reciclado posible?

How will the data be licensed to permit the widest re-use possible?

La licencia de los datos será *Creative Commons 4.0.* Resumiendo, esta licencia permite:

- **Compartir** — copiar y redistribuir el material en cualquier medio o formato
- **Adaptar** — volver a mezclar, transformar y crear a partir del material para cualquier finalidad, incluso comercial.

Siempre y cuando se cumpla:

- **Reconocimiento** — Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

No hay restricciones adicionales — No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

¿Cuándo serán accesibles los datos para reciclado? Sin embargo, hace que la publicación se retrase, problemas con patentes, especificar por qué y cuánto tiempo se hará, con la intención de que los datos deberán ser accesibles tan pronto como sea posible.

When will the data be made available for re-use? If an-embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Como la publicación del proyecto y los datos se realizará en el repositorio de UCrea, una vez se publiquen se podrá acceder a ellos. La parte del proyecto del año 2018 tiene como fecha estimada para su publicación a partir del 16 de enero de 2019. Sin embargo, la relacionada con el año 2019 no estará disponible hasta el 29 de enero de 2020 según la estimación realizada.

Por otro lado, el estudio será gestionado por dicho repositorio y, por lo tanto, no se sabe cuánto tardará en solucionarse en este momento.

¿Cuánto tiempo se pretende que los datos puedan ser reciclados?

How long is it intended that the data remains re-usable?

El tiempo que el repositorio UCrea mantenga accesible los datos.

¿Se describen los procesos que aseguran la calidad de los datos?

Are data quality assurance processes described?

Sí. Como se ha dicho anteriormente, se incluyen los *scripts* de todos los procesos que se han hecho a la hora de trabajar con los datos.

3. Allocation of resources

¿Cuáles son los costes de hacer los datos de vuestro proyecto *FAIR*?

What are the costs for making data FAIR in your project?

Los datos ya cumplen los estándares FAIR:

- **Encontrable (findability):** Se proporcionan identificadores persistentes y únicos en los metadatos.
- **Accesible (accessibility):** Es accesible públicamente mediante el repositorio UCrea.
- **Interoperable (interoperability):** Se proporcionan los datos junto con metadatos siguiendo el estándar *Dublin Core* extendido explicado para facilitar la interoperabilidad.
- **Reciclable (reusability):** Los datos son accesibles públicamente para que cualquiera pueda acceder a ellos y usarlos en futuros proyectos.

4. Data security

¿Cómo se prevee la seguridad de los datos (incluyendo recuperación de datos, así como seguridad de almacenamiento y transferencia de datos personales)? ¿Están los datos almacenados de forma segura en repositorios certificados para la preservación y curación a largo plazo?

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)? Is the data safely stored in certified repositories for long term preservation and curation?

La seguridad de los datos y el estudio dependerá totalmente de la de la Universidad de Cantabria. No compete a este proyecto hacerse cargo de su seguridad.

5. Ethical aspects

¿Existen problemas legales o éticos que puedan causar problemas a la hora de compartir los datos? Estos pueden ser discutidos en el contexto de la revisión ética. Si es relevante, incluye referencias al capítulo de ética en la DoA

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Los datos no tienen ninguna información sensible que pueda suponer problemas éticos o legales.

¿En los cuestionarios existe información para el consentimiento de preservación a largo plazo, así como compartir los datos personales?

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

No procede. No hay datos personales.