CREXDATA

Critical Action Planning over Extreme-Scale Data

D1.2 Data Management Plan
Version 2.1

Documentation Information

Contract Number	101092749
Project Website	https://crexdata.eu/
Contractual Deadline	M18, 06.2024
Dissemination Level	Public
Nature	Report
Author	Jordi Roca, Xavi Llort (HYDS)
Contributors	Jens Pottebaum (UPB) Oliver Krüger (FDDO) Enes Derin (FDDO) Merlin Stampa (DRZ) Ivana Kruijff (DRZ) Ilona Láng-Ritter (FMI) Arnau Montagud (BSC) Thalia Diniaco (BSC) Miguel Ponce de León (BSC) Jonathan Orama (BSC) Manolis Kaliorakis (Kpler) Georgios Grigoropoulos (Kpler)
Reviewer	Antonios Deligiannakis (TUC)
Keywords	Data Management, FAIR principles, Data Sources, Data Sets

CREXDATA has received funding from the European Union's Horizon Europe programme under grant agreement number 101092749.

Change Log

Version	Author	Date	Description Change
0.1	Jordi Roca (HYDS)	23/03/2023	Creation
0.2	Jordi Roca, Xavi Llort (HYDS), Jens Pottebaum (UPB), Oliver Krüger (FDDO), Ilona Láng-Ritter (FMI), Arnau Montagud, Miguel Ponce de León (BSC), Manolis Kaliorakis (Kpler).	22/05/2023	Compilation of datasets from partners.
0.3	Jordi Roca, Xavi Llort (HYDS), Jens Pottebaum (UPB), Oliver Krüger (FDDO), Ilona Láng-Ritter (FMI), Arnau Montagud, Miguel Ponce de León (BSC), Manolis Kaliorakis (Kpler).	31/05/2023	Final summary and minor changes
1.0	Jordi Roca, Xavi Llort (HYDS), Jens Pottebaum (UPB), Oliver Krüger (FDDO), Ilona Láng-Ritter (FMI), Arnau Montagud, Miguel Ponce de León (BSC), Manolis Kaliorakis (Kpler).	23/06/2023	New template and minor changes from internal revision.
1.1	Jordi Roca (HYDS) Ivana Kruijff, Merlin Stampa (DRZ) Enes Derin (FDDO) Miguel Ponce de León, Thalia Diniaco, Jonathan Orama (BSC) Georgios Grigoropoulos (Kpler)	05/04/2024	Revision and addition of new datasets from partners
1.4	Jordi Roca (HYDS)	31/05/2024	Minor additions
1.7	Jordi Roca (HYDS) Ivana Kruijff, Merlin Stampa (DRZ) Enes Derin (FDDO) Miguel Ponce de León, Thalia Diniaco, Jonathan Orama (BSC) Georgios Grigoropoulos (Kpler)	18/06/2024	Minor changes after the internal review
2.0	Antonios Deligiannakis (TUC).	28/06/2024	Final version
2.1	Jordi Roca (HYDS), Miguel Ponce de León, Jonathan Orama (BSC), Georgios Grigoropoulos (Kpler), Deniz Özcan (UPB)	04/11/2024	Addressed additions required by the reviewers.

Contents

Executive Summary

This deliverable is an update of the Data Management Plan in CREXDATA provided at Month 18 of the project. It is based on the first version delivered at Month 6 and it will receive its final form in Month 36. To build the initial Data Management Plan we utilized the respective template of the European Research Council (ERC): ERC DMP. We also followed the guidelines for FAIR (findable, accessible, interoperable, reusable) data management in Horizon Europe programs. In this version we have updated and added several sections according to the actual progress of the project.

We provide a summary of the datasets that are input to or output of the project and describe our initial provisions for:

- making data findable
- making data openly accessible
- making data interoperable
- increasing data reusability

We further provide responses to questions related to allocation of resources, data security and ethical issues where and when they arise in the scope of CREXDATA.

1 Introduction

CREXDATA deals with the planning and decision making based on data of extreme scale and complexity. Three use cases of real-time critical situation management will be carried out involving numerous datasets of very different nature. Those datasets, that could be either input or output of algorithms, software tools and hardware devices, will be integrated in a Prediction-as-a-Service (PaaS) system as outcome of the project. The three use cases are:

- The Weather Emergencies Use Case is devoted to improving situational awareness in weather emergency situations, so that informed decisions are taken by civil protection avoiding disaster impacts. Two levels of data sources are considered: (i) local environment data including stationary sensors systems, as well as mobile robotic sensor platforms deployed in case of an incident and (ii) global environment data referring to current and forecasted weather data issued by Copernicus covering specific fields as forest fires, health, etc.
- The Health Crisis Use Case considers two different scales: (i) epidemiological models will be used to build digital twins of the movement and infection of populations and (ii) mechanistic multiscale models will simulate different treatments of patients as drug administration or the use of mechanical ventilators in severe patients with collapsed lungs among other interventions.
- Finally, the Maritime Use Case aims to develop a vessel routing and route forecasting solutions in emergency situations that performs for all vessels of a fleet simultaneously (instead of on-demand request per vessels).

The purpose of this document is to describe the data embraced in CREXDATA and the procedures related with its management. It is organized according to the Horizon Europe DMP template[1]. In the following sections, for each use case we first provide an elaborated data summary, and we then proceed with explaining our provisions for making the corresponding datasets compliant with the FAIR principles.

This document will reflect the information available at this point of the project and will be updated in future versions. In particular, some new datasets (not present in M6 version of the document) have been included in this version according to the progress of the project. Find them below, classified by use case.

For the Weather Emergencies use case new data are:

- MIKE+ simulations
- Geosphere Austria layers (Innsbruck)
- Weather stations of Tyrol (Innsbruck)
- Water level data (Innsbruck)
- Precipitation data (Innsbruck)
- Weather stations measurements (Dortmund)
- Sewer network (Dortmund)
- Digital Elevation Model (Dortmund)
- Traffic network (Dortmund)
- Parking lots (Dortmund)

1 Can be found looking for “Data Management Plan” in the document searcher of the following web page: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON

- Rivers and streams data (Dortmund)
- Flood plains (Dortmund)
- Development WebODM (Dortmund)
- Type of development (Dortmund)
- Object plans (Dortmund)
- Vital data Firefighters (Dortmund)
- Social media posts (6 datasets)
- Disaster data streams (2 datasets)

For the Health Crisis use case the only added dataset is:

- COVID-19 reports from the ISCIIIs

For the Maritime use case the added datasets are:

- Synthetic AIS Dataset of Vessel Proximity Events
- GPS Coordinates for AR application (HoloLens 2)

2 Weather Emergencies Data and Management Procedures

In the Weather Emergencies use case, different roles in the emergency management are considered for contribution:

- Control center staff: Floods and fire emergencies are typically coordinated in control rooms with the need of situational awareness. This can be provided by large-scale services (EFAS, EFFIS, NWP models, …) and stationary sensor system (weather stations, weather radars, …). ARGOS is a multi-hazard early warning system developed by HYDS, which gathers and process all these kinds of data in real-time. In Argos, data fusion is only done with regards to mapping of geographical data sources and within mathematical models for specific topics.

In this use case, three implementations of ARGOS will be setup in Dortmund (Germany), Innsbruck (Austria) and Helsinki (Finland).

- On-site commands: Mobile robotic sensor platforms (UGV and UA) with their local viewpoint can complement large-scale data sources with different types of cameras, laser scanners, radar, etc. Augmented reality devices are also new data sources that may provide useful insight into an emergency situation for on-site people.

The corresponding use case for those technologies will take place at the German Rescue Robotics Center, the existing indoor and outdoor test bed will be used. For this test bed, exact terrain information, building information models etc. are available for reproducible evaluation settings. Both UAVs and UGVs can be operated within the area.

2.1 Weather Emergencies Data Summary

Due to the nature of this particular use case, is it worth to distinguish between:

- Data sources: capable and available to produce data within the project as part of demonstrator systems (web services, sensor systems, etc), taking benefit of open data policies of public institutions.
- Data sets: available for research purposes in the project, acquired from stakeholders (incident logs, emergency calls, weather data history, satellite image archives, etc) or generated within the project.

2.1.1 Data Sources

At this stage of the use case, various data sources are being considered although the concrete specifics for several of them is still to be defined. As mentioned in the introduction, two levels of sources are considered:

- Local environment data sources:
  - Robot localization system:

Description	Position, orientation, velocity and acceleration of a robot, derived by fusing data from various sensors. E.g., odometry is often obtained by fusing data from wheel encoders, IMUs, and gyroscopes. Can be enhanced or substituted by visual odometry

	(extracting camera movement from consecutive images). Global pose (I.e., relative to a known reference frame) is typically obtained by combining data from GNSS receivers (e.g., GPS), magnetometers, odometry and matching range sensor data (e.g., from a lidar) to a known map.
Purpose	Visualization of robot localization by control center staff
Type and Format	Pose data is transferred using designated ROS message types, e.g., geometry_msgs/Pose3D.
Processing and Size	A single pose message is <1kB, overall data rate depends on required update frequency.
Origin	DRZ robots
Use beyond project	Not intended

- Video stream per camera:

Description	RGB video stream by FPV and wide-angle camera
Purpose	Support on on-site operations
Type and Format	H-264, MPEG-4
Processing and Size	Common video stream rates (e.g. 30fps)
Origin	DRZ robots
Use beyond project	Not intended

- Thermal video stream:

Description	Images with temperature data, e.g., derived from near-infrared spectrum.
Purpose	Support on on-site operations
Type and Format	H-264, MPEG-4
Processing and Size	Common video stream rates (e.g. 30fps)
Origin	DRZ robots
Use beyond project	Not intended

- 3D environmental model:

Description	Map generated from robotic sensors. Commonly derived by combining localization data (see above) with range measurements (from cameras or lidars).
Purpose	Support on on-site operations
Type and Format	Some solutions build the map “live” (Simultaneous Localization and Mapping – SLAM). Others build the model offline by processing collected data (most often images) in batch, which is much slower but often more accurate (e.g., using WebODM).
Processing and Size	Typically < 10 Hz
Origin	DRZ robots
Use beyond project	Not intended

- AR (HoloLens 2):

Description	Video stream, incl. annotations of AR users in the scene.
Purpose	Support on on-site operations
Type and Format	MPEG4
Processing and Size	Typically > 10 Hz
Origin	Hololens 2
Use beyond project	Not intended

- MIKE+ simulations:

Description	Simulations of water level evolution during potential flood events.
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	dfs, res, csv
Processing and Size	Less than 100 kB per simulation.
Origin	Generated at UPB using ARGOS precipitation input.
Use beyond project	Studies on flooding in pilot sites.

- Geosphere Austria:

Description	Different layers reporting locations of vulnerable elements (schools, bridges, sewage installations, …)
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	Served as a Web Map Service layer
Processing and Size	Served as a Web Map Service layer
Origin	Geosphere Austria
Use beyond project	Multiple uses in geoscience.

- Weather stations of Tyrol:

Description	Weather station data: temperature, precipitation, wind…
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	Served through an API
Processing and Size	Served through an API
Origin	Austria government Open data
Use beyond project	Multiple uses in geoscience.

- Water level data of HYDRO Tyrol:

Description	Water level data
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	Served through an API
Processing and Size	Served through an API
Origin	Austria government Open data
Use beyond project	Multiple uses in geoscience.

- Precipitation data of Hydrographic Service of Tyrol

Description	Water level data

Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	Served through an API
Processing and Size	Served through an API
Origin	Hydrographic service of Tyrol
Use beyond project	Multiple uses in geoscience.

- Global environment data sources:
  - EFAS:

Description	EFAS is the European Flood Awareness System. It is operational since 2012 in collaboration with several European organizations responsible for producing and providing the flood information. It provides pan-European overview maps of flood reporting points with hydrological probabilistic forecasts.
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	NetCDF
Processing and Size	Acquired as a web service.
Origin	EFAS
Use beyond project	Institutions in charge of flood risk.

  - EFFIS:

Description	EFFIS is the European Fire Forecast Information System. Since 1998, EFFIS is supported by a network of experts from the countries in what is called the Expert Group on Forest Fires. In 2015, EFFIS became one of the components of the Emergency Management Services in the EU Copernicus program. It provides pan-European maps of fire indices based on meteorological inputs.
Purpose	To be used as an input for ARGOS implementations (Task 2.1).
Type and Format	NetCDF
Processing and Size	Acquired as a web service.
Origin	EFFIS

Use beyond project	Institutions in charge of fire risk.

- ECMWF Forecasts:

Description	High-resolution weather forecasts, which uses weather observations, and prior information about the Earth system, ECMWF's highest-resolution model. 4 forecast runs per day (at hours 00/06/12/18). Hourly steps to step 90 for all four runs.
Purpose	For the training and development of machine-learning-based tools in Finland or in other use cases and locations (Task 2.1).
Type and Format	GRIB, NetCDF.
Processing and Size	The full output of a simulation depends on the selected locations.
Origin	ECMWF (European Center of Medium-Range Weather Forecasts).
Use beyond project	All weather-related activities.

- Weather stations measurements in Germany:

Description	Measurements of temperature, wind and precipitation registered by stations of the German Weather Service (DWD) network
Purpose	To be used as an input for ARGOS implementation Dortmund (Task 2.1).
Type and Format	Numerical data conveniently georeferenced and timestamped in JSON format.
Processing and Size	New measurements may be available every 10 minutes, downloading and processing are straightforward. Less than 1MB per day of data is expected.
Origin	Re-use: DWD Open data portal.
Use beyond project	In any field where Meteorology may play a role.

- Weather radar:

Description	Radar based precipitation estimates derived from OPERA European network.
Purpose	To be used as an input for ARGOS implementations (Task 2.1).

Type and Format	NetCDF.
Processing and Size	New measurements may be available every 15 minutes. Around 250MB per day of data is expected.
Origin	EUMETSAT
Use beyond project	OPERA data cannot be shared outside project.

2.1.2 Data Sets

The following data sets have been identified to play a role in the Weather Emergency use case:

- Weather stations measurements:

Description	Data from the Brückstraße (inner City Dortmund) weather station for 2009 - 2022 at 10-minute intervals. Measurements of temperature, wind and precipitation registered by the station.
Purpose	Basis for machine learning (e.g., FMI), finding correlations between weather (1st rain, later maybe heatwaves) and number of emergency operations (all forms: ambulances, technical response, fire, etc.); Output: forecasts of impact caused by different levels of extreme weather
Type / Format	Excel-List; *.xlsx & GeoJSON
Size	3 KB
Origin	Explore — Open Data Dortmund
Use beyond project	For all activities where weather-related impacts are investigated and analyzed for the City of Dortmund.

- Sewer network: nodes, edges, pump locations, flow direction:

Description	Canal geodata refers to localized information specifically concerning waterways and canals. The data contains details about lids, shafts, the type of water flowing through (mixed water, rainwater, and/or wastewater) and the flow direction.
Purpose	For the training and development of machine-learning-based tools. Insight into flow direction and the composition of water (mixed water, rainwater, wastewater) assists in anticipating areas prone to flooding during intense rainfall occurrences.
Type / Format	Spatial Data; *.shp & Excel-List; *xlsx
Size	Spatial Data; *.shp & Excel-List; *xlsx
Origin	Urban drainage office Dortmund

Use beyond project	Distribution outside of the project is not intended

- Digital Elevation Model, Elevation model, terrain slope:

Description	Digital elevation models (DEMs) are digital representations of the topography of the earth's surface.
Purpose	For the training and development of machine-learning-based tools. DEMs provide elevation data that helps in simulating and predicting flood extents and depths during extreme weather events such as heavy rainfall or storm surges.
Type / Format	Tag Image File Format; *.tif
Size	1 GB
Origin	Digitales Geländemodell - Gitterweite 1 m (nrw.de) Cadastral office Dortmund; Not yet available TIM-online (nrw.de) Geobasisdaten (nrw.de)
Use beyond project	DEMs can assist in assessing the extent of damage caused by weather-related events. By comparing pre- and post-event DEMs, authorities can quantify changes in elevation, identify areas affected by flooding or erosion.

- Traffic network (especially road network): Nodes, edges: means of mass transportation networks, underground parking spaces:

Description	Traffic network data includes information about the nodes (intersections, junctions) and edges (road sections) of a city's road network. This data is normally presented in the form of lines,but can be further refined to include additional information such as road widths.
Purpose	Traffic network data provides information about road infrastructure, including intersections and road sections. During weather emergencies such as floods or storms, this data helps emergency responders identify alternative routes, avoid flooded or blocked roads and plan efficient evacuation routes to ensure public safety.
Type / Format	Spatial Data; *.shp
Size	300 MB
Origin	Relation: Dortmund (1829065) | OpenStreetMap
Use beyond project	By announcing information through traffic management systems, websites and mobile applications, authorities can help citizens to avoid hazards.

- Rivers and streams data:

Description	River and stream data provide information on the location of these watercourses.
Purpose	River and stream data help assess the potential for flooding by providing information on watercourse locations and their proximity to populated areas. This data aids in identifying areas prone to flash floods or river overflow during heavy rainfall events.
Type / Format	Spatial Data; *.shp
Size	30 MB
Origin	Relation: Dortmund (1829065) | OpenStreetMap GEOportal.NRW
Use beyond project	Emergency managers can use this information to coordinate response efforts to mitigate the impact of riverine flooding on communities and infrastructure.

- Development: WebODM (3D model, ortophoto):

Description	This data includes high-resolution aerial imagery, digital surface models (DSM), digital terrain models (DTM) and 3D point clouds. They are typically obtained by processing aerial images or drone images and provide detailed insights into the terrain and surface characteristics of specific areas. The 3D modelling of the city includes data on the buildings and their storeys and heights. Orthoimagery is the use of aerial or satellite imagery to represent a geographic area. These data are processed in such a way that they provide a geometrically correct representation of the Earth's surface by correcting for distortions due to the characteristics of the terrain and the movements of aircraft.
Purpose	This data provides detailed insights into terrain characteristics, surface features, and building structures, allowing for comprehensive risk assessment and vulnerability mapping.
Type / Format	CityGML & Tag Image File Format; *.tif
Size	45 GB; 2 GB & 5 GB
Origin	Cadastral office Dortmund; not yet available TIM-online (nrw.de) Geobasisdaten (nrw.de)

Use beyond project	By visualizing terrain and surface characteristics, emergency managers can prioritize response efforts, allocate resources effectively, and coordinate rescue operations in affected areas, ensuring a swift and coordinated response to weather-related emergencies.

- Type of development, critical infrastructure, vulnerable population:

Description	This data includes locations such as hospitals, schools and retirement homes, stadiums, museums.
Purpose	Data pinpointing critical infrastructure such as hospitals, schools, and retirement homes enables emergency responders to prioritize resource allocation and coordinate rescue and evacuation efforts effectively.
Type / Format	Spatial Data; *.shp & GeoJSON
Size	100 KB
Origin	Explore — Open Data Dortmund
Use beyond project	Emergency managers can use this data to develop response plans, establish evacuation routes, and allocate resources effectively to areas most at risk during weather-related emergencies.

- Floodplains / Flood hazard map:

Description	Floodplains provide information on geographical areas that flood during heavy rainfall or flooding.
Purpose	Floodplain data provides critical information about areas prone to flooding during heavy rainfall or flooding events. By mapping floodplains, emergency planners can assess the extent of flood risk, identify vulnerable communities, and prioritize resources for flood preparedness and mitigation efforts.
Type / Format	Spatial Data; *.shp & Tag Image File Format; *.tif
Size	400 MB – 5 GB
Origin	GEOportal.NRW Starkregengefahrenkarte TN100 (digistadtdo.de)
Use beyond project	By mapping floodplain boundaries and regulating development in flood-prone areas, authorities can reduce the risk of flood damage and promote resilient.

- Emergency Cases FDDO 2021-2023:

Description	List of operations of Fire Department of Dortmund (FDDO) from 01/2021 to 12/2023 (around 25.000 incidents); Information

	includes time (down to minutes), which units are responding to the emergency, location of emergency (address and coordinates), age of person involved… Level of detail available to partners is yet to be determined.
Purpose	Basis for machine learning (e.g. FMI), finding correlations between weather (1st rain, later maybe heatwaves) and number of emergency operations (all forms: ambulances, technical response, fire, etc.); Output: forecasts of impact caused by different levels of extreme weather
Type / Format	Excel-List; *.xlsx
Size	Around 20MB for one year
Origin	Fire Department Dortmund Re-used: For every operation the responders have to fill in a report which adds to information, which is generated automatically by the emergency call system / operations control system; key facts for each incident are listed in a database of operations from where the dataset is exported by request
Use beyond project	Distribution outside of the project is not intended

- Object plans:

Description	Building plans contain information about the structure, layout and features of a building. These plans show evacuation areas, escape routes, the location of emergency exits, fire-fighting equipment, emergency supply points and other important facilities in the building.
Purpose	Building plans provide essential information for emergency preparedness by detailing evacuation areas, escape routes, emergency exits, and the location of critical facilities such as fire-fighting equipment and emergency supply points.
Type / Format	PDF-Data; *.pdf
Size	7 KB
Origin	Fire Department Dortmund
Use beyond project	Distribution outside of the project is not intended.

- Traffic over time and Parking garages:

Description	Data on the parking time of cars in car parks over a certain period of time.
Purpose	Understanding parking patterns during different weather conditions can help emergency responders anticipate traffic congestion and plan accordingly. For example, during severe weather events, knowing typical parking durations can aid in estimating the number of vehicles likely to be in a given area.
Type / Format	Spatial Data; *.shp & Excel-List; *xlsx & GeoJSON
Size	19 KB
Origin	Parkhäuser und Parkplätze — Open Data Dortmund
Use beyond project	Analyzing parking time data can provide insights into patterns of vehicle usage and parking behavior following different events.

- Vital Data Firefighters:

Description	The data is based on monitoring the heart rate of individuals during a simulation of a real-life scenario.
Purpose	For the training and development of machine-learning-based tools Monitoring heart rate during simulations helps ensure the safety and well-being of individuals involved in emergency response exercises. By tracking vital signs, emergency responders can promptly identify signs of distress among participants and take appropriate action to prevent injuries or health complications.
Type / Format	Excel-List; *xlsx
Size	23 KB
Origin	Fire Department Dortmund; not yet available
Use beyond project	Distribution outside of the project is not intended

- Damage clearance tasks in Finland

Description	Daily, weather-related damage clearance tasks performed by regional rescue departments aggregated on a county level in Finland. Case aggregation: 1-4 daily cases shown as 0. Case amounts of 5 and above are shown without aggregation. Dates with zero cases do not appear in the search results. Dataset is available from 2001 onwards and updated irregularly, approximately once a year covering the data from the previous calendar year.
Purpose	For the training and development of machine-learning-based tools in Finland or in other use cases and locations. (Task 2.1)

Type / Format	Different possibilities, for example, csv, JSON
Size	~10-20MB
Origin	Re-used from: Emergency Services Academy / Ministry of Interior Finland. Aggregation by FMI
Use beyond project	For all activities where weather-related impacts are investigated and analyzed.

- Warnings for flooding in Dortmund:

Description	ARGOS platform for Dortmund will issue flood warnings for the city and its vulnerable elements, based on different meteorological inputs including some from specific algorithms.
Purpose	Output of the ARGOS platform
Type / Format	It will be text data in csv format, not exactly defined yet.
Size	Only in heavy rain episodes, less than 1MB per day is expected.
Origin	Generated within the project by the ARGOS platform itself
Use beyond project	To anyone studies heavy rain episodes from the point of view of Civil Protection

- User data in ARGOS platform:

Description	User in ARGOS platform for the implementations of Dortmund, Innsbruck and Helsinki which include personal data as full name, e-mail and possibly phone number.
Purpose	Actual use of the ARGOS platform
Type / Format	Text entries in a database
Size	Less than 1MB is expected.
Origin	Generated within the project by the ARGOS platform
Use beyond project	None.

- Interviews within the Weather Emergency Use Case:

Description	Interviews conducted to explore needs and requirements of the stakeholder regarding the CREXDATA technologies
Purpose	Used to collect customer and system requirements for the development of CREXDATA technologies and the overall system
Type and Format	Protocols saved as PDF files on the CREXDATA Microsoft Teams

Processing and Size	~ 50-100 MB
Origin	Participants of the interviews, stakeholder of the CREXDATA weather emergency use case
Use beyond project	None.

- Trials and evaluation activities within the Weather Emergency Use Case:

Description	Conducting trials and evaluation activities as part of the Weather Emergency Use Case to verify and validate project results and developments
Purpose	Requirements analysis for the CREXDATA technologies and evaluation of the project results together with relevant stakeholders
Type and Format	written protocols, completed surveys (PDF, *.docx, *.xml; saved on the CREXDATA Microsoft Teams) and video recordings of the trials for documentation purpose (mp4, saved on a local UPB server)
Processing and Size	Protocols and surveys: < 50 MB video recordings (only documentation purpose): ~17,94 GB
Origin	participants of the trials; video taken by observation cameras
Use beyond project	None.

- Surveys of T6.2:

Description	Survey results collected as part of the exploitation strategy
Purpose	To be used as an input for exploitation and business planning. (Task 6.2).
Type and Format	Alphanumeric characters and dependencies between them
Processing and Size	No data processing only data collection via https://umfragen.uni-paderborn.de/index.php/surveyAdministration/ and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

- Presence and individual Workshops:

Description	Workshop results collected as part of a joint collaboration on the exploitation strategy
Purpose	To be used as an input for exploitation and business planning. (Task 6.2).
Type and Format	Alphanumeric characters, pictures, posters and dependencies between them
Processing and Size	No data processing only data consolidation and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

- Social media post of floodings and other disaster events (CrisisLexT6):

Description	A collection of English tweets across 6 large events in 2012 and 2013, with about 10,000 tweets labeled by relatedness (as "on-topic", or "off-topic") with each event. Events included, hurricane, floods, bombing, tornado, and explosion. (only floods are used). Data sampled from twitter (X) by keywords and geographical regions or coordinates.
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Text in comma-separated values (.csv) format.
Size	3.1MB (Zipped)
Origin	Re used from: Data collected for public research https://www.crisislex.org/data-collections.html#CrisisLexT6
Use beyond project	Not intended for use outside the project, but publicly available for research

- Social media post of floodings and other disaster events (CrisisLexT26):

Description	A multilingual collection of tweets collected during 26 large crisis events in 2012 and 2013, with about 1,000 tweets labeled per crisis for informativeness (i.e. “informative," or "not informative"), information type, and source. ~28,000 labeled tweets, sampled by keywords. Events included, floods, earthquakes, fires, explosions, typhoons, meteor, bombings, building collapse, haze, train crash, helicopter crash, shootings.
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Text in comma-separated values (.csv) format

Size	4.6MB (Zipped)
Origin	Re used from: Data collected for public research https://www.crisislex.org/data-collections.html#CrisisLexT26
Use beyond project	Not intended for use outside the project, but publicly available for research

- Social media post of floodings and other disaster events (CrisisLexMMD):

Description	A collection of tweets and images collected during seven major natural disasters including earthquakes, hurricanes, wildfires, and floods that happened in the year 2017 across different parts of the World. Annotated by relevance, damage severity, humanitarian categories. (only made use of textual data)
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Text as raw tweets in json and annotated tweets in tsv.
Size	Approximately 1.8GB, but lesser if using only textual data
Origin	Re used from: Data collected for public research https://crisisnlp.qcri.org/crisismmd
Use beyond project	Not intended for use outside the project, but publicly available for research

- Social media post of floodings and other disaster events (CrisisNLP):

Description	A multilingual collection of tweets collected during ~19 major crisis events including, earthquakes, cyclone, typhoon, hurricane, volcano, floods, epidemics, landslides, etc. in 2013-2015
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Text in tab-separated values (tsv) and comma-separated values (csv) formats
Size	Approximately 2.5GB
Origin	Re used from: Data collected for public research https://crisisnlp.qcri.org/lrec2016/lrec2016.html
Use beyond project	Not intended for use outside the project, but publicly available for research

- Social media post of floodings and other disaster events (HumanAID):

Description	A collection of manually annotated tweets that has been collected during 19 major natural disaster events including earthquakes, hurricanes, wildfires, and floods, which happened from 2016 to 2019 across different parts of the World.
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Text in tab-separated values (tsv) format
Size	~3MB (Zipped)
Origin	Re used from: Data collected for public research https://crisisnlp.qcri.org/humaid_dataset
Use beyond project	Not intended for use outside the project, but publicly available for research

- Disaster response messages:

Description	A multilingual collection of 30,000 messages drawn from events including an earthquake in Haiti in 2010, an earthquake in Chile in 2010, floods in Pakistan in 2010, superstorm Sandy in the U.S.A. in 2012, and news articles spanning a large number of years and 100s of different disasters.
Purpose	For the training of language models for relevance and event type prediction in social media posts. (T4.5)
Type / Format	Apache Arrow (.arrow) format
Size	7.2 MB
Origin	Re used from: Data collected for public research https://huggingface.co/datasets/disaster_response_messages
Use beyond project	Not intended for use outside the project, but publicly available for research

- Disaster data streams:

Description	Multi-stream datasets from several disasters, covering Twitter, Reddit, Facebook, and online news sources gathered from the NELA News Collection.
Purpose	For evaluating the modules on question answering and retrieval. (T4.5)
Type / Format	.json files
Size	3.6 GB
Origin	Re used from: Publicly available through https://crisisfacts.github.io/

Use beyond project	Not intended for use outside the project, but publicly available for research

2.2 FAIR Weather Emergencies Data

Data management will be guided by FAIR principles in this use case. However, many data sources belong to institutions with its own Open Data policy; so, in these cases, data is made Findable, Accessible, Interoperable and Re-usables (FAIR) by the origin institution.

2.2.1 Data Sources

- Local environment data sources: At this stage, the use and management of this sources is still on definition.
- Global environment data sources:
  - EFAS:
    EFAS provides real-time data to its members and near-real time data can be accessed through a public viewer (https://www.efas.eu/efas_frontend/#/home).

  - EFFIS:
    EFFIS provides real-time data to its members and near-real time data can be accessed through a public viewer (https://effis.jrc.ec.europa.eu/apps/effis_current_situation/).

  - ECMWF forecasts:
    A subset of ECWMF real-time forecast data is made available to the public free of charge. Their use is governed by the Creative Commons CC-4.0-BY license and the ECWMF Terms of use. This means that the data may be redistributed and used subject to appropriate attribution (https://www.ecmwf.int/en/forecasts/datasets/open-data).

  - Weather stations measurements in Germany:
    They are available through the DWD Open Data Portal. A searcher can be used for data discovery. No registration is needed to access the data.

  - Weather radar:
    OPERA network estimation will be used through a specific agreement between EUMETSAT and CREXDATA and for inner-project purposes only, otherwise a commercial license is needed.

2.2.2 Data Sets

- Emergency Cases FDDO 2020-2021:
  Distribution of this data outside of the project is not intended because it is sensitive data in terms of security.

- Sewer network: nodes, edges, pump locations, flow direction:

Distribution of this data outside of the project is not intended because it is sensitive data in terms of security.

- Object plans:

Distribution of this data outside of the project is not intended because it is sensitive data in terms of security.

- Vital Data Firefighters:
  Distribution of this data outside of the project is not intended because it is sensitive data in terms of security.

- Sensors (weather): Type, position:

Findable	Findable using tags: Wetter, Klima, Wetterstation, Temperatur, Luft
Accessible	Datasets area already available at Open Data Dortmund: Explore — Open Data Dortmund
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Digital Elevation Mode, Elevation model, terrain slope

Findable	Findable by using tags: Geländeneigung, Geländestufen, Höhenschichten, Oberflächenmodell
Accessible	Datasets area already available at the federal level: TIM-online (nrw.de) Geobasisdaten (nrw.de) Digitales Geländemodell - Gitterweite 1 m (nrw.de)
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Traffic network (especially road network): Nodes, edge

Findable	Findable by using the location for Dortmund
Accessible	Datasets area already available at OpenStreetMap: Relation: Dortmund (1829065) | OpenStreetMap

Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Rivers & streams

Findable	Findable by using the location for Dortmund
Accessible	Datasets area already available at OpenStreetMap: Relation: Dortmund (1829065) | OpenStreetMap
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Development: WebODM (3D mode, orthophoto)

Findable	Findable by using tags: 3D-Gebäudemodelle, Orthophoto, Luftbilder, Luftaufnahmen
Accessible	Datasets area already available at the federal level:Geobasisdaten (nrw.de)
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Type of development, critical infrastructure, vulnerable population

Findable	Findable by using tags: Orte von Interesse
Accessible	Datasets area already available at Open Data Dortmund: Explore — Open Data Dortmund
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Floodplains/ Flood hazard map

Findable	Findable by using tags: Wasser, Hochwasser
Accessible	Datasets area already available at the federal level: GEOportal.NRW
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Traffic over time: Parking garages

Findable	Findable by using tags: Parken, Parkhaus, Parkplätze
Accessible	Datasets area already available at Open Data Dortmund: Parkhäuser und Parkplätze — Open Data Dortmund
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Damage clearance tasks in Finland

Findable	Details of available existing weather-related impact datasets: http://openwms.fmi.fi/geoserver/silva/wfs?service=wfs&version=2.0.0&request=GetCapabilities Saved queries: http://openwms.fmi.fi/geoserver/silva/wfs?service=wfs&version=2.0.0&request=DescribeStoredQueries
Accessible	Aggregated data is openly available through FMI open data webpages: https://en.ilmatieteenlaitos.fi/open-data-sets-available
Interoperable	Standardized format
Re-usable	The dataset can be used for example for the development of impact-based weather warnings or EWS tools. Aggregated data is openly available through FMI open data webpages and is updated yearly.

- Warnings for flooding in Dortmund:
  FAIR provisions for this dataset are not defined yet.

- Social media post of floodings and other disaster events (CrisisLexT6):

Findable	
Accessible	Dataset is freely available at: https://www.crisislex.org/ https://www.crisislex.org/data-collections.html#CrisisLexT6
Interoperable	Datasets can be freely accessed and are stored in an open format following the CC BY 4.0 guidelines.
Re-usable	License: CC BY 4.0

- Social media post of floodings and other disaster events (CrisisLexT26):

Findable	
Accessible	Dataset is freely available at: https://www.crisislex.org/ https://www.crisislex.org/data-collections.html#CrisisLexT26
Interoperable	Datasets can be freely accessed and are stored in an open format following the CC BY 4.0 guidelines.
Re-usable	License: CC BY 4.0

- Social media post of floodings and other disaster events (CrisisLexMMD):

Findable	
Accessible	Dataset is freely available at: https://crisisnlp.qcri.org/crisismmd
Interoperable	Datasets can be freely accessed and are stored in an open format following the terms of use at https://crisisnlp.qcri.org/terms-of-use.html
Re-usable	Data usage terms available: https://crisisnlp.qcri.org/terms-of-use.html

- Social media post of floodings and other disaster events (CrisisNLP):

Findable	
Accessible	Dataset is freely available at: https://crisisnlp.qcri.org/lrec2016/lrec2016.html
Interoperable	Datasets can be freely accessed and are stored in an open format following the terms of use at https://crisisnlp.qcri.org/terms-of-use.html
Re-usable	Data usage terms available:

	https://crisisnlp.qcri.org/terms-of-use.html

- Social media post of floodings and other disaster events (HumanAID):

Findable	
Accessible	Dataset is freely available at: https://crisisnlp.qcri.org/humaid_dataset
Interoperable	Datasets can be freely accessed and are stored in an open format following the terms of use at https://crisisnlp.qcri.org/terms-of-use.html
Re-usable	Data usage terms available: https://crisisnlp.qcri.org/terms-of-use.html

- Disaster response messages:

Findable	
Accessible	Dataset is freely available at: https://huggingface.co/datasets/disaster_response_messages
Interoperable	Datasets can be freely accessed and are stored in an open format.
Re-usable	

- Disaster data streams:

Findable	
Accessible	Dataset is freely available at: https://crisisfacts.github.io/#datasets http://trecis.org https://github.com/crisisfacts/utilities/blob/main/00-Data/00-CrisisFACTS.Downloader.ipynb
Interoperable	Datasets can be freely accessed and are stored in an open format.
Re-usable	

2.3 Data Security in the Weather Emergencies Use Case

Many data sources and datasets used in the Weather Emergencies use case belong to third parties so they are stored and protected according to their own protocols. Others are generated within the project and security measures are in place. Find them below for each component:

ARGOS platform

Currently, ARGOS is hosted on the AWS Cloud from Amazon and is replicated across two Availability Zones to ensure high availability and reliability. AWS has implemented several measures to ensure the redundancy and security of its cloud infrastructure.

Users can securely access the System from any location with internet access using a modern web browser. The connection established between the user's browser and the System servers is secured using HTTPS. HTTPS uses Transport Layer Security (TLS) as the cryptographic protocol to ensure end-to-end communications between two computer networks (following best practices and security recommendations, only TLS versions 1.2 and 1.3 are used).

In data exchange with providers, usually providers offer APIs that support HTTPS, ensuring secure transmission of data in transit. Alternatively, ARGOS offers an SFTP server that utilizes SSH as its transport mechanism creating a secure tunnel through which the data is delivered securely.

Data that is stored. but not actively accessed or transmitted, is protected by encryption providing and additional layer of protection against unauthorized access or data breaches.

Database backups are generated regularly through automated processes (at least every 12 hours) to ensure data integrity and facilitate disaster recovery. These backups are encrypted at the time of generation using AES-256 both on-site and off-site. For off-site storage, an Amazon S3 bucket is used.

The System's web viewer is only accessible to authenticated users using a username and password. The credentials are stored in a database following security best practices and recommendations. The passwords are never stored in plaintext and in its place a salted hash is used. The hashing is performed using the PBKDF2 key derivation function and HMAC-SHA-256 hash function with a work factor of 600 thousand iterations following the current OWASP (Open Worldwide Application Security Project) recommendations as of 2023.

The System complies with GDPR informing the users of the management of protected data and asking for their consent. At the login page, unless previously accepted, a pop-up window will require the acceptance of the use of cookies before entering the system. Cookies are needed to provide the basic functionalities of the System such as authentication. Also, on the login page and inside the system, the Privacy Policy and Terms of Use are linked, including all the details required by regulations. The data transfer and storage are done inside the AWS Data Centres located in Europe. AWS comply with GDPR and other data protection regulations. For further details on the GDPR compliance of AWS services please visit: https://aws.amazon.com/compliance/gdpr-center

On top, HYDS (ARGOS developer) comply with the GDPR, having support from an external specialized company that periodically assesses the company processes. The System resides physically in the Paris Region and therefore the data centres are located within the EEA (European Economic Area) and comply with GDPR.

CREXDATA's SharePoint

The data on CREXDATA's SharePoint is secure because it is stored on the servers-based Paderborn University licenses. These servers have their own backup protocol, which ensures regular backups to prevent data loss. Access to the data is also strictly limited to authorized users. Access to the CREXDATA Sharepoint on Microsoft Teams is managed by the UPB team. Only members of the CREXDATA project have access. Critical datasets that

require their own data usage agreement can also be made accessible to only relevant team members in restricted folders.

3 Health Crisis Data and Management Procedures

To control and manage a worldwide crisis such as the COVID-19 pandemic, we require the resolution of very complex and heterogeneous problems including, but not restricted to, the detection of outbreaks and the tracking of the COVID-19 pandemic; the design and evaluation of effective pharmaceutical and non-pharmaceutical interventions; the study of viral evolution for the detection of emerging mutations of clinical relevance; and the prediction of potential drug targets as well as the development and repurposing of novel therapeutic strategies. All these problems require high-quality, massive, real-time data.

Mechanistic modelling and epidemiological simulations are necessary tools to understand the dynamics of complex systems of a very different nature, to forecast their evolution in different scenarios and the interventions needed to have the desired outcome. In this project, we will consider two scenarios:

- Scenario 1: Epidemiological models have been developed that allow us to follow the evolution of viral transmission by providing fairly accurate quantitative predictions. This development is due to an increase in computing power; the availability of a large amount of data that characterize, at a very high spatio-temporal resolution, the socio-economic activity of regions and countries; population mobility; and the developments of mathematical models of epidemic spread.
- Scenario 2: Mechanistic multiscale models will be used to build a toolbox aimed at having a digital twin for the treatment of patients. Multiscale models are being widely used for their capacity of bridging many temporal and spatial scales from the cell-level, down to proteins and genes, to study cancer and SARS-CoV-2 infection. ***CREXDATA*** will expand them from cell-level up to the organ-level to build a multiscale, multicellular, spatiotemporal model of lung tissue infected by the SARS-CoV-2 virus.

3.1 Health Data Summary

The following data sets have been identified to play a role in the use case:

- COVID-19 Flow-Maps:

Description	COVID-19 Flow-Maps, a cross-referenced Geographic Information System that integrates regularly updated time-series accounting for population mobility and daily reports of COVID-19 cases in Spain at different scales of time spatial resolution.
Purpose	Calibrate model parameters and validate simulations for the COVID-19 epidemic spread in Spain
Type / Format	Apache Parquet, geoJSON, CSV, NetCDF
Size	Daily Mobility 8G COVID-19 Daily Reports 250MB Geographical Layers 250MB
Origin	Re-used: Mobility data comes from a MITMA study:

	https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data COVID-19 cases comes from Several different sources, check: https://github.com/bsc-flowmaps/fm-covid-19 Geographical Layers comes from Several different sources, check: https://github.com/bsc-flowmaps/fm-layers
Use beyond project	Any researcher interested in the study human mobility, epidemics and their interplay.

- COVID-19 reports from the ISCIII:

Description	This dataset includes the number of cases, number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence for all Spain from 2020-01-01 until 2022-03-27. The reports were obtained from the declaration of COVID-19 cases to the National Epidemiological Surveillance Network (RENAVE) through the computer platform via the SiViES Web (Spanish Surveillance System) managed by the National Center for Epidemiology (CNE). This information came from the epidemiological case survey that each Autonomous Community completed when a case of COVID-19 was identified.
Purpose	Calibrate model parameters and validate simulations for the COVID-19 epidemic spread in Spain
Type / Format	CSV, NetCDF
Size	4MB
Origin	COVID-19 cases, hospitalization and defunction by sex and age comes from the ISCIII oficial reports: https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos
Use beyond project	Any researcher interested in the study the COVID-19 pandemic or epidemics in general.

All these data records were published in “COVID-19 Flow-Maps an open geographic information system on COVID-19 and human mobility for Spain” [link].

- Epidemic Simulations:

Description	This dataset consists of simulated spatio-temoral time series of spread of an infectious disease across different regions
Purpose	The datasets are the output of the epidemic simulations and are used to evaluate/validate the forecast of the simulations and also

	to investigate *in-silico* scenarios that can help mitigate the spread of a new emergent infectious disease.
Type / Format	NetCDF file format, or similar for multidimensional data arrays.
Processing and Size	A single simulation of six months takes ~30 seconds and produces an output of ~150MB. During the parameter calibration or model exploration [200-1000] can be evaluated in parallel. The full output of a simulation for 1 year is ~250MB.
Origin	Generated within the project. Data is generated by the simulation engine which include the Julia Package EpiSim.jl: https://github.com/Epi-Sim/EpiSim.jl (simulator) https://github.com/Epi-Sim/MMCACovid19Vac.jl (vaccination model engine) https://github.com/Epi-Sim/MMCACovid19Vac.jl (Model Exploration WorFlow)
Use beyond project	In a real scenario, simulated data could potentially be used as a forecast or as source of counterfactual to be used by policy makers.

- Anonymized patient omics molecular data:

Description	Publicly available pseudo-anonymized raw experimental data from patients.
Purpose	This dataset bundles different studies that will be the input used to analyze and personalize our multiscale models. It potentially consists of transcriptomics, genomics, copy number variations and proteomics data.
Type / Format	Mostly CSV, but also some in TSV and JSON
Processing and Size	Static data of 1 TB in total
Origin	Re-used from several publications
Use beyond project	Any researcher interested in the study of lung cancer.

- Anonymized patient pulmonary 3D positional data:

Description	Publicly available pseudo-anonymized image data from patients.
Purpose	This dataset bundles different studies that will be the input used to have complex 3Dal setups for our multiscale models. Once analyzed, this dataset will have positional data for each of the alveoli of the patients.
Type / Format	SVG and CSV

Processing and Size	Static data of 200 GBs
Origin	Re-used from several publications
Use beyond project	Any researcher interested in studying lung shapes and sizes or interested in using the positional data to setup their simulations.

- Simulated multiscale infection data:

Description	Model forecasts of drug and ventilation mechanisms that explain different alveolus and immune cells dynamics in COVID19 patient data
Purpose	This dataset will be outputs of the health crisis use case. Both for the ensemble runs of the model's parameter exploration and for the hero runs of the long simulations.
Type / Format	CSV, TXT and XML
Processing and Size	~100 GB/min for ensembles, ~4.5 GB/min for single simulation, potentially totalizing many TBs
Origin	Generated within the project via Simulation of PhysiBoSS and Alya in MareNostrum4
Use beyond project	To perform summary statistics of the parameter exploration methods and, potentially, as training data for training AI methods.

- Surveys of T6.2:

Description	Survey results collected as part of the exploitation strategy
Purpose	To be used as an input for exploitation and business planning (Task 6.2).
Type and Format	Alphanumeric characters and dependencies between them
Processing and Size	No data processing only data collection via https://umfragen.uni-paderborn.de/index.php/surveyAdministration/ and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

- Presence and individual Workshops:

Description	Workshop results collected as part of a joint collaboration on the exploitation strategy

Purpose	To be used as an input for exploitation and business planning. (Task 6.2).
Type and Format	Alphanumeric characters, pictures, posters and dependencies between them
Processing and Size	No data processing only data consolidation and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

3.2 FAIR Health Crisis Data

The following approaches have been set regarding the FAIR principles in the Health Crisis use case:

- COVID-19 Flow-Maps:

Findable	Machine-accessible metadata file describing the reported data: https://doi.org/10.6084/m9.figshare.15198123
Accessible	Datasets area already available at Zenodo and GitHub https://zenodo.org/communities/flow-maps/?page=1&size=20 https://github.com/bsc-flowmaps/
Interoperable	Datasets can be freely accessed and are stored in an open format
Re-usable	By using this data to train/calibrate and validate models and simulations of the COVID-19 pandemic in Spain. Datasets can also be used for retrospective and counter factual analysis

- Epidemic Simulations:

Findable	We expect to provide Machine-accessible metadata.
Accessible	Simulated Datasets associated with publications will be available through Zenodo or similar resources, , together with the code and all the input files required to reproduce the simulated data.
Interoperable	Simulated Datasets are reported in the NetCDF open format
Re-usable	Simulated data can be used to benchmark our simulator against other approaches

- Anonymized patient omics molecular data:

Findable	Each dataset has its own findable metadata as per EGA and GEO requirement.

Accessible	Datasets available at GEO (https://www.ncbi.nlm.nih.gov/geo/) and EGA (https://ega-archive.org/).
Interoperable	Datasets can be freely accessed and are stored in an open format.
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Anonymized patient pulmonary 3D positional data:

Findable	Each dataset has its own findable metadata as per EGA and GEO requirement.
Accessible	Datasets available at GEO (https://www.ncbi.nlm.nih.gov/geo/) and EGA (https://ega-archive.org/).
Interoperable	Datasets can be freely accessed and are stored in an open format.
Re-usable	By accessing it from these publicly available repositories, by bundling them with other studies to perform summary studies, by personalizing models with them.

- Simulated multiscale infection data:

Findable	We will provide machine-accessible metadata.
Accessible	Simulated datasets associated with publications will be available at Zenodo.
Interoperable	We will use open formats and comprehensive metadata.
Re-usable	The dataset will be available for benchmarking other tools and other exploration methods, such as AI-based ones.

3.3 Data Security in the Health Crisis Use Case

All data and simulation outputs are stored and processed on the MareNostrum 5 supercomputing system, which provides high-performance data storage capabilities and ensures data integrity throughout computational tasks. MareNostrum 5 offers both robust storage capacity and a reliable backup infrastructure, including regular snapshots and redundancy mechanisms, minimizing the risk of data loss. Access to the data is secured by MareNostrum's multi-layered security protocols, which include authentication and access controls to ensure that only authorized personnel can view or modify files. In addition, all sensitive data are stored in encrypted directories, complying with established data protection and privacy regulations.

4 Maritime Data and Management Procedures

The purpose of this use case is to develop the first solution, combining hardware and software development, that will be able to use data coming from a vessel's VDR, which will be fused with global views of data creating reliable digital twins of vessels. It will develop the first weather and emergency routing and route forecasting solutions that will be performed for all vessels of a fleet simultaneously (instead of on-demand requests per vessels) and that will rely on big data and AI technologies.

- ***Scenario 1:*** First, we will develop an IoT device, named IoT-Voyage Data Streamer (VDS) that will be deployed on board vessels and will be able to connect to the VDR (Voyage Data Recorder) of a vessel, which can be perceived as a vessel's black box.
- ***Scenario 2:*** Since (i) we cannot deploy this device in more than one vessel, (ii) maritime events of extreme hazard are sparse, and (iii) extreme weather conditions are also relatively rare, we will create streams of simulated vessel positions, events, and weather conditions in order to simulate and optimize the navigation of a vessel under certain conditions (e.g., traffic, weather conditions). We will use the route forecast data of a vessel, based on historical information, and the real-time kinematic features of its motion, and we will fuse this data with the real-time IoT-VDS data (if it was installed on board the vessel) in order to detect critical events.

4.1 Maritime Data Summary

The following data sets have been identified to be relevant in the use case:

- Single Ground Based AIS Receiver Vessel Tracking Dataset:

Description	This dataset contains all decoded messages collected within a 24h period (starting from 29/02/2020 10PM UTC) from a single receiver located near the port of Piraeus (Greece). All vessels' identifiers such as IMO and MMSI have been anonymized and no down-sampling procedure, filtering or cleaning has been applied.
Purpose	AIS uncompressed data can be used to train predictive traffic models at a fine-grained temporal resolution at a local scale, as captured from a single receiver.
Type / Format	Comma-Separated Values (CSV)
Size	8.2 MB
Origin	Re-used: AIS data from a single receiver located near the port of Piraeus (Greece) / MarineTraffic (Kpler)
Use beyond project	Research, Academia; Defense; Industry - Maritime Intelligence; Industry - Big Data; Industry – AI; Industry – Software Engineering; Industry – Supply Chain.

- Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance:

Description	This dataset contains ships' information collected through the Automatic Identification System, integrated with a set of complementary data having spatial and temporal dimensions aligned. The dataset contains four categories of data: Navigation data, vessel-oriented data, geographic data, and environmental data. It covers a time span of six months, from October 1st, 2015 to March 31st, 2016 and provides ships positions within the Celtic sea, the Channel and Bay of Biscay (France). The dataset is proposed with predefined integration and querying principles for relational databases. These rely on the widespread and free relational database management system PostgreSQL, with the adjunction of the PostGIS extension, for the treatment of all spatial features proposed in the dataset.
Purpose	The dataset is realistic and comprehensive, representative of operational information needs in different Maritime Intelligence, Surveillance, and Reconnaissance scenarios, including safety and security of navigation and preservation of protected areas. It will suit various researches and potential applications and reuses, including: Maritime traffic analysis: The AIS data volume is large enough to support the development of sea traffic models to characterize the maritime traffic and for estimating other patterns of life. Impact assessment of human activities at sea: The dataset may provide a valuable support for understanding the impact of human activities at sea, facilitating the development of sustainable fishery policies (e.g., by developing fishing pressure models1). Environmental assessment: It might be beneficial to ocean dynamics estimation study, to assess the quality of specific ocean state variables (e.g., currents). Assessment and benchmarking of multi-source information fusion algorithms: The dataset gathers highly heterogeneous data, mixing raw and processed information, archival data and streaming data. Since these heterogeneous data are aligned in time and space, they are a perfect test bed to experiment multi-source information fusion algorithms.
Type / Format	Comma-Separated Values (CSV); ESRI Shapefiles
Size	1.8 GB
Origin	Re-used. Areas covered: Celtic sea, the Channel and Bay of Biscay (France),

	Generated by: Naval Academy Research Institute, Brest, France; Arts et Metiers ParisTech, France; NATO STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy
Use beyond project	Research, Academia; Defense; Industry - Maritime Intelligence; Industry - Big Data; Industry – AI; Industry – Software Engineering; Industry – Supply Chain.

- The Piraeus AIS Dataset for Large-scale Maritime Data Analytics:

Description	The AIS dataset comes along with spatially and temporally correlated data about the vessels and the area of interest, including weather information. It covers a time span of over 2.5 years, from May 9th, 2017 to December 26th, 2019 and provides anonymized vessel positions within the wider area of the port of Piraeus (Greece), one of the busiest ports in Europe and worldwide. The dataset consists of over 244 million AIS records, an average of more than 10,000 records per hour, which makes it an ideal input for large-scale mobility data processing and analytics purposes.
Purpose	The dataset is representative of operational information regarding a crowdy seaport, ideal for a variety of large-scale maritime data processing and analytics scenarios, including, among others, mobility profiling, offline, and streaming data analytics. The dataset can be useful to both practitioners and researchers working on projects regarding maritime data management and awareness. Due to its diversity and long spatio-temporal coverage, the dataset can support the training of vessel profiling models, from movement behaviour (e.g. routine vs. unusual movement) to activity profiles (e.g. anchoring vs. cruising vs. fishing, in case of fishery boats).
Type / Format	Comma-Separated Values (CSV); ESRI Shapefiles
Size	4.5 GB
Origin	Re-used: The dataset was created by combining data collected by a terrestrial AIS receiver and publicly available datasets
Use beyond project	Research, Academia; Defense; Industry - Maritime Intelligence; Industry - Big Data; Industry – AI; Industry – Software Engineering; Industry – Supply Chain.

- Hellenic Trench AIS Data:

Description	Data from Automatic Identification System (AIS) transmissions received from both satellite and terrestrial receivers of the Marine Traffic network (www.marinetraffic.com) for one year (31 July 2015 to 31 July 2016) along the Hellenic Trench, the core habitat of the eastern Mediterranean.

Purpose	Environmental impact mitigation, Marine conservation; Collision risk analysis and mitigation with endangered marine mammals.
Type / Format	Comma-Separated Values (CSV)
Size	1.0 GB
Origin	Re-used: The dataset was created by combining data collected by a terrestrial AIS receiver and publicly available datasets Generated by: MarineTraffic (Kpler) Made Available by: Pelagos Cetacean Research Institute, Greece and IFAW, Great Britain
Use beyond project	Research, Academia; Defense; Industry - Maritime Intelligence; Industry - Big Data; Industry – AI; Industry – Software Engineering; Industry – Supply Chain.

- Synthetic AIS Dataset of Vessel Proximity Events:

Description	As sea routes become more congested and vessel speeds increase, the likelihood of significant accidents during a ship's operational life rises. The increasing congestion on sea lanes elevates the probability of accidents and especially collisions between vessels. The development of solutions and models for the analysis, early detection and mitigation of vessel collision events is a significant step towards ensuring future maritime safety. In this context, a synthetic vessel proximity event dataset is created using real vessel AIS messages. The synthetic dataset of trajectories with reconstructed timestamps is generated so that a pair of trajectories reach simultaneously their intersection point, simulating an unintended proximity event (collision close call). The dataset aims to provide a basis for the development of methods for the detection and mitigation of maritime collisions and proximity events, as well as the study and training of vessel crews in simulator environments. The dataset consists of 4658 samples/AIS messages of 213 unique vessels from the Aegean Sea.
Purpose	Collision risk analysis and mitigation between vessels, evaluation of collision detection and avoidance algorithms for maritime applications
Type / Format	Comma-Separated Values (CSV)
Size	1.9 MB

Origin	The dataset consists of 4658 samples/AIS messages of 213 unique vessels from the Aegean Sea. The steps that were followed to create the collision dataset are: Given 2 vessels X (vessel_id1) and Y (vessel_id2) with their current known location (LATITUDE [lat], LONGITUDE [lon]): Check if the trajectories of vessels X and Y are spatially intersecting. If the trajectories of vessels X and Y are intersecting, then align temporally the timestamp of vessel Y at the intersect point according to X's timestamp at the intersect point. The temporal alignment is performed so the spatial intersection (nearest proximity point) occurs at the same time for both vessels. Also for each vessel pair the timestamp of the proximity event is different from a proximity event that occurs later so that different vessel trajectory pairs do not overlap temporarily. Two csv files are provided. vessel_positions.csv includes the AIS positions vessel_id, t, lon, lat, heading, course, speed of all vessels. Simulated_vessel_proximity_events.csv includes the id, position and timestamp of each identified proximity event along with the vessel_id number of the associated vessels. The final sum of unintended proximity events in the dataset is 237. Examples of unintended vessel proximity events are visualized in the respective png and gif files.
Use beyond project	Research, Academia; Defense; Industry - Maritime Intelligence; Industry - Big Data; Industry – AI; Industry – Software Engineering.

- GPS Coordinates for AR application (HoloLens 2) [TUC/AR]

Description	GPS coordinates relayed from Android devices to pinpoint and provide user locations in the HoloLens 2 AR interface
Purpose	To enable real-time user location tracking within the AR application
Type / Format	String format containing concatenated latitude and longitude values
Size	Data is processed in real-time, with a frequency depending on the device's settings (typically every few seconds)
Origin	User's Android device via the app's GPS data retrieval functionality
Use beyond project	Not intended

- Surveys of T6.2:

Description	Survey results collected as part of the exploitation strategy
Purpose	To be used as an input for exploitation and business planning. (Task 6.2).
Type and Format	Alphanumeric characters and dependencies between them
Processing and Size	No data processing only data collection via https://umfragen.uni-paderborn.de/index.php/surveyAdministration/ and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

- Presence and individual Workshops:

Description	Workshop results collected as part of a joint collaboration on the exploitation strategy
Purpose	To be used as an input for exploitation and business planning. (Task 6.2).
Type and Format	Alphanumeric characters, pictures, posters and dependencies between them
Processing and Size	No data processing only data consolidation and saved in CREXDATA Microsoft Teams.
Origin	Participants of the workshop
Use beyond project	None.

4.2 FAIR Maritime Data

The following approaches have been set regarding the FAIR principles in the Maritime use case:

- Single Ground Based AIS Receiver Vessel Tracking Dataset:

Findable	Digital Object Identifier (DOI): 10.5281/zenodo.3754480 Findable using tags: AIS; Vessel Tracking Dataset
Accessible	Zenodo platform (Open Access)
Interoperable	AIS data standard adhering to the ITU Recommendation M.1371-5 (02/2014) https://www.itu.int/rec/R-REC-M.1371-5-201402-I/en
Re-usable	License: CC BY-NC-ND 4.0

- Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance:

Findable	Digital Object Identifier (DOI): 10.5281/zenodo.1167595 Findable using tags: Maritime data Automatic Identification System (AIS), Vessel positions, Sea state, Weather forecast, METOC data, Maritime protected areas, Ports register, Vessel register, Fleet register, Maritime geographic features
Accessible	Zenodo platform (Open Access)
Interoperable	AIS dataset adheres to the ITU Recommendation M.1371-5 (02/2014) https://www.itu.int/rec/R-REC-M.1371-5-201402-I/en and NMEA 4.0 standards Geographical, contextual and environmental data adhere to the ESRI standards
Re-usable	License: CC BY-NC-SA 4.0

- The Piraeus AIS Dataset for Large-scale Maritime Data Analytics:

Findable	Digital Object Identifier (DOI): 10.5281/zenodo.5792100 Findable using tags: AIS Data, Maritime, Piraeus, Big Data Analytics, Trajectory Dataset
Accessible	Zenodo platform (Open Access)
Interoperable	AIS dataset adheres to the ITU Recommendation M.1371-5 (02/2014) https://www.itu.int/rec/R-REC-M.1371-5-201402-I/en Geographical, contextual and environmental data adhere to the ESRI standards
Re-usable	AIS Data License: CC BY-NC-SA 4.0 Geographic-related Data License: Regions: CC BY 3.0 Harbours, Receiver, Location, Spatial, Coverage: CC BY-NC-SA 4.0 Piraeus Port, Islands, Territorial, Waters: ODbL 1.0 Weather Data License: NOAA Data Policy Synopses meta-data License: CC BY-NC-SA 4.0

- Hellenic Trench AIS Data:

Findable	Digital Object Identifier (DOI): 10.17882/57040
Accessible	SEANOE (SEA scieNtific Open data Edition) platform (Open Access)
Interoperable	AIS dataset adheres to the ITU Recommendation M.1371-5 (02/2014) https://www.itu.int/rec/R-REC-M.1371-5-201402-I/en
Re-usable	AIS Data License: CC BY-NC-SA 4.0

- Synthetic AIS Dataset of Vessel Proximity Events:

Findable	Digital Object Identifier (DOI): 10.5281/zenodo.8358664 Findable using tags: AIS, maritime accidents, Trajectory, Vessel, collision
Accessible	Zenodo platform (Open Access)
Interoperable	AIS data standard adhering to the ITU Recommendation M.1371-5 (02/2014) https://www.itu.int/rec/R-REC-M.1371-5-201402-I/en
Re-usable	AIS Data License: CC BY-NC-SA 4.0

4.3 Data Security in the Maritime Use Case

Maritime datasets are preserved, and curated following database/stream management system best practices as prescribed by the corresponding vendors and open-source communities (Zenodo). Steps taken by Kpler to ensure data privacy and confidentiality include data minimization, augmentation and anonymization. Additionally, for Kpler proprietary streaming AIS datasets used in the context of the CREXDATA project, secure data storage environments are used. Access control is implemented through end-to-end security mechanisms for authentication. Data storage and management involve input sanitization and logging breach protection. Protocols are in place for data retention, recovery and breach responses.

5 Summary

CREXDATA will deal with data sources and datasets of very different nature as input for demonstrator systems as well as for research purposes; output datasets will also be generated within the three use cases involving Weather Emergencies, Health Crises and Maritime domains.

This Data Management Plan summarizes the aforementioned data and describes initial provisions for the way they will be handled, shared and rendered findable, easily accessible, interoperable and reusable for relevant scientific, research or other communities. The Data Management Plan will be enhanced and amended, where needed, in future updates of this deliverable.