



Horizon Europe

Project: 101079789

D3.3 - Report on Data Collection and Generation

WP 3 – Data and Ethics

WP Leader:	CNR
Date:	MAY 2025
Nature:	DEC
Dissemination level:	Public

Document Information

Grant Number	Agreement	101079789	Acronym	EIRENE PPPPPP
Full title	Environmental Exposure Assessment Research Infrastructure Preparatory Phase Project			
Project URL	https://www.eirene.eu/			
Project Officer	Andreas Holtel, Andreas.HOLTEL@ec.europa.eu			

Delivery date	Contractual	31/5/2025	Actual	30/5/2025
Status	Final			
Nature	DEC			
Dissemination level	Public			

Responsible Partner	CNR			
Responsible Author	Nicola Pirrone		E-mail	nicola.pirrone@iia.cnr.it
	Partner	Phone	+39.0984.493239
Other partners	VITO, RECETOX			

Document History

Institution	Date	Version
CNR	2025/05/31	v.01

Table of Contents

1. Introduction.....	4
2. Data Collection.....	4
2.1. Methods of Data Collection.....	4
2.2. Advanced Tools and Technologies for Environmental Data Collection.....	6
2.3. Advanced Tools and Technologies for Health Data collection.....	8
2.4. Hurdles in Data Collection.....	11
3. Data Generation.....	12
3.1. Methods for data Generation.....	12
3.2. Tools and Techniques for Data Generation.....	14
3.3. Challenges in Data Generation.....	16
3.4. Data Generation in the exposome domain.....	17
3.5. Ethical, Legal and Security implications in Data Generation.....	18
4. Data Collection and Generation in EIRENE.....	20
5. Conclusions.....	22
6. References.....	22
Annex 1 - List of Services.....	23

1. Introduction

The activity carried out in Task 3.3 was focused on possible processes for collecting new data and federating existing datasets. Reliable and effective data pipelines will be designed for processing raw data and data provided by external data providers to build the EIRENE RI Virtual Infrastructure. The EIRENE RI distributed infrastructure will allow scientists to process data using complex algorithms and tools as well as support policymakers, NGOs, and citizens in their assessments. This Deliverable gives an overview on methods for data Collection and Data Generation, the latter becoming more and more fundamental for research, providing the raw material needed to gain data insights, build models, and make informed decisions. A short paragraph describes Data Collection and Generation in EIRENE that is based over an ongoing activity. Editing of this Deliverable was assisted by GEMINI the Google AI-powered tool.

2. Data Collection

Data collection refers to the process of gathering data from various sources and through different methods. Most often data are numerical values (quantitative) but sometime can be descriptive (qualitative). In the environmental domain quantitative data refers to numerical information that is collected and analysed to understand environmental processes, conditions, or impacts. Quantification of parameters occurs through scientific instruments or monitoring systems. Quantitative data is crucial for understanding and assessing exposome.

On the other side quantitative and qualitative data in the health domain refers to numerical and textual information collected and analysed to understand various aspects of health, disease patterns, medical treatments, and outcomes. This data is essential for evidence-based decision-making, public health research, and policy formulation.

Methods for data collection follow different approaches when they occur in different domains such as environmental and health studies.

2.1. Methods of Data Collection

The following methods for data collection are of interest for EIRENE and can occur in parallel, one following another or alone.

Observational Studies

In observational studies, researchers observe and measure variables without actively intervening or manipulating the subjects. This contrasts with experimental studies, where researchers control and manipulate variables. A variable can be measured through sensors or instruments and the goal is to identify associations between variables, but it's crucial to remember that observational studies cannot establish cause-and-effect relationships.

Experiments

Data collection within experimental studies focuses on rigorous control and precise measurement of parameters. When doing an experimental data collection, it's fundamental to isolate of the effects of the independent variable on the dependent variable. This often involves creating controlled environments (e.g., laboratory settings) or carefully manipulating variables in field experiments. Manipulation of the independent variable might occur carefully and minimizing bias and ensuring comparability can be ensured through randomization techniques. Also, replication of experiments gives the reliability of the results.

Surveys, Questionnaires and Interviews

These methods gather data on attitudes, beliefs, and behaviours through structured questionnaires. When preparing surveys, questionnaires, and interviews, careful planning is essential for collecting

accurate and valuable data. A clear goal can help formulation of specific questions and guide the following analysis. Question types can be closed-ended questions, which provide predefined response options (e.g., multiple-choice) or open-ended questions, which allow respondents to provide free-form answers. While the former can be supported by numerical responses (e.g., number of respondents who have selected a colour), the latter require much attention in the analysis. This method of data collection requires clear, concise, and unbiased language, needs the definition of the target population and selection of a representative sample (appropriate sampling method) and request a careful attention to ethical and legal aspects laying on the privacy of the interviewed people.

Online resources

Data are even more available through servers connected to the web. Generally speaking sources of data can be web articles, scientific journals, reports, databases or archives. These online resources can be collected through several methods, which are non-exhaustively hereafter reported:

- **Manual download:** this is not suited for large datasets and can be useful for old scanned documents which require a deep pre-processing like the Optical Character Recognition (OCR) and the dataset structuring.
- **Web Scraping:** is the process of automatically extracting data from websites. It is done by a code (or software) that parses the HTML structure of web pages to locate and retrieve specific information, typically reported in tables. The code sends HTTP requests to web servers, retrieves the HTML content, and then parses the content to collect the data. These data can be saved in different formats. Web scraping is very flexible, extracting any available data and well suited for historical data not connected with API. One of the main disadvantages is that scraping without permission can raise ethical and legal issues.
- **APIs (Application Programming Interfaces):** are tool that allow communication between different interoperable systems. APIs can give access to data or web services. They provide predefined endpoints, which allow requests by means of structured formats, typically XML or JSON. Data retrieved with APIs are structured and do not need pre-processing before archiving. Most notable is the legal aspect as the API is provided by the data provider who as in principle defined the licence. Not all websites offer APIs and some may have usage limitations or require payment for access
- **Databases:** while datasets represent a collection of data in the form of table, they do not have a relation each other and cannot be queried. On the other side a database holds structured data as well data structures and relationships, indexes, views, etc. Datasets are stored in static files while databases are managed by Database Management Systems (DBMSs) that allow for querying, updating, and organizing data efficiently. The following table reports key differences between a Dataset and a Database (**table 1**).

Table 1. Key differences between Dataset and Database

Feature	Dataset	Database
Definition	A collection of related data, typically in a single file or format.	A system used to store, manage, and organize data, often with a DBMS.
Structure	Simple (usually a table or file).	Complex (multiple tables, relationships, and data structures).
Storage	Stored as files (e.g., CSV, JSON).	Stored in a database management system (DBMS).
Management	Typically static, used for analysis or modelling.	Dynamic, supports queries, updates, and transactions.
Purpose	For analysis, machine learning, or research.	For storing, organizing, and retrieving large datasets.
Examples	A CSV file, Excel sheet, JSON file.	MySQL, PostgreSQL, MongoDB, Oracle DB.
Size & Complexity	Smaller, simpler datasets.	Larger, more complex with multiple datasets and tables.

Databases can be built with observations, experiments or derive from open-source shared data. Results of researches are published through scientific journal or specialized books or reports, which are archived in digital libraries or made available through archives. Data sharing is becoming popular with the availability of open-source platforms that allow users to submit or share data in a collaborative manner.

2.2. Advanced Tools and Technologies for Environmental Data Collection

Environmental data collection can occur with various methodologies and technologies, which are selected according the monitoring scope. The following sections will summarize only advanced methodologies and technologies presuming that classical approaches are well known.

Remote sensing drones

Drones can be unmanned aerial vehicles (UAVs) or unmanned surface vehicles (USVs) (terrestrial or marine drones), which can be equipped with various sensors and cameras to gather environmental parameters in a cost-effective way. Large and inaccessible areas can be monitored, providing very high-resolution data that is often superior to traditional methods or satellite imagery. Without detailing the carriers, a non-exhaustive list of sensors can be reported looking the specific environmental parameter being monitored:

- **Cameras:** Standard cameras (RGB) or thermal infrared cameras can capture images in red, green, and blue wavelengths, providing visual data for land cover mapping, habitat assessment, and change detection, the former while detects and measures the infrared radiation emitted by objects, the latter.
- **Multispectral Sensors:** Capture data multiple bands of the electromagnetic spectrum beyond visible light (e.g., near-infrared, red-edge), which allows for the assessment of environmental health, quality, and conditions.
- **Hyperspectral Sensors:** Similar to multispectral sensors but capture data in hundreds of narrow, continuous spectral bands, enabling detailed analysis.
- **Thermal Sensors:** Detect infrared radiation, allowing for the measurement of surface temperatures.
- **LiDAR (Light Detection and Ranging):** Uses laser pulses to detect the fluorescence signal returned by a sensor that can measure an environmental parameter.
- **Air Quality Sensors:** Can measure various atmospheric pollutants such as particulate matter, ozone, nitrogen dioxide, and sulphur dioxide.
- **Water Quality Sensors:** Some drones can be equipped to collect water samples or carry sensors to measure parameters in water.

Internet of Things (IoT) devices

Some sensor can be directly connected to the Internet for transmitting measurements. The connectivity can occur with established wireless infrastructures (Wi-Fi), short range gateway devices (Bluetooth), cellular networks (*G, LTE-M, NB-IoT), satellite or other long-range, low-power communication technologies. Pervasive monitoring, interoperability, large amount of collected data and low energy requirement are main challenges, while weakness can include connectivity issues, privacy and security, calibration and maintenance.

Citizen observatories

Citizen observatories are community-based environmental monitoring initiatives that actively involve citizens in the process of collecting and sharing environmental data. They represent a powerful application of citizen science, leveraging the collective capacity of the public to expand the spatial and temporal coverage of environmental monitoring efforts. Despite citizen observatories are recognized as a valuable tool for environmental data collection, some criticism is reserved for environmental observations which require very high competences and knowledge.

Artificial Intelligence and Machine Learning

Environmental Data Collection using Artificial Intelligence (AI) and Machine Learning (ML) represents a significant advancement in how we monitor and understand our planet. By leveraging the power of these technologies, we can analyse vast amounts of environmental data with greater efficiency, accuracy, and insight than traditional methods allow.

AI-powered sensors can autonomously collect and pre-process environmental data in real-time, reducing the need for manual data gathering. For example, AI algorithms can be integrated into water quality sensors to automatically identify anomalies or specific pollutants.

AI can enable drones and robots to navigate autonomously, capture high-resolution imagery and sensor data over large or inaccessible areas (e.g., forests, oceans, disaster zones), and even perform tasks like sample collection. While ML algorithms can analyse the collected data and make much robust the dataset by analysing massive datasets, filling gaps and detecting anomalies or outliers.

AI and ML are powerful at analysing the massive datasets generated by Earth observation satellites. They can automatically identify land cover changes, monitor deforestation, track ice melt, detect pollution plumes, and assess vegetation health with greater speed and accuracy than manual interpretation.

The highest contribute of AI and ML is on improvement of data quality and efficiency. ML algorithms can identify unusual patterns or outliers in environmental data (Anomaly Detection), flagging potential errors or significant environmental events that require further investigation.

AI can integrate data from diverse sources (e.g., ground sensors, satellites, weather models) (Data Fusion) to create a more comprehensive and reliable understanding of environmental conditions.

ML techniques can be used to estimate missing data points based on existing patterns (Data Gap Filling), ensuring more complete datasets for analysis.

Blockchain and smart contracts

The Blockchain is a shared ledger with growing lists of records (blocks) that are securely linked together via cryptographic hashes. The ledger is replicated, shared, and synchronized and digital data is geographically distributed across many sites, countries, or institutions.

Smart contracts are digital transaction protocols that automatically facilitate, verify, or enforce the negotiation or performance of a contract, according to the agreement terms.

Blockchain provides a decentralized and tamper-proof ledger for storing environmental data. With blockchain data alteration do not occur, ensuring the integrity and trustworthiness of the information. Blockchain can track the origin and journey of environmental data from the sensor to the end-user. This is particularly valuable for supply chain monitoring, ensuring the ethical and sustainable sourcing of materials. Integrating blockchain with IoT sensors allows for real-time environmental data monitoring. The immutability of the blockchain ensures that the collected data is verified and trustworthy. For example, air quality monitoring data from a network of sensors can be securely recorded and accessed by relevant parties.

Smart contracts, can automate the process of validating and verifying environmental data based on predefined conditions. This reduces the need for manual checks and minimizes errors. They can be used to create automated processes based on verified data from environmental monitoring systems and enforce sustainability agreements within supply chains, ensuring that all participants adhere to environmental standards.

Virtual reality (VR) and augmented reality (AR)

Environmental data collection can be significantly enhanced and transformed through the use of Virtual Reality (VR) and Augmented Reality (AR) technologies, which are immersive technologies offer novel ways to visualize, interact with, and even collect environmental information.

VR creates fully immersive, simulated environments that users can interact with. In the context of environmental data collection, VR can be used for Data Visualization – 3D visualizations of complex environmental datasets. Instead of looking at graphs or maps on a 2D screen, for example – and for Simulation and Modelling for better understanding environmental processes and the potential impacts of different scenarios.

AR overlays digital information onto the real world, enhancing the user's perception of their surroundings through devices like smartphones, tablets, or specialized AR glasses. This technology can guide users through environmental monitoring protocols, providing step-by-step instructions and visual cues for collecting data accurately. It can also empower citizen observatories by providing intuitive tools for collecting and submitting environmental observations, such as air quality readings, biodiversity sightings, or pollution reports, contributing to larger datasets.

APIs and Web Scraping Tools

API tools are provided by data providers. For example, weather APIs provide current weather data, forecasts, and historical information while mapping APIs can enable embedding maps, geocoding addresses, and calculating routes. Web scraping can be completed with:

- Browser Extensions (Web Scraper) that allow users to visually select and extract data from web pages within their browser;
- Online Scraping Services that provide visual interfaces for building scraping workflows;
- Programming Libraries that offer flexibility and control for building custom scraping scripts;
- Desktop Software that can allow download of entire websites for offline browsing and data extraction.

2.3. Advanced Tools and Technologies for Health Data collection

Surveys and Questionnaires

Health data collection through surveys and questionnaires is a widely used and valuable method for gathering information about individuals' health status, behaviours, attitudes, and experiences. These tools can be administered in various formats and settings, providing insights into a broad range of health-related topics.

Surveys and Questionnaires in Health Data Collection can be:

- Self-Administered Questionnaires: Individuals complete these independently, either on paper or electronically (online surveys). They offer anonymity and allow respondents to answer at their own pace.
- Interviewer-Administered Questionnaires: A trained interviewer asks the questions and records the responses, either in person or over the phone. This method can yield higher response rates and allow for clarification of questions.
- Cross-Sectional Surveys: Data is collected at a single point in time to provide a snapshot of the health status or behaviours of a population at that moment.
- Longitudinal Surveys: Data is collected from the same individuals repeatedly over a period of time to track changes in health outcomes, behaviours, or attitudes.
- Retrospective Surveys: Participants are asked to recall past health events, behaviours, or exposures. These can be prone to recall bias.
- Prospective Surveys: Data is collected about current or future events, behaviours, or exposures and then linked to future health outcomes.

With Surveys and Questionnaires, a wide array of health-related information can be gathered: demographics, health status, symptoms and health complaints, health behaviours, healthcare access and utilization, psychological and social factors, environmental exposures, knowledge, attitudes, and beliefs and experiences with healthcare. Surveys and Questionnaires are powerful tools for collecting a wide range of health data. While they have limitations, careful design, implementation, and

consideration of potential biases can yield valuable insights into the health of individuals and populations, informing public health initiatives, clinical practice, and research.

Clinical Interviews

Clinical interviews are a fundamental method for collecting health data, particularly in mental health and behavioural health settings, but also relevant in general medicine for gathering detailed patient histories and understanding the psychosocial context of physical health. They involve a structured or semi-structured conversation between a trained healthcare professional (e.g., psychiatrist, psychologist, social worker, physician) and a patient to gather information relevant to their health and well-being.

Clinical Interviews pose significant challenges in data collection as they are:

- Purposeful: The primary goal is to collect specific information to understand the patient's symptoms, history, functioning, and context. This information is crucial for diagnosis, treatment planning, and monitoring progress.
- Interactive: Unlike self-report questionnaires, clinical interviews involve a dynamic exchange between the interviewer and the patient. The interviewer can probe for more detail, clarify responses, and observe non-verbal cues.
- Flexible: Interviews can range from highly structured (following a specific set of questions in a fixed order) to unstructured (allowing the conversation to flow more naturally based on the patient's responses). Semi-structured interviews, which use a guide of topics and questions but allow for flexibility in follow-up, are common.
- Observational Component: The clinician observes the patient's behaviour, affect (emotional expression), speech, and thought processes during the interview, which can provide valuable clinical information beyond the verbal responses.
- Relationship-Based: A therapeutic alliance and rapport between the clinician and the patient are crucial for fostering trust and encouraging honest and comprehensive disclosure.

Clinical Interviews can provide rich and detailed information, allowing for in-depth exploration of the patient's experiences and perspectives. They are flexible and adaptable because the interviewer can tailor questions and probes based on the patient's responses and emerging information. Several times observation of non-verbal cues can provides insights into the patient's emotional state and thought processes beyond their verbal communication. Clinical Interviews increase establishment of rapport by building a therapeutic relationship, facilitating more open and honest disclosure. In addition, the interviewer can clarify ambiguous responses and ask follow-up questions to obtain more comprehensive information, he/she can help to understand the social, cultural, and personal context of the patient's health issues. In summary, Clinical Interviews can allow for a holistic evaluation of the patient's physical, psychological, and social well-being.

Medical Records Review

Medical records review is a crucial method for health data collection, involving the systematic examination of patient charts and electronic health records (EHRs) to extract relevant information for various purposes, including research, quality improvement, public health surveillance, and clinical audits. It provides a wealth of real-world data on patient demographics, diagnoses, treatments, outcomes, and healthcare utilization. EHRs are becoming even more frequent despite paper-based records are sometime maintained. A large amount of past collected information is available and several initiatives to digitalize such significant data is occurring. Digitalization incur in several limitations like harmonization, address missing data, check protocols, to cite few.

Clinical Trials

Clinical trials are research studies conducted in humans to evaluate the safety and efficacy of new medical interventions — such as drugs, vaccines, medical devices, or behavioural modifications. Health data collection is a cornerstone of clinical trials, providing the evidence needed to determine

if an intervention is safe and effective. The rigor and quality of this data collection are paramount to the integrity and validity of the trial results.

A variety of methods are employed to collect health data in clinical trials, ensuring accuracy, completeness, and consistency:

- **Case Report Forms (CRFs):** These are standardized documents (paper-based or electronic - eCRFs) used to record all protocol-required information for each participant. eCRFs offer advantages like real-time data entry, automated checks, and improved data quality.
- **Electronic Data Capture (EDC) Systems:** Integrated software platforms used for designing CRFs, entering, managing, and reporting clinical trial data. They often include features for data validation and audit trails.
- **Patient-Reported Outcomes (PROs):** Data collected directly from patients about their health status, symptoms, and quality of life using questionnaires, diaries (paper or electronic - ePROs), or interviews.
- **Clinical Assessments:** Standardized evaluations conducted by trained study staff, such as physical examinations, neurological assessments, and cognitive tests.
- **Laboratory Tests:** Collection and analysis of biological samples (e.g., blood, urine, tissue) to measure specific parameters.
- **Imaging Techniques:** Use of X-rays, MRIs, CT scans, and other imaging modalities to assess disease status and treatment response.
- **Physiological Monitoring:** Continuous or intermittent monitoring of vital signs (e.g., blood pressure, heart rate) using electronic devices. Wearable devices and sensors are increasingly used for remote monitoring in decentralized clinical trials.
- **Interactive Response Technology (IRT):** Systems used for randomization of participants, managing drug supply, and collecting some patient data via phone or web.
- **Direct Data Capture (DDC):** Data generated directly from electronic devices (e.g., lab instruments, ECG machines) and automatically transferred to the trial database, reducing manual data entry.
- **Hospital Records and Pharmacy Dispensing Records:** Review of existing medical records to gather information on medical history, concomitant medications, and healthcare utilization.

The specific data collected in a clinical trial depends on the nature of the intervention being studied and the trial's objectives. However, common categories include:

- **Baseline Data:** Collected before the intervention begins. This includes demographic information (age, sex, ethnicity), medical history, current health status, relevant disease-specific measures, and lifestyle factors.
- **Efficacy Data:** Measures the effect of the intervention on the disease or condition being studied. This can include clinical outcomes (e.g., reduction in symptoms, disease progression), physiological measures (e.g., blood pressure, lab results), and patient-reported outcomes (PROs).
- **Safety Data:** Monitors participants for any adverse events (AEs) or serious adverse events (SAEs) that may occur during the trial. This includes detailed documentation of any health problems, changes in lab values, or other indicators of harm.
- **Pharmacokinetic/Pharmacodynamic (PK/PD) Data:** In drug trials, this data describes how the drug is absorbed, distributed, metabolized, and excreted by the body (PK) and the drug's effects at the biological level (PD).
- **Quality of Life (QoL) Data:** Assesses the impact of the intervention and the disease on the participant's overall well-being and daily functioning, often collected through questionnaires.
- **Healthcare Resource Utilization:** Data on hospitalizations, doctor visits, medications used, and other healthcare services consumed during the trial.
- **Genetic or Biomarker Data:** Increasingly collected to understand how genetic factors or specific biomarkers might influence a participant's response to the intervention.

Biological Sampling

Biological sampling is a fundamental method for health data collection, involving the collection of biological materials from individuals for analysis. These samples can provide crucial information about a person's health status, disease processes, genetic makeup, exposure to environmental factors, and response to medical interventions. The types of samples collected and the analyses performed vary widely depending on the research question or clinical need.

Common types of biological samples can include samples of blood, urine, tissue, saliva, stool, hair and nail, cerebrospinal fluid, semen and breath. Biological sampling uses standardized protocols, trained personnel, appropriate tracking and secure data management.

Ethical Considerations

When collecting Health Data, ethical considerations are of paramount importance. Given the sensitive nature of health information and its potential impact on individuals and communities, it's crucial to adhere to ethical principles and guidelines to protect privacy, ensure fairness, and maintain trust. Here are key ethical considerations in health data collection:

- storage and use of human biological samples have to comply with all national and EU-level ethical and legal rules;
- access of researchers to personal data and biological samples have to be defined by Ethics, Legal and Societal Implication (ELSI) guidelines;
- informed consent must be required when health and biological data are collected, which include what data will be collected, how they will be used, who will have access, and for how long it will be stored; data should only be used for the specific purposes for which they were collected and for which consent was given. Using data for new, unrelated purposes raises ethical concerns;
- robust measures must be in place to protect data from unauthorized access, use, disclosure, alteration, or destruction. This includes physical, technical, and administrative safeguards;
- whenever possible, data should be anonymized or de-identified to remove any direct identifiers that could link the information back to an individual;

2.4. Hurdles in Data Collection

When collecting data some key points might be considered, which can pertain to data quality, data access and availability and data bias.

Data Quality

Inaccuracy have cascading negative effects on analysis, leading to erroneous assumptions. Among sources of inaccuracy there are data entry mistakes like typos entering data or misreading or wrong field displacement. Where possible the inaccuracy has to be reduced following Standard Operative Procedures or Quality Checks, the former established for instruments and analytical procedures and the latter for other generated data.

Misunderstand is frequent whit inaccurate preparation of surveys. Participants might also provide incorrect information intentionally (social desirability bias), or have poor recall.

Few technical issues can affect data quality like for example inaccurate reading of sensor (which produce measurement spikes), erroneous setting of time (local time vs UTC), data corruption while transferring, bad instrumental calibration, lacking training or inconsistent application of protocols.

Data Access and Sharing

Data access and data sharing are crucial in the context of data collection. Technical barriers might be considered, which are related with data silos and data fragmentation as well as with storage in proprietary or unusual formats.

More on Data Access and Sharing can be found in the companion Deliverable *D3.4 – Data Access and Sharing*.

Bias

Data collection bias is a systematic error that occurs during the data gathering process, resulting in data that is not truly representative of the population or phenomenon being studied. This can lead to skewed findings, inaccurate conclusions, and flawed decision-making. Bias can creep in at various stages of data collection, influenced by factors ranging from the study design and sampling methods to the way questions are asked and responses are recorded.

Most bias in data collection depends from human cognitive biases, limitations in data collection tools or methods, resource constraints, or a lack of careful planning and oversight in the research design. It can be generated by erroneous cohort selections or not representative monitoring site.

3. Data Generation

Data generation refers to the process of creating or producing new data. This is a fundamental aspect of research, analysis, and machine learning, providing the raw material needed to gain insights, build models, and make informed decisions. Main scopes of data generation can be low data availability, imbalance datasets, or privacy problems.

Generating datasets for data privacy is critical for changing information so that the original data in the source dataset cannot be discovered. Anonymization – by removing personal identifying information such as names, addresses, and phone numbers – or pseudonymization – by substituting identifying information with pseudonyms – are the most common processes for data privacy generation.

On the other hand, data augmentation is a fundamental machine learning technique that improves dataset quality by producing additional data from existing data.

There is also a variety of artificial datasets that can be structured in the form of tabular data or unstructured that produce media data. Also, artificial text data can be generated by natural language processing applications.

Data can be generated through various methods, broadly categorized as primary data collection, utilizing secondary data sources, and generating synthetic data. Synthetic data creation is generally suitable for testing model functionalities or while expanding existing datasets can occur through ad-hoc models or through statistical methods (<https://www.turing.com/kb/synthetic-data-generation-techniques>).

3.1. Methods for data Generation

Data generation through simulation is a powerful technique for creating artificial datasets (or synthetic datasets) that mimic the characteristics and behaviour of real-world systems or phenomena. Unlike collecting primary data or utilizing existing secondary data, simulation allows for the creation of data under controlled conditions, enabling researchers and analysts to explore various scenarios, test hypotheses, and train models without the limitations or costs associated with real-world data acquisition.

The core idea behind data generation through simulation is to build a computational model that represents the essential aspects and dynamics of a system. By running this model, synthetic data is produced based on the rules, parameters, and interactions defined within the simulation. This generated data can then be used for a multitude of purposes, especially when real data is scarce, expensive to obtain, contains privacy concerns, or doesn't cover the full range of possible situations.

Several types of simulation approaches are commonly used for data generation:

- **Monte Carlo Simulation:** This method relies on repeated random sampling to generate outcomes based on defined probability distributions for input variables. By running a large number of iterations, Monte Carlo simulations can produce a distribution of possible results, providing insights into the likelihood of different outcomes in systems with inherent uncertainty. This is widely used in finance, physics, engineering, and project management to model risk and predict potential results.
- **Agent-Based Modelling (ABM):** ABM focuses on simulating the actions and interactions of autonomous individual entities (agents) within an environment. The behaviour of the overall system emerges from the collective actions and interactions of these agents, which follow defined rules. ABM is particularly useful for generating data that reflects complex adaptive systems, such as market dynamics, the spread of diseases, or social phenomena, where macro-level patterns arise from micro-level behaviours.
- **Discrete-Event Simulation (DES):** DES model systems as a sequence of events occurring at distinct points in time. The simulation clock advances from one event to the next, and the system's state changes only when an event occurs. This approach is well-suited for generating data related to processes and workflows, such as customer flow in a service system, supply chain logistics, or manufacturing operations, helping to identify bottlenecks and optimize performance.
- **System Dynamics:** This methodology focuses on understanding the behaviour of complex systems over time by modelling the feedback loops, delays, and stocks and flows within the system. System dynamics models are typically represented using stock and flow diagrams and causal loop diagrams. Running these simulations generates data that shows how the system's variables change over time in response to different policies or external factors, useful for strategic planning and understanding long-term trends.

Creating artificial data that mimics the statistical properties and characteristics of real-world data but does not contain actual yet observed information. Synthetic data is particularly useful for training machine learning models, testing systems, and addressing privacy concerns when real data is sensitive or scarce. Methods include a) **Generative AI Models**, utilizing models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and transformer models (like GPT); b) **Statistical Models**, generating data based on known statistical distributions (e.g., normal, uniform) or by modelling the relationships and correlations observed in real data; c) **rule-Based Approaches**, Creating synthetic data based on predefined rules, logic, or simulations that reflect the characteristics of the data being modelled; and d) **Data Augmentation**, a technique primarily used to increase the size and diversity of existing datasets, particularly in machine learning. It involves applying transformations to existing data points to create new, slightly modified samples. While not generating entirely new data from scratch like some other synthetic data methods, it expands the dataset based on existing observations.

Data can also be generated randomly by producing data values that are not determined by a predictable pattern or sequence. This technique is fundamental in various fields, including statistics, computer science, simulations, and testing, where the need for varied and unpredictable data is crucial.

The core principle of random data generation lies in the absence of a discernible order or relationship between successive data points. This is often achieved through algorithms or processes designed to produce outputs that are as close to truly random as possible.

There are generally two main types of random data generation:

- **True Random Number Generation (TRNG):** This method relies on physical phenomena that are inherently random and unpredictable, such as atmospheric noise, thermal noise in electronic components, or radioactive decay. TRNGs produce highly unpredictable sequences of numbers, but they can be slower and require specialized hardware.

- **Pseudo-Random Number Generation (PRNG):** This is the most common method used in computing. PRNGs use mathematical algorithms to generate sequences of numbers that appear random but are in fact deterministic. They start with an initial value called a "seed," and the algorithm then produces a sequence based on this seed. While not truly random, good PRNGs generate sequences that are statistically random and suitable for most applications. The same seed will always produce the same sequence, which can be useful for reproducibility in simulations and testing.

Finally, procedural data generation is a method of creating data algorithmically, rather than manually. It involves defining a set of rules, parameters, and algorithms that a computer follows to automatically generate a potentially vast amount of varied and complex data. This technique is widely used across various fields, particularly where the manual creation of large datasets or content is impractical or undesirable.

The core principle of procedural generation lies in using procedures (algorithms) to determine the characteristics and form of the data. Instead of storing pre-made data, the system stores the rules and a starting "seed" (often a number). The algorithm then uses this seed and the rules to generate the data on the fly or in advance. This means that a small set of rules and a seed can produce a large and unique output.

3.2. Tools and Techniques for Data Generation

Synthetic data generation tools are increasing their availability on the market and some of them are displayed in **figure 1**. Whereas the following non-exhaustive **table 2** summarizes each tool reported in the figure.

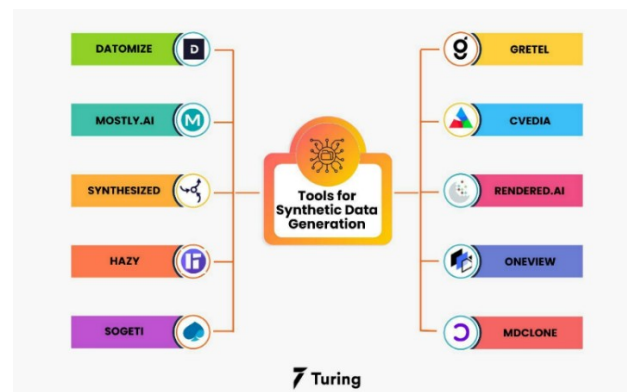


Figure 1. Example of data generation tools (Credit: <https://www.turing.com/kb/synthetic-data-generation-techniques>).

Table 2. Tools for synthetic data generation (Credit: [turing.com](https://www.turing.com))

Tool	Description
Datomize	It has an AI or ML model which is majorly used by world-class banks all over the globe. With Datomize, you can easily connect your enterprise data services and process high-intensity data structures and dependencies with different tables. This algorithm will help you in extracting behavioral features from the raw data and you can create identical data twins with the original data.
MOSTLY.AI	It is a synthetic data tool that enables AI and high-priority privacy while extracting structures and patterns from the original data for preparing completely different datasets.
Synthesized	It is an all-in-one AI dataOps solution which will help you with data augmentation, collaboration, data provisioning, and secured sharing. This tool generates different versions of the original data, and also tests them with multiple test data. This helps in identifying the missing values and finding sensitive information.
Hazy	It is a synthetic data generation tool that aims to train raw banking data for fintech industries. It will let

the developers ramp up their analytics workflows by avoiding any fraudulence while collecting real customer data. You can generate complex data during financial service generations and store it in silos within the company. But, sharing real financial data for research purposes is severely limited and restricted by the government.

Sogeti	It is a cognitive-based solution that helps you with data synthesis and processing. It uses Artificial Data Amplifier technology which reads and reasons with any data type, whether it's structured or unstructured. ADA uses deep learning methods to mimic recognition capabilities and sets it apart.
Gretel	It is the tool that is specifically built to create synthetic data. It is a self-proclaimed tool that generates statistically equivalent datasets without giving out any sensitive customer data from the source. While training the model for data synthesis, it compares the real-time information by using a sequence-to-sequence model for enabling the prediction while generating new data.
CVEDIA	It provides synthetic computer vision solutions for improved object recognition and AI rendering. It is used for a variety of tools, and IoT services for developing AI applications and sensors. It is packed with different machine language algorithms.
Rendered.AI	It generates physics-based synthetic datasets for satellites, robotics, healthcare, and autonomous vehicles. It is a no-code configuration tool and API for engineers to make quick changes and analytics on datasets. They can perform data generation on the browser and it will enable easy operation on ML workflows without much computing power.
Oneview	It is a data science tool that uses satellite images and remote sensing technologies for defense intelligence. Using mobiles, satellites, drones, and cameras, this algorithm will help object detection even where there are blurred images or lower resolutions. It will provide accurate and detailed annotations on the virtually created imagery which will closely resemble the real-world environment.
MDCClone	It is a dedicated tool that is majorly used in healthcare businesses for generating an abundance of patient data which will allow the industry to harness the information for personalized care. But, for accessing clinical data, researchers should depend on mediators and the process was slow and limited. It offers a systematic approach for democratizing healthcare data for research, synthesis, and analytics without disturbing sensitive data.

A few Python-based libraries (**Table 3**) can serve synthetic data generation for specific requirements. Selection of an appropriate library for the kind of data required to be generated preserve from unexpected results.

Table 3. Example of Python-based libraries for generating artificial datasets

Python Library	Purpose
DataSynthesizer, Sym Py	Increasing data points
Fakeer, Pydbgen, Mimesis	Create fake names, addresses, contact, or date information
Synthetic Data Vault (SDV)	Create relational data
Platipy	Create entirely fresh sample data
TimeSeriesGenerator, Synthetic Data Vault	Timeseries data
Gretel Synthetics, Scikit-learn	Automatically generated data
Mesa	Complex scenarios
Zpy, Blender	Image data
Blender	Video data

In the case of Data privacy datasets two libraries can be used data generation, the SmartNoise library (<https://pypi.org/project/smartnoise-synth/>) and the Synthcity library (<https://pypi.org/project/synthcity/>), the latter offering various methods of generating synthetic data but not exclusively methods for ensuring privacy. The SmartNoise library offers a set of methods for synthesizing data privately, ensuring the privacy of the original data during the synthetic data generation process.

Details on installation, methods and commands can be obtained online.

3.3. Challenges in Data Generation

The current methods for generating synthetic datasets are notable revealing challenges and shortcomings (Hao et al.,). Most common limitations are the followings.

Bias in Synthetic Data

Major incongruity between synthetic datasets and their real counterparts, encompasses disparities in feature and class distribution, and other statistical parameters. This bias drives models to provide misleading simulations compromising their accuracy to replicate real-world phenomena.

Incomplete Data

Synthetic datasets may be missing parts or contain only partial information. This happens due to imperfections in the generation process or because it doesn't fully capture the real-world changes. This lack of complete information can make it hard for models to accurately handle situations in the real world where data is incomplete, affecting how reliable and useful the model is.

Inaccurate Data

Synthetic datasets can contain errors, inaccuracies, or noise that don't match the truth of real-world data. This can be caused by problems in the generation algorithm or intentional noise injection. If a model learns these errors, it can make biased or incorrect predictions and won't perform well or reliably when used with real data.

Insufficient Noise

Synthetic data is often too "clean" and doesn't include the natural noise, errors, and variations found in real-world data. Real data is always a bit messy due to various factors. Without enough realistic noise, a model trained on synthetic data might struggle to work effectively in realistic environments.

Over-Smoothing

During generation, some methods might simplify or smooth the data too much, removing the fine details and diversity found in real datasets. This can make it difficult for a model to learn and understand the complex variations that exist in genuine data.

Neglecting Time and Dynamics

Some synthetic data generation methods don't adequately capture how data changes over time or its dynamic behaviours, which are very important in real datasets. If the synthetic data doesn't accurately simulate these time-based patterns, models trained on it might not be effective in real-world applications where timing and change are crucial.

Lack of Inconsistency

Synthetic datasets often lack the natural variations and inconsistencies present in real-world data, which come from different sources, times, or conditions. Because synthetic data might be too uniform, models trained on it can struggle to adapt to the diverse changes originating from different sources or periods, reducing how well they perform (generalize) on varied real datasets.

Data Reliability

The effectiveness of any machine learning or deep learning model relies completely on the quality of its data source. When using synthetic data, its quality is strongly linked to the quality of the original data it's based on and how well the data generation model works. It's crucial to check for biases in the original data, as these can easily carry over into the synthetic data. Therefore, you must thoroughly validate and verify the quality of the synthetic data before using it for predictions.

Handling Outliers

Synthetic data is designed to imitate real-world data but cannot be a perfect copy. Consequently, synthetic datasets might fail to include certain outliers that are present in genuine data. These rare outliers can sometimes hold more significance than typical data points.

Requires Expertise, Time, and Effort

Even though creating synthetic data can sometimes be simpler and less expensive than collecting real data, the process itself still demands a specific level of skill, takes time, and requires significant effort to manage effectively.

Gaining User Acceptance

Synthetic data is a relatively new idea, and people who are unfamiliar with its benefits might be hesitant to trust predictions made by models trained on it. To increase user acceptance, it's necessary to first educate people about the value and potential of synthetic data.

Essential Quality Checks

The purpose of generating synthetic data is to accurately mirror real-world data. Because of this, performing careful quality checks and controlling the output is essential. For complex datasets created automatically by algorithms, it is absolutely necessary to confirm the data's correctness before using it in machine learning or deep learning models.

3.4. Data Generation in the exposome domain

Collecting data for exposome research is a monumental undertaking aimed at capturing the entirety of environmental exposures an individual experienced throughout their life and understanding how these exposures interact with their biological makeup to influence health outcomes. As the exposome encompasses a vast array of factors, ranging from chemical pollutants and physical stressors to lifestyle choices, social environments, and internal biological responses, the importance of collecting comprehensive exposome data lies in its potential to shift our understanding of disease aetiology beyond single-exposure or genetic-centric views. By considering the complex interplay of multiple environmental factors over time, researchers hope to gain deeper insights into the origins of chronic diseases, identify critical periods of susceptibility, and ultimately inform more effective prevention and intervention strategies. Data collection in the exposome domain is often prohibitively complex, costly, and limited by privacy concerns.

Generating artificial or synthetic data is emerging as a valuable approach in exposome research to complement and address the significant challenges associated with collecting and analysing real-world exposome data.

Artificial data generation offers a potential solution by creating synthetic datasets that mimic the statistical properties, patterns, and relationships observed in real exposome data without containing sensitive or identifiable information from actual individuals. This approach can help overcome several limitations of real data and accelerate exposome research by:

- **Addressing Data Scarcity:** Real exposome data, especially for rare exposures or specific populations, can be scarce. Artificial data can be generated to augment existing datasets, providing researchers with larger and more diverse data pools for analysis and model training (Hu et al.,).
- **Protecting Privacy:** Generating synthetic data that retains the statistical characteristics of real data while removing direct identifiers is a crucial benefit for privacy-sensitive exposome studies, particularly those involving personal monitoring data, health records, and genomic information. This facilitates data sharing and collaboration among researchers while mitigating privacy risks (Safarlou et al.,).
- **Enabling Research on Sensitive Topics:** Artificial data can be used to simulate scenarios involving sensitive or stigmatizing exposures or health outcomes, allowing researchers to investigate these areas without the ethical concerns associated with using real data.

- **Facilitating Model Development and Testing:** Large, labelled synthetic datasets can be generated to train and validate complex computational models, including machine learning and AI algorithms, aimed at identifying exposure-disease associations, predicting health risks, and understanding complex exposure mixtures. This is particularly useful for developing robust and generalizable models.
- **Simulating Hypothetical Scenarios:** Artificial data allows researchers to simulate hypothetical exposure scenarios and their potential health impacts, which may be difficult or impossible to study using real-world data. This can aid in risk assessment, policy evaluation, and the exploration of potential interventions.
- **Balancing Datasets:** Real exposome datasets can suffer from class imbalance, where certain exposures or health outcomes are underrepresented. Artificial data can be generated to create more balanced datasets, improving the performance of analytical models.

Various techniques are being explored and adapted for generating artificial exposome data, often leveraging advancements in machine learning and statistical modelling. Refer to paragraph 3.1. Methods of Data Generation for an unexhaustive review of generative methods.

Despite the potential benefits, generating high-quality and reliable artificial data for exposome research presents its own set of challenges:

- **Ensuring Data Realism and Representativeness:** The generated artificial data must accurately reflect the complex distributions, correlations, and temporal dynamics of real exposome data. If the synthetic data does not sufficiently capture the nuances of real exposures, models trained on it may not generalize well to real-world scenarios.
- **Maintaining Data Utility:** While protecting privacy is crucial, the generated data must retain sufficient utility for research purposes. Over-anonymization or inaccurate generation can destroy valuable signals and limit the insights that can be gained.
- **Validating Synthetic Data:** Rigorous methods are needed to validate the quality and representativeness of generated artificial data against real datasets. This involves comparing statistical properties, relationships between variables, and the performance of models trained on both real and synthetic data.
- **Capturing Temporal and Longitudinal Aspects:** The exposome is inherently dynamic and changes over time. Generating synthetic data that accurately captures these temporal patterns and individual exposure trajectories is particularly challenging.
- **Accounting for Causality:** While artificial data can mimic correlations, it is challenging to ensure that causal relationships between exposures and health outcomes are preserved in the generated data, which is crucial for etiological research.
- **Addressing Bias:** If the real data used to train generative models contains biases, these biases can be reflected and potentially amplified in the generated artificial data. Careful attention is needed to detect and mitigate bias in both the training data and the generation process.

In conclusion, generating artificial data holds significant promise for advancing exposome research by providing a means to overcome data limitations, protect privacy, and facilitate the development of sophisticated analytical tools. However, it is crucial to address the technical and methodological challenges to ensure that the generated data is realistic, representative, and suitable for drawing valid scientific conclusions about the complex relationship between the environment and human health.

3.5. Ethical, Legal and Security implications in Data Generation

Data generation presents significant ethical, legal and security implications, also considering that AI technologies can process vast amounts of personal data. While offering significant benefits, the data generation can raise crucial questions about fairness, privacy, intellectual property, and system vulnerabilities.

Ethical Implications:

At the core of data generation ethics is the principle of responsible creation and use.

Generated data can inherit and even amplify biases present in the original datasets used for training generative models. This can lead to not realistic outcomes in applications trained on such data, introducing errors, biases or giving wrong connections between health and exposure to pollutants. Ensuring fairness requires careful attention to the diversity and representativeness of training data and implementing methods to detect and mitigate bias in generated outputs.

Even when aiming to create synthetic data that is not directly traceable to individuals, there's a risk of re-identification, especially if the synthetic data closely mirrors the original dataset or if the original dataset contains unique data points. The potential for inferring sensitive information from generated data, even if anonymized or synthesized, remains a significant ethical challenge.

It is also crucial to be transparent about when and how data is generated, particularly when it is used for decision-making processes that impact individuals. Establishing clear lines of accountability for the quality, biases, and potential harms caused by generated data is essential.

Finally, the ability to generate highly realistic fake data, such as deepfakes or fabricated reports, raises serious ethical concerns about the potential for spreading misinformation, manipulating public opinion, and facilitating fraudulent activities.

Legal Implications:

The legal landscape surrounding data generation is still evolving, grappling with how existing laws apply to novel data creation methods.

Regulations like GDPR place strict requirements on the collection, processing, and storage of personal data. When real data is used to train generative models, ensuring compliance with these laws, particularly regarding consent, data minimization, and the right to be forgotten, is paramount. The application of these laws to synthetic data, especially concerning the risk of re-identification, is an ongoing area of discussion (EDPS, 2025).

The generated data itself, as well as the data used for training, can raise intellectual property issues. Questions arise about the ownership of generated data, whether it can be copyrighted, and whether the use of copyrighted data in training sets constitutes infringement.

Moreover, determining liability for harms caused by systems trained on generated data can be challenging. If biased generated data leads to discriminatory outcomes, or if malicious content is generated, identifying the responsible party (e.g., the data generator, the model developer, or the deploying entity) is complex.

While the increasing use of generated data highlights the need for clearer legal frameworks and regulations specifically addressing data generation practices, including standards for data quality, bias mitigation, and security. The European Union's Data Act, for example, aims to provide more control to users over the data generated by their connected products.

Security implications:

Data generation processes and the resulting datasets can be vulnerable to various security threats:

- **Data Poisoning:** Malicious actors can attempt to poison the training data used for generative models, introducing vulnerabilities or biases into the generated outputs. This is particularly concerning in security-sensitive applications, such as training models for threat detection.
- **Model Theft and Reverse Engineering:** The generative models themselves are valuable assets and can be targets for theft or reverse engineering, potentially exposing the underlying data or allowing malicious generation of data.
- **Privacy Attacks on Generated Data:** Despite efforts to ensure privacy, generated data can still be susceptible to attacks aimed at re-identifying individuals or inferring sensitive information,

especially if the data generation process does not employ robust privacy-preserving techniques.

- **Security of the Generation Infrastructure:** The systems and platforms used for data generation must be secured against unauthorized access, data breaches, and tampering to ensure the integrity and confidentiality of both the training data and the generated data.
- **Use of Generated Data for Malicious Purposes:** As mentioned earlier, the ability to generate realistic fake data can be exploited for malicious activities such as creating deepfakes for impersonation or disinformation, generating malicious code, or automating phishing attacks.

Addressing these ethical, legal, and security implications, especially in the exposome domain, requires a multi-faceted approach involving the development of ethical guidelines, clear legal frameworks, robust security measures, and a commitment to responsible innovation in data generation practices.

4. Data Collection and Generation in EIRENE

The EIRENE RI will be founded on newly developed computational, data interpretation and modelling tools provided through the Knowledge Platform and specific Virtual Labs, operated at multiple research clouds.

To make an effective and holistic analysis possible of data on exposure, susceptibility and effect markers together with data on environmental and social stressors, food basket and consumer product contamination, high quality data managed following interoperable standards are required. Such data can then be processed with bioinformatics tools, omics-based approaches, remote sensing technologies, GIS-based computational platforms and exposure models using advanced artificial intelligence and machine learning methods.

As of its conceptual definition the exposome addresses its assessments by means of multidisciplinary data, which include for example factors indoor and outdoor environment, socioeconomics, lifestyle, and the individual's ability to cope with various stressors such as infection or stress.

The data used in EIRENE RI can be classified into three major categories:

- **Environmental:** (monitoring data from measurement sites/campaigns). Data acquisition should be carefully and extensively documented, according to the FAIR principles. It is not expected that this data will be sensitive, but the data owner/provider must be clearly identifiable.
- **Human studies:** (statistics from population-based studies on morphometry, health and epidemiology). Data access will be via national nodes to ensure that all appropriate data protection processes comply with international and national legislation and that ethics requirements are met.
- **Laboratory Services/Capabilities:** (data generated in EIRENE physical and virtual laboratories) this is effectively a shop window for the infrastructure's technical expertise and capacity. Visiting scientists will need to be made aware of the local/national legislation which covers the work they do in other countries.

In detail, they can include environmental, social, economic and health disciplines (**table 4**).

Table 4. Possible data categories generated or collected in EIRENE.

Data Category	Notes	Examples
---------------	-------	----------

Environmental (Earth Observation)	Include monitoring data on chemicals and ancillary parameters, which will have the geographical component.	pollution levels, land-use change, water quality, soil quality, vegetation
Human studies (Census)	Includes demographic and social statistics.	population, fertility, mortality
Human studies: Laboratory analyses (Nucleotide sequence-based)	Including genomics, epigenomics, metagenomics and transcriptomics.	genome, gene and transcript sequences
Human studies: Laboratory analyses (Biological and biochemical markers)	Includes nucleic acid-based biomarkers.	gene mutations or polymorphisms and quantitative gene expression analysis, peptides, proteins, lipids metabolites
Human studies: Health & lifestyle	Including health, lifestyle and nutrition, social environment, and psychology and stress.	health status, lifestyle
Human studies: External and internal exposure	Include modelled environmental data based on location profile of data subject, human biomonitoring data assessing internal chemical (and metabolite) exposure	Exposure to a particular chemical or chemical group

Within this EIRENE PPP a survey was carried out among the participating national nodes to capture available datasets and services currently offered and that will participate as Virtual Access.

The survey brought a comprehensive catalog of services, which were categorized under the six EIRENE pillars: Chemical Profiling, Toxicological Profiling, Biological Profiling, Environmental Data & Samples, Human Data & Samples, and Tools (see diagram below), and according to mode/s of access. **Figure 2** reports a summary of EIRENE Services by pillar and service type that will be offered. The full preliminary list of these services is reported in **Annex 1**.

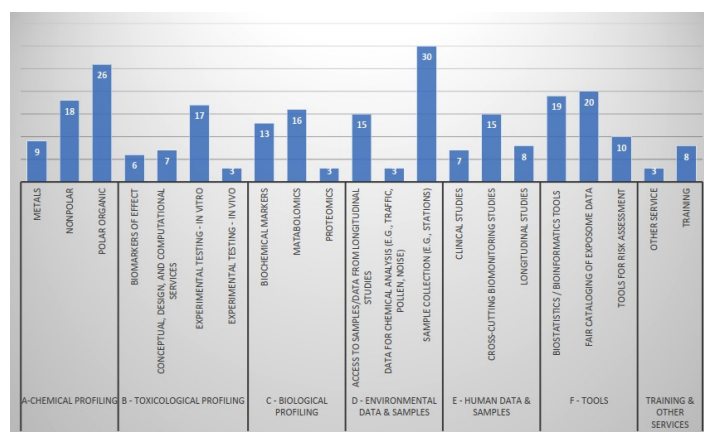


Figure 2. EIRENE Services by Pillar & Service Type

5. Conclusions

The landscape of data access methods and technologies is undergoing a profound transformation, evolving from direct, low-level database interactions to highly abstracted, distributed, and intelligent paradigms. EIRENE RI will strongly benefit from this change of paradigm: the advent of APIs and Object-Relational Mappers (ORMs) make possible abstracting the underlying data sources and allowing applications to interact with data as services or objects. This progression can empower

users, accelerating research cycles, and facilitating the rise of microservices architectures, fundamentally changing how data is used and managed within applications. Also, the proliferation of diverse data storage environments—from traditional relational databases to flexible NoSQL systems, expansive data lakes, and scalable cloud and serverless platforms—has further diversified data access strategies. Each environment presents unique characteristics regarding schema enforcement (from rigid schema-on-write to agile schema-on-read), scalability, and consistency models. The increasing adoption of cloud platforms, in particular, has emerged as a unifying layer, abstracting infrastructure complexities and offering managed services that simplify data access and provide sophisticated access control mechanisms. Serverless databases represent the pinnacle of this abstraction, offering elastic scalability and a pay-as-you-go model, further streamlining data management. EIRENE RI can adopt hybrid architectures that blend batch and real-time capabilities to achieve both comprehensive historical analysis and immediate operational awareness, enabling a more holistic and timely understanding of business operations.

Finally, secure data access remains a critical, non-negotiable aspect of EIRENE RI data strategy. The foundational principles of Confidentiality, Integrity, and Availability, enforced through robust authentication, authorization, and encryption mechanisms (see proposed solution in *D3.4 - Data Access and Sharing*), are paramount. The principle of least privilege, ensuring users and systems have only the minimum necessary access, is a cornerstone of effective authorization, significantly reducing potential attack surfaces. The dynamic and distributed nature of modern data environments, especially in cloud and serverless contexts, necessitates an adaptive and automated approach to security. Continuous monitoring, regular updates, and a strong emphasis on employee education are vital to maintaining a robust security posture against evolving threats.

Looking ahead, the evolution of data access will likely continue its trajectory towards greater abstraction, automation, and intelligence and the EIRENE RI might be driven by such evolution. Artificial intelligence and machine learning will also play an increasingly significant role in optimizing data access performance, predicting bottlenecks, and automating security responses. The shift towards cloud-native and serverless architectures will deepen, requiring to fully embrace the shared responsibility model and develop specialized security practices for these multi-domain environments. The concept of federated data governance, as championed by data mesh architectures, will become more prevalent, allowing to balance centralized standards with domain-specific autonomy.

6. References

- Hao S., Han W., Jiang T., Li Y., Wu H., Zhong C., Zhou Z., Tang H. 2024. Synthetic data in AI: Challenges, applications, and ethical implications. arXiv preprint arXiv:2401.01629.
- Hu et al., 2022. Methodological Challenges in Spatial and Contextual Exposome-Health Studies. *Crit Rev Environ Sci Technol.* 53(7):827–846. doi: 10.1080/10643389.2022.2093595
- Safarlou et al., 2023. The ethical aspects of exposome research: a systematic review. *Exposome*, 3(1): osad004. doi: 10.1093/exposome/osad004
- EDPS, 2025. European Data Protection Supervisor. Synthetic Data. Available: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en
- Hu J., Bowen C.M., 2024. Advancing microdata privacy protection: A review of synthetic data methods. *Wiley Interdisciplinary Rev.: Comput. Statist.*, vol. 16, no. 1, 2024, Art. no. e1636.

Annex 1 - List of Services

Note that some services are shown more than once since multiple installations sometimes offer the same service. The installations are not listed to preserve space.

Country	Institution	Installation / Core facility	Pillar	Service Type	Description of Service	Mode of Access
Belgium	KU Leuven	Human Sentinel Surveillance Platform (https://www.humansentinel.net/en/about): for training and capacity building	F - Tools	Tools for risk assessment	Training, capacity building, and e-learning are integral components of the Human Sentinel Surveillance Platform (HSSP), designed to support structured and scientifically harmonized exposure data collection in occupational and environmental health. This training is essential for strengthening workforce capacity and enabling effective use of the HSSP platform as a full-service research infrastructure. It ensures methodological coherence across surveillance activities and supports the development of standardized protocols for exposure monitoring.	Virtual
Belgium	KU Leuven	Human Sentinel Surveillance Platform (https://www.humansentinel.net/en/about): for training and capacity building	Training & Other Services	Training	Training, capacity building, and e-learning are integral components of the Human Sentinel Surveillance Platform (HSSP), designed to support structured and scientifically harmonized exposure data collection in occupational and environmental health. This training is essential for strengthening workforce capacity and enabling effective use of the HSSP platform as a full-service research infrastructure. It ensures methodological coherence across surveillance activities and supports the development of standardized protocols for exposure monitoring.	Virtual
Belgium	KU Leuven	Human Sentinel Surveillance Platform (https://www.humansentinel.net/en/about): Creating network from OHS professionals, providing large scale biomonitoring data sampling and risk assessment and health surveillance	E - Human Data & Samples	Cross-cutting biomonitoring studies	The Human Sentinel Surveillance Platform (HSSP)-(https://www.humansentinel.net/en/about) backed by a trained collaborative network of OSH professionals in Belgium, uniquely positioned to conduct exposure studies across Belgium, offering a scalable and technologically advanced system designed to address both current and emerging environmental and workplace challenges. At its core, HSSP aims to enable large-scale data and biomonitoring sampling and advanced data analysis, raw data reporting and generating high-resolution insights into exposure patterns and associated health risks. HSSP is a fast-access service platform for policymakers, research institutions, and stakeholders to reach targeted populations via the trained network that enables comprehensive ergonomic, physical, chemical, and psychosocial risk evaluations.	Virtual
Belgium	KU Leuven	LC-MS/MS, GC-FID, GS-MS	E - Human Data & Samples	Cross-cutting biomonitoring studies	Analyzing VOCs in air samples. A key asset is the in-house developed and validated GC-FID method enabling rapid, single-run analysis of approximately 188 VOCs—including aliphatic, aromatic, halogenated hydrocarbons, esters, ketones, glycol ethers, and alcohols—used routinely in occupational hygiene monitoring and exposure assessment studies (e.g., dermal exposure, glove permeation, and measurement campaigns)	Remote
Belgium	UAntwerpen	LC-MS/MS	E - Human Data & Samples	Cross-cutting biomonitoring studies	Developed and validated in-house target methods for quantification of several classes of compounds such as metabolites of plasticizers (phthalates and alternatives), organophosphate flame retardants (PFRs), bisphenols, current-use pesticides, etc, using a LC-MS/MS platform. such methods are routinely used in used routinely in exposure	Remote

					assessment studies (e.g., national and international cohort studies)	
Belgium	UAntwerpen	LC-MS/MS	F - Tools	Fair cataloging of exposome data	Developed and validated in-house target methods for quantification of several classes of compounds such as metabolites of plasticizers (phthalates and alternatives), organophosphate flame retardants (PFRs), bisphenols, current-use pesticides, etc, using a LC-MS/MS platform. such methods are routinely used in used routinely in exposure assessment studies (e.g., national and international cohort studies)	Remote
Belgium	UAntwerpen	LC-MS/MS, LC-HRMS	E - Human Data & Samples	Cross-cutting biomonitoring studies	Services include Suspect Screening Analysis (SSA) and Non-Targeted Analysis (NTA) approaches for the identification of known and emerging substances in biological and environmental matrices.	Remote
Belgium	UAntwerpen	LC-MS/MS, LC-HRMS	F - Tools	Fair cataloging of exposome data	Services include Suspect Screening Analysis (SSA) and Non-Targeted Analysis (NTA) approaches for the identification of known and emerging substances in biological and environmental matrices.	Remote
Belgium	UAntwerpen	LC-MS/MS	F - Tools	Tools for risk assessment	The in-house targeted methods based on LC-MS/MS currently under development will enable participation in hypothesis-driven studies, to measure the extent to which specific endogenous metabolites are altered in specific exposure conditions. We will target specific classes of polar metabolites (e.g. organic acids, amino acids), or non-polar lipids (e.g. endocannabinoids, phospholipids, etc)	Remote
Belgium	UAntwerpen	LC-MS/MS, LC-HRMS	F - Tools	Tools for risk assessment	In-house developed and validated untargeted methods based on LC-HRMS and LC-MSMS enable participation in hypothesis-generating studies, to unravel the endogenous metabolites which are altered in specific exposure conditions	Remote
Belgium	UGent	LC-MS/MS, LC-HRMS	F - Tools	Tools for risk assessment	In-house validated methods for targeted analysis of mycotoxins (regulated + emerging) in diverse matrices: food, feed, indoor dust, intravenous blood (serum, plasma, whole blood), microsampling blood (VAMS), urine, faeces and in vitro (cell/fungal/binders) systems.	Remote
Belgium	UGent	LC-MS/MS, LC-HRMS	Training & Other Services	Training	In-house validated methods for targeted analysis of mycotoxins (regulated + emerging) in diverse matrices: food, feed, indoor dust, intravenous blood (serum, plasma, whole blood), microsampling blood (VAMS), urine, faeces and in vitro (cell/fungal/binders) systems.	Remote
Belgium	UGent	GLORIA-GEZONDHEIDSMONITOR	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	The GLORIA-GEZONDHEIDSMONITOR is a core facility at Ghent University. Requests to use the data are based on a procedure and fled according to a fee-for-data principle.	Virtual
Belgium	UGent	GLORIA-GEZONDHEIDSMONITOR	E - Human Data & Samples	Longitudinal studies	The GLORIA-GEZONDHEIDSMONITOR is a core facility at Ghent University. Requests to use the data are based on a procedure and fled according to a fee-for-data principle.	Virtual
Belgium	VITO	OLINK	E - Human Data & Samples	Cross-cutting biomonitoring studies	OLINK Probe-panel and Amplification-based targeted protein quantification (Target 48, Target 96 and Flex panels)	Remote
Belgium	VITO	European HBM dashboard and Tools for harmonized HBM data reporting. See https://hbm.vito.be/	E - Human Data & Samples	Cross-cutting biomonitoring studies	Harmonization, validation, and calculation of derived variables and summary statistics for Human Biomonitoring data in a consistent way according to a codebook agreed by the community. Reporting of summary statistics in European HBM dashboard and in IPCHEM platform by EU Commission.	Virtual
Belgium	VITO	European HBM dashboard and Tools for harmonized HBM data reporting. See https://hbm.vito.be/	F - Tools	biostatistics / bioinformatics tools	Harmonization, validation, and calculation of derived variables and summary statistics for Human Biomonitoring data in a consistent way according to a codebook agreed by the community. Reporting of summary statistics in European HBM dashboard and in IPCHEM platform by EU Commission.	Virtual

Belgium	VITO	Tools for harmonized HBM data reporting, and Personal Exposure and Health data platform. See https://hbm.vito.be/	E - Human Data & Samples	Cross-cutting biomonitoring studies	GDPR-compliant storage and FAIRification of Human Biomonitoring data.	Virtual
Belgium	VITO	Tools for harmonized HBM data reporting, and Personal Exposure and Health data platform. See https://hbm.vito.be/	F - Tools	biostatistics / bioinformatics tools	GDPR-compliant storage and FAIRification of Human Biomonitoring data.	Virtual
Belgium	VITO	Personal Exposure and Health data platform . See https://hbm.vito.be/	E - Human Data & Samples	Cross-cutting biomonitoring studies	Provide access to extended HBM datasets (incl. exposure data, health data, behaviour, etc) on an individual level.	Virtual
Belgium	VITO	Personal Exposure and Health data platform . See https://hbm.vito.be/	F - Tools	Fair cataloging of exposome data	Provide access to extended HBM datasets (incl. exposure data, health data, behaviour, etc) on an individual level.	Virtual
Belgium	VITO	LC-MS/MS, Including semi-automated sample handling	E - Human Data & Samples	Cross-cutting biomonitoring studies	Non-targeted analysis (NTA), Suspect screening (SS)	Remote
Belgium	VITO	Terrascope	D - Environmental Data & Samples	Sample collection (e.g., stations)	Sentinel earth observation, NO2, CO, CH2O, CH4, SO2 measurements. Land use and environmental aspects (water, air quality, geo cond, urban & green env)	Virtual
Cyprus	CUT	CYPRUS INTERNATIONAL ISNTITUTE FOR ENVIRONMENTAL AND PUBLIC HEALTH	F - Tools	biostatistics / bioinformatics tools	exposomics tools for repeated measured to observe changes in biomarkers of exposure/effect	Virtual
Cyprus	CUT	CYPRUS INTERNATIONAL ISNTITUTE FOR ENVIRONMENTAL AND PUBLIC HEALTH	E - Human Data & Samples	Cross-cutting biomonitoring studies	human data/samples on height, weight, BMI, waist circumference, blood pressure, 3-day nutritional analysis (dietary recall based), general urine analysis in children (hepatic and kidney function)	Hybrid
Cyprus	CUT	CYPRUS INTERNATIONAL ISNTITUTE FOR ENVIRONMENTAL AND PUBLIC HEALTH	E - Human Data & Samples	Longitudinal studies	human data/samples on height, weight, BMI, waist circumference, blood pressure, 3-day nutritional analysis (dietary recall based), general urine analysis in children (hepatic and kidney function)	Hybrid
Cyprus	CUT	CYPRUS INTERNATIONAL ISNTITUTE FOR ENVIRONMENTAL AND PUBLIC HEALTH	E - Human Data & Samples	Clinical studies	human data/samples on height, weight, BMI, waist circumference, blood pressure, 3-day nutritional analysis (dietary recall based), general urine analysis in children (hepatic and kidney function)	Hybrid
Cyprus	CUT	CYPRUS INTERNATIONAL ISNTITUTE FOR ENVIRONMENTAL AND PUBLIC HEALTH	D - Environmental Data & Samples	Sample collection (e.g., stations)	indoor air analysis temp, noise, CO2, PM, VOC	Physical
Czech Republic	Masaryk University - EIRENE-CZ	Central Lab - TAL	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Environmental monitoring networks - Air and water monitoring networks - from sampling plan through establishment to maintenance and coordination of long-term activities. Sampling of the environmental matrices - air (including different particle fractions), water, sediments, soils, biota.	Hybrid
Czech Republic	Masaryk University - EIRENE-CZ	Data services - BAL	F - Tools	biostatistics / bioinformatics tools	Spectrometric data analysis and processing - Galaxy Tools for untargeted mass spectroscopy analyses. https://doi.org/10.5281/zenodo.10612856 , https://doi.org/10.5281/zenodo.13692108	Hybrid
Czech Republic	Masaryk University - EIRENE-CZ	Central Lab - MGL	F - Tools	biostatistics / bioinformatics tools	Analysis of microbial data - biostatistic and bioinformatics tools.	Hybrid
Czech	Masaryk	Population Studies	E - Human Data & Samples	Longitudinal	CELSPEC CS - platform includes an extensive epidemiological database, which contains	Hybrid

Republic	University - EIRENE-CZ		Samples	studies	data from different types of cohorts from the Central Europe region: The Next Generation (TNG) birth cohort, Young Adults cohort, Prospective seroconversion coronavirus cohort (PROSECO), and other cross-sectional biomonitoring studies. Laboratory service - Human samples processing including DNA/RNA isolation. Biobanking - collection, automatic processing and long-term storage of biological sample aliquots.	
Czech Republic	Masaryk University - EIRENE-CZ	Population Studies	E - Human Data & Samples	Cross-cutting biomonitoring studies	CELSPAC CS - platform includes an extensive epidemiological database, which contains data from different types of cohorts from the Central Europe region: The Next Generation (TNG) birth cohort, Young Adults cohort, Prospective seroconversion coronavirus cohort (PROSECO), and other cross-sectional biomonitoring studies. Laboratory service - Human samples processing including DNA/RNA isolation. Biobanking - collection, automatic processing and long-term storage of biological sample aliquots.	Hybrid
Czech Republic	Masaryk University - EIRENE-CZ	Data services	F - Tools	Tools for risk assessment	portals and tools, access to databases and information system: GENASIS - https://www.genasis.cz/ - global IS providing a comprehensive information on environmental contamination.	Hybrid
France	EHESP	Breizh Exposome	F - Tools	biostatistics / bioinformatics tools	Annotation of LC-HRMS data	Virtual
France	ONIRIS	LABERCA/HBM PLATFORM	F - Tools	biostatistics / bioinformatics tools	Annotation of LC-HRMS data	Virtual
Germany	UFZ	Other services - Training	Training & Other Services	Training	High throughput screening in vitro bioassays	Remote
Italy	CNR	Trace Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Access to environmental samples on trace metal and ionic species; carbonaceous aerosol fractions in particulate matter for classifying Organic (OC) and Elemental (EC) Carbon.	Physical
Italy	CNR	Trace Lab	F - Tools	Fair cataloging of exposome data	Access to environmental samples on trace metal and ionic species; carbonaceous aerosol fractions in particulate matter for classifying Organic (OC) and Elemental (EC) Carbon.	Physical
Italy	CNR	Trace Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Monte Curcio (MCU) GAW Environmental-Climate Observatory; Italian National Network "Reti speciali" measuring all criteria air pollutants; possibilities to provide on-site training to technicians and scientists.	Physical
Italy	CNR	Trace Lab	F - Tools	Fair cataloging of exposome data	Monte Curcio (MCU) GAW Environmental-Climate Observatory; Italian National Network "Reti speciali" measuring all criteria air pollutants; possibilities to provide on-site training to technicians and scientists.	Physical
Italy	CNR	Trace Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	User-friendly platform to access to online data from the Monte Curcio GAW site and from "Reti Speciali" national network.	Virtual
Italy	CNR	Trace Lab	F - Tools	Fair cataloging of exposome data	User-friendly platform to access to online data from the Monte Curcio GAW site and from "Reti Speciali" national network.	Virtual
Italy	CNR	Trace Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Ion chromatography (IC); Inductively Coupled Plasma-tandem Mass Spectrometry (ICP-MS/MS); Ion Chromatography followed by Inductively Coupled Plasma-tandem Mass Spectrometry (IC-ICP-MS/MS);TD-AAS; Thermal-Optical Transmittance (TOT) methods.	Physical
Italy	CNR	Trace Lab	D - Environmental	Sample	Quantification of heavy metal and ionic species; Determination of carbonaceous aerosol	Physical

			Data & Samples	collection (e.g., stations)	fractions in particulate matter for classifying Organic (OC) and Elemental (EC) Carbon.	
Italy	CNR	Mercury Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	High throughput measurement of total and speciated Hg in environmental (air, water, soil, biota, waste) matrices.	Physical
Italy	CNR	Mercury Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Global Mercury Observation System GMOS/GOS4M network with over 40 sites in both southern and northern hemispheres. Access to the GOS4M platform for data handling/automated QA/QC and storage of raw data.	Virtual
Italy	CNR	Mercury Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Access to GOS4M data / platform that allows to access to historical data since 2012 from satellite, off-shore and in-situ monitoring platforms	Virtual
Italy	CNR	Mercury Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Direct Thermal Decomposition – Gold Amalgamation – Cold Vapor Atomic Absorption Spectroscopy (CVAAS); Cold Vapor Atomic Fluorescence Spectrometry (CVAFS).	Physical
Italy	CNR	Mercury Lab	D - Environmental Data & Samples	Sample collection (e.g., stations)	Quantification of total and speciated Hg in environmental (air, water, soil, biota, waste) and biological (urine, hair, human breast milk) matrices.	Physical
Italy	CNR	GOS4M Knowledge Hub	D - Environmental Data & Samples	Sample collection (e.g., stations)	Online catalog (metadata) for available datasets on mercury in the atmosphere, oceans and marine biota, as well as ancillary parameters, tool for discovery and download datasets.	Virtual
Italy	CNR	GOS4M Knowledge Hub	F - Tools	Fair cataloging of exposome data	Online catalog (metadata) for available datasets on mercury in the atmosphere, oceans and marine biota, as well as ancillary parameters, tool for discovery and download datasets.	Virtual
Italy	ISS	Trace Lab Metals	D - Environmental Data & Samples	Sample collection (e.g., stations)	Analysis of human samples (serum, urine, blood, exhaled breath condensate, hair, dermal wipes) for metals including Hg, nanoparticles of metals, species of metals	Physical
Italy	ISS	Trace Lab Metals	D - Environmental Data & Samples	Sample collection (e.g., stations)	Inductively coupled plasma-mass spectrometry (ICP-MS, both ICAp-Q ICP-MS and SF-ICP-MS); field-flow fractionation with multi angle light scattering coupled to inductively plasma mass spectrometry (FFF-MALS-ICP-MS) and Single Particle ICP-MS; high performance liquid chromatography coupled to inductively plasma mass spectrometry (HPLC-ICP-MS), ion chromatography coupled to inductively coupled plasma mass spectrometry (IC-ICP-MS), multicollector coupled to inductively plasma mass spectrometry (MC-ICP-MS); direct mercury analyser (DMA-80)	Physical
Italy	ISS	Trace Lab Metals	D - Environmental Data & Samples	Sample collection (e.g., stations)	Determination of metals including Hg, nanoparticles of metals, species of metals	Physical
Italy	ISS	Trace Lab Organics	D - Environmental Data & Samples	Sample collection (e.g., stations)	Analysis of human samples (blood, urine and breast milk) for organic pollutants including POPs	Physical
Italy	ISS	Trace Lab Organics	D - Environmental Data & Samples	Sample collection (e.g., stations)	High-resolution gas chromatography coupled with high resolution mass spectrometry (HRGC-HRMS); liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS); ultra performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS); non target screening (NTS) with high resolution mass spectrometry coupled to GC and LC modules	Physical
Italy	ISS	Trace Lab Organics	D - Environmental	Sample	Determination of PCDDs, PCDFs, PCBs, pesticides, PBDEs, HBCDs, PFAS, PAHs and their	Physical

			Data & Samples	collection (e.g., stations)	metabolites; NTS	
Italy	ASI	MADS	D - Environmental Data & Samples	Sample collection (e.g., stations)	The Multimission Access Data System (MADS) is a platform that will support the individual Ground Segments by providing to users a cloud-based, unique point of access to products from different missions with possibilities to browse catalogs, to plan new acquisitions, to access data through standard M2M Interfaces, to run their applications on the cloud (users to the data). Currently under development, initial functionalities available in 2024.	Virtual
Italy	ASI	MADS	F - Tools	Fair cataloging of exposome data	The Multimission Access Data System (MADS) is a platform that will support the individual Ground Segments by providing to users a cloud-based, unique point of access to products from different missions with possibilities to browse catalogs, to plan new acquisitions, to access data through standard M2M Interfaces, to run their applications on the cloud (users to the data). Currently under development, initial functionalities available in 2024.	Virtual
Italy	ASI	MADS	D - Environmental Data & Samples	Sample collection (e.g., stations)	Online catalog (metadata) for available standard products from Italian national satellite missions, tool for discovery and download (upon registration) datasets. Currently it is possible to access products from the COSMO-SkyMed, PRISMA and SAOCOM missions from different, dedicated websites and with different modalities. The development of MADS will provide just one user interface to discover and access all data of present and future ASI missions.	Virtual
Italy	ASI	MADS	F - Tools	Fair cataloging of exposome data	Online catalog (metadata) for available standard products from Italian national satellite missions, tool for discovery and download (upon registration) datasets. Currently it is possible to access products from the COSMO-SkyMed, PRISMA and SAOCOM missions from different, dedicated websites and with different modalities. The development of MADS will provide just one user interface to discover and access all data of present and future ASI missions.	Virtual
Italy	ASI	ASI-NPM	D - Environmental Data & Samples	Sample collection (e.g., stations)	The ASI-sponsored NPM (Network for Particulate Measurement) is a component of the MAIA mission and it is made of surface monitors located inside the Italian target areas (see attached file), periodically observed during the MAIA mission, measuring PM2.5 sulphate, nitrate, elemental carbon, organic carbon, and dust (calculated using concentrations of Fe, Al, Ca, Si, and Ti). The network is jointly developed by ASI, CNR and the Regional Environmental Agencies (ARPAs) and it is based on the already available surfaces monitors operated by CNR and the ARPAs whose territory is within the MAIA Target Areas. It will be progressively upgraded to comply with MAIA ground measurement requirements.	Virtual
Italy	ASI	ASI-NPM	F - Tools	Fair cataloging of exposome data	The ASI-sponsored NPM (Network for Particulate Measurement) is a component of the MAIA mission and it is made of surface monitors located inside the Italian target areas (see attached file), periodically observed during the MAIA mission, measuring PM2.5 sulphate, nitrate, elemental carbon, organic carbon, and dust (calculated using concentrations of Fe, Al, Ca, Si, and Ti). The network is jointly developed by ASI, CNR and the Regional Environmental Agencies (ARPAs) and it is based on the already available surfaces monitors operated by CNR and the ARPAs whose territory is within the MAIA Target Areas. It will be progressively upgraded to comply with MAIA ground measurement requirements.	Virtual
Italy	ASI	ASI-NPM	D - Environmental Data & Samples	Sample collection (e.g.,	Access to data from the ASI-sponsored NPM	Virtual

				stations)		
Italy	ASI	ASI-NPM	F - Tools	Fair cataloging of exposome data	Access to data from the ASI-sponsored NPM	Virtual
Italy	ASI	ASI-Air Quality and Health Knowledge Hub	D - Environmental Data & Samples	Sample collection (e.g., stations)	Under development, to provide access to ASI-sponsored projects and investigations on Air Quality monitoring, forecasting and associated effects and risks on population health. Projects for the development of MAIA-related products and prototype services will be activated in 2024.	Virtual
Italy	ASI	ASI-Air Quality and Health Knowledge Hub	F - Tools	Fair cataloging of exposome data	Under development, to provide access to ASI-sponsored projects and investigations on Air Quality monitoring, forecasting and associated effects and risks on population health. Projects for the development of MAIA-related products and prototype services will be activated in 2024.	Virtual
Italy	ASI	HYPERHEALTH Knowledge Hub	D - Environmental Data & Samples	Sample collection (e.g., stations)	HYPERHEALTH is a project being developed (April 2022-April 2024) through a partnership that includes ASI, University of Pisa, CNR, and SiHealth Photonics S.r.l company, centered on the use of PRISMA Hyperspectral data. The ultimate objective is to develop and validate (in a test area in Tuscany) the HyperHealth prototype service (mobile app) providing a PRISMA-based assessment of environmental health risk connected with pollen maps, health-relevant atmospheric components (e.g. CO ₂ , CWV) and solar UV radiation.	Virtual
Italy	ASI	HYPERHEALTH Knowledge Hub	F - Tools	Fair cataloging of exposome data	HYPERHEALTH is a project being developed (April 2022-April 2024) through a partnership that includes ASI, University of Pisa, CNR, and SiHealth Photonics S.r.l company centred on the use of PRISMA Hyperspectral data. The ultimate objective is to develop and validate (in a test area in Tuscany) the HyperHealth prototype service (mobile app) providing a PRISMA-based assessment of environmental health risk connected with pollen maps, health-relevant atmospheric components (e.g. CO ₂ , CWV) and solar UV radiation.	Virtual
Italy	ASI	PRIMARY Knowledge Hub	D - Environmental Data & Samples	Sample collection (e.g., stations)	PRIMARY is a project being developed (April 2022-April 2024) through a partnership that includes ASI, University of Tor Vergata (Rome), CNR, University of L'Aquila, and SERCO company. It is centred on the use of PRISMA Hyperspectral data and neural algorithms for the generation of the products of interest for the Rome urban area (test area). Expected main product is the abundance of chemical species in the aerosol; the current list includes inorganic and organic particulate, Black carbon, mineral dust, marine salt and the mixing ratio (ppm) of the above. The project will also made available a full wealth of ground and airborne measurements, used for the calibration and validation of the satellite-derived products.	Virtual
Italy	ASI	PRIMARY Knowledge Hub	F - Tools	Fair cataloging of exposome data	PRIMARY is a project being developed (April 2022-April 2024) through a partnership that includes ASI, University of Tor Vergata (Rome), CNR, University of L'Aquila, and SERCO company. It is centred on the use of PRISMA Hyperspectral data and neural algorithms for the generation of the products of interest for the Rome urban area (test area). Expected main product is the abundance of chemical species in the aerosol; the current list includes inorganic and organic particulate, Black carbon, mineral dust, marine salt and the mixing ratio (ppm) of the above. The project will also made available a full wealth of ground and airborne measurements, used for the calibration and validation of the satellite-derived products.	Virtual
Netherlands	VU/ A UMC	Population Studies	D - Environmental Data & Samples	Access to samples/data from longitudinal	Access to cohort study or survey data on an individual level	Hybrid

				studies		
Netherlands	UU	Population Studies	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Access to cohort study or survey data on an individual level	Hybrid
Netherlands	A UMC	exposure maps	D - Environmental Data & Samples	Data for chemical analysis (e.g., traffic, pollen, noise)	Databases and exposure maps on environmental factors (e.g. pollutants, temperature, noise, socio-economic, lifestyle)	Remote
Netherlands	UU	exposome maps	D - Environmental Data & Samples	Data for chemical analysis (e.g., traffic, pollen, noise)	Databases and exposure maps on environmental factors (e.g. pollutants, temperature, noise, socio-economic, lifestyle)	Remote
Netherlands	LACDR	FAIR cataloging of MoA data. Unit of access is high throughput transcriptomics datasets per compound or high throughput imaging datasets per compound	F - Tools	Fair cataloging of exposome data	FAIR cataloging of exposome data (e.g. cohorts, algorithms)	Remote
Norway	NIPH	GC- and LC/MS/MS	E - Human Data & Samples	Cross-cutting biomonitoring studies	Occurrence data/ internal exposure in human samples	Hybrid
Norway	NIPH	GC- and LC/MS/MS	F - Tools	Tools for risk assessment	Occurrence data/ internal exposure in human samples	Hybrid
Norway	NIPH	GC- and LC/MS/MS	E - Human Data & Samples	Longitudinal studies	Risk assessment	Hybrid
Norway	NIPH	GC- and LC/MS/MS	F - Tools	Fair cataloging of exposome data	Risk assessment	Hybrid
NORWAY	NILU	GC- and LC/MS/MS	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Occurrence data/ external exposure in environmental samples	Hybrid
NORWAY	NILU	GC- and LC/MS/MS	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Occurrence data/ external exposure in environmental samples	Hybrid
NORWAY	NILU	ICP-MS	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Occurrence data/ external exposure in environmental samples	Hybrid
Norway	NIPH	?	E - Human Data & Samples	Cross-cutting biomonitoring studies	Exposure estimates	Hybrid
Slovenia	JSI	JSI/O2	D - Environmental Data & Samples	Access to samples/data from longitudinal	Chemical analysis, biobank, training	Hybrid

				studies		
Slovenia	JSI	JSI/O2	E - Human Data & Samples	Longitudinal studies	Chemical analysis, biobank, training	Hybrid
Slovenia	JSI	JSI/O2	F - Tools	biostatistics / bioinformatics tools	Chemical analysis, biobank, training	Hybrid
Slovenia	JSI	JSI/O2	D - Environmental Data & Samples	Data for chemical analysis (e.g., traffic, pollen, noise)	chemical analysis, training biobank	Hybrid
Slovenia	JSI	JSI/O2	E - Human Data & Samples	Longitudinal studies	chemical analysis, training biobank	Hybrid
Slovenia	JSI	JSI/O2	F - Tools	biostatistics / bioinformatics tools	chemical analysis, training biobank	Hybrid
Slovenia	JSI	JSI/O2	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	chemical analyses, training, biobank	Hybrid
Slovenia	JSI	JSI/O2	E - Human Data & Samples	Longitudinal studies	chemical analyses, training, biobank	Hybrid
Slovenia	JSI	JSI/O2	F - Tools	biostatistics / bioinformatics tools	chemical analyses, training, biobank	Hybrid
Slovenia	JSI	JSI/O2	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	chemical analysis, training, biobank	Hybrid
Norway	NIPH	?	F - Tools	Tools for risk assessment	Exposure estimates	Hybrid
Slovenia	JSI	JSI/O2	E - Human Data & Samples	Cross-cutting biomonitoring studies	chemical analysis, training, biobank	Hybrid
Slovenia	JSI	JSI/O2	F - Tools	biostatistics / bioinformatics tools	chemical analysis, training, biobank	Hybrid
Spain	IDAEA-CSIC	Barcelona	E - Human Data & Samples	Longitudinal studies	The group has expertise on these types of studies and may organize others at demand	Hybrid
Spain	IDAEA-CSIC	Barcelona	E - Human Data & Samples	Cross-cutting biomonitoring studies	The group has expertise on these types of studies and may organize others at demand	Hybrid
Spain	IDAEA-CSIC	Barcelona	F - Tools	biostatistics / bioinformatics tools	The group may advice on the use of these methods	Hybrid

Spain	IDAEA-CSIC	Barcelona	F - Tools	Tools for risk assessment	The group may advice on the use of these methods	Hybrid
UK	University of Birmingham	Phenome Centre Birmingham	D - Environmental Data & Samples	Sample collection (e.g., stations)	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results. Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.	
UK	University of Birmingham	Phenome Centre Birmingham	F - Tools	biostatistics / bioinformatics tools	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results. Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.	
UK	University of Birmingham	Phenome Centre Birmingham	D - Environmental Data & Samples	Sample collection (e.g., stations)	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results. Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.	
UK	University of Birmingham	Phenome Centre Birmingham	F - Tools	biostatistics / bioinformatics tools	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results. Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.	
UK	University of Birmingham	Phenome Centre Birmingham	D - Environmental Data & Samples	Sample collection (e.g., stations)	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results. Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.	
UK	University of Birmingham	Phenome Centre Birmingham	F - Tools	biostatistics / bioinformatics tools	Access to world renowned experts in the field, state-of-the-art instrumentation, outstanding operational support, and a high level of service including: Advice and support to design and plan metabolomics studies. Remote support / discussion of results.	

					Analysis performed using state-of-the-art mass spectrometry and data analysis tools. Guidance to interpret the study results and ensure biological questions answered. Outstanding operational and scientific support to deliver results within a defined timescale.
UK	University of Birmingham	Phenotypic Screening Platform	F - Tools	biostatistics / bioinformatics tools	Robotic platform for high-throughput analysis of panes of chemicals across a range of cell lines (primary and immortalised, 3D cultures etc) Support in experimental design, sample optimisation, statistical optimisation, etc.
UK	University of Birmingham	Phenotypic Screening Platform	Training & Other Services	Other Service	Robotic platform for high-throughput analysis of panes of chemicals across a range of cell lines (primary and immortalised, 3D cultures etc) Support in experimental design, sample optimisation, statistical optimisation, etc.
UK	University of Birmingham	Centre for Health Data Science - Semantic data Analysis	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Robotic platform for high-throughput analysis of panes of chemicals across a range of cell lines (primary and immortalised, 3D cultures etc) Support in experimental design, sample optimisation, statistical optimisation, etc.
UK	University of Birmingham	Centre for Health Data Science - Semantic data Analysis	E - Human Data & Samples	Clinical studies	Robotic platform for high-throughput analysis of panes of chemicals across a range of cell lines (primary and immortalised, 3D cultures etc) Support in experimental design, sample optimisation, statistical optimisation, etc.
UK	University of Birmingham	Centre for Health Data Science - Semantic data Analysis	F - Tools	Fair cataloging of exposome data	Robotic platform for high-throughput analysis of panes of chemicals across a range of cell lines (primary and immortalised, 3D cultures etc) Support in experimental design, sample optimisation, statistical optimisation, etc.
UK	University of Birmingham	Centre for Health Data Science - Translational phenomics	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	
UK	University of Birmingham	Centre for Health Data Science - Translational phenomics	E - Human Data & Samples	Clinical studies	
UK	University of Birmingham	Centre for Health Data Science - Translational phenomics	F - Tools	Fair cataloging of exposome data	
UK	University of Birmingham	Centre for Health Data Science - Ontologies and Standards	E - Human Data & Samples	Clinical studies	
UK	University of Birmingham	Centre for Health Data Science - Ontologies and Standards	F - Tools	Fair cataloging of exposome data	
UK	University of Birmingham	Centre for Health Data Science - Multi-omics integrative analysis	F - Tools	biostatistics / bioinformatics tools	
UK	University of Birmingham	FAIR Toxicology data management - ontologies, metadata templates, data capture templates, databasing	F - Tools	Tools for risk assessment	
UK	University of Birmingham	FAIR Toxicology data management - ontologies, metadata templates, data capture templates, databasing	Training & Other Services	Training	
UK	University of Birmingham	FAIR Toxicology data management - ontologies, metadata templates, data capture templates, databasing	Training & Other Services	Other Service	

UK	University of Birmingham	Tools for NAMs documentation and validation / databasing	F - Tools	Tools for risk assessment		
UK	University of Birmingham	Tools for NAMs documentation and validation / databasing	Training & Other Services	Training		
UK	University of Birmingham	Sensor networks -Air, water, soil, trees, plants (including CO2 elevated)	D - Environmental Data & Samples	Sample collection (e.g., stations)		
UK	University of Birmingham	Birmingham Clinical Trials Unit	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Late-phase trials: Trial design, including appropriate sample size; Protocol development and case report form design; Research costs and funding applications Establishing collaborative networks of investigators; Gaining research ethics and regulatory approval; Randomisation procedures; Strategies for patient recruitment Trial co-ordination and data management; Research governance and quality assurance procedures; Database design and computerised trial management software; Incorporation of health economics and quality of life measures Independent data monitoring; Statistical analysis and reporting are not possible to answer a question about the effectiveness of an intervention and/or where the quality of data collection and data management is required to be at the standard of that of a randomised controlled trial.	Hybrid
UK	University of Birmingham	Birmingham Clinical Trials Unit	E - Human Data & Samples	Clinical studies	Late-phase trials: Trial design, including appropriate sample size; Protocol development and case report form design; Research costs and funding applications Establishing collaborative networks of investigators; Gaining research ethics and regulatory approval; Randomisation procedures; Strategies for patient recruitment Trial co-ordination and data management; Research governance and quality assurance procedures; Database design and computerised trial management software; Incorporation of health economics and quality of life measures Independent data monitoring; Statistical analysis and reporting are not possible to answer a question about the effectiveness of an intervention and/or where the quality of data collection and data management is required to be at the standard of that of a randomised controlled trial.	Hybrid
UK	University of Birmingham	Birmingham Clinical Trials Unit	F - Tools	biostatistics / bioinformatics tools	Late-phase trials: Trial design, including appropriate sample size; Protocol development and case report form design; Research costs and funding applications Establishing collaborative networks of investigators; Gaining research ethics and regulatory approval; Randomisation procedures; Strategies for patient recruitment Trial co-ordination and data management; Research governance and quality assurance procedures; Database design and computerised trial management software; Incorporation of health economics and quality of life measures Independent data monitoring; Statistical analysis and reporting are not possible to answer a question about the effectiveness of an intervention and/or where the quality of data collection and data management is required to be at the standard of that of a randomised controlled trial.	Hybrid
UK	University of Birmingham	Birmingham Clinical Trials Unit	Training & Other Services	Training	Late-phase trials: Trial design, including appropriate sample size; Protocol development and case report form design; Research costs and funding applications Establishing collaborative networks of investigators; Gaining research ethics and regulatory approval; Randomisation procedures; Strategies for patient recruitment Trial co-ordination and data management; Research governance and quality assurance procedures; Database design and computerised trial management software;	Hybrid

					Incorporation of health economics and quality of life measures Independent data monitoring; Statistical analysis and reporting are not possible to answer a question about the effectiveness of an intervention and/or where the quality of data collection and data management is required to be at the standard of that of a randomised controlled trial.	
UK	University of Birmingham	Birmingham Clinical Trials Unit	Training & Other Services	Other Service	Late-phase trials: Trial design, including appropriate sample size; Protocol development and case report form design; Research costs and funding applications Establishing collaborative networks of investigators; Gaining research ethics and regulatory approval; Randomisation procedures; Strategies for patient recruitment Trial co-ordination and data management; Research governance and quality assurance procedures; Database design and computerised trial management software; Incorporation of health economics and quality of life measures Independent data monitoring; Statistical analysis and reporting are not possible to answer a question about the effectiveness of an intervention and/or where the quality of data collection and data management is required to be at the standard of that of a randomised controlled trial.	Hybrid
UK	University of Birmingham	Birmingham Health Partners (BHP) Centre for Regulatory Science and Innovation	D - Environmental Data & Samples	Access to samples/data from longitudinal studies		Hybrid
UK	University of Birmingham	Birmingham Health Partners (BHP) Centre for Regulatory Science and Innovation	E - Human Data & Samples	Clinical studies		Hybrid
UK	University of Birmingham	Birmingham Health Partners (BHP) Centre for Regulatory Science and Innovation	F - Tools	biostatistics / bioinformatics tools		Hybrid
UK	University of Birmingham	Birmingham Health Partners (BHP) Centre for Regulatory Science and Innovation	Training & Other Services	Training		Hybrid
UK	University of Birmingham	Human Biomaterials Resource Centre (HBRC)	D - Environmental Data & Samples	Access to samples/data from longitudinal studies	Researchers may apply to access samples already stored, or to set up bespoke collections linked to new or ongoing clinical trials / patient interventions. High quality storage space for large collections. These collections may, or may not be stored under the HTA licence depending upon the types of biomaterials included and the status of any existing ethical approvals. A hosting service agreement for each collection.	
UK	University of Birmingham	Human Biomaterials Resource Centre (HBRC)	E - Human Data & Samples	Clinical studies	Researchers may apply to access samples already stored, or to set up bespoke collections linked to new or ongoing clinical trials / patient interventions. High quality storage space for large collections. These collections may, or may not be stored under the HTA licence depending upon the types of biomaterials included and the status of any existing ethical approvals. A hosting service agreement for each collection.	
UK	University of Birmingham	Human Biomaterials Resource Centre (HBRC)	F - Tools	Fair cataloging of exposome data	Researchers may apply to access samples already stored, or to set up bespoke collections linked to new or ongoing clinical trials / patient interventions. High quality storage space for large collections. These collections may, or may not be stored under the HTA licence depending upon the types of biomaterials included and the status of any existing ethical approvals. A hosting service agreement for each collection.	
UK	University of Birmingham	Human Biomaterials Resource Centre	Training & Other Services	Training	Researchers may apply to access samples already stored, or to set up bespoke	

Birmingham	(HBRC)	Services	collections linked to new or ongoing clinical trials / patient interventions. High quality storage space for large collections. These collections may, or may not be stored under the HTA licence depending upon the types of biomaterials included and the status of any existing ethical approvals. A hosting service agreement for each collection.
------------	--------	----------	--