



## Data management and protection in occupational and environmental exposome research - A case study from the EU-funded EXIMIOUS project

Manosij Ghosh<sup>a,\*</sup>, Katrijn Broothaerts<sup>a</sup>, Steven Ronsmans<sup>a</sup>, Ingrid Barcena Roig<sup>b</sup>, Jef Scheepers<sup>b</sup>, Mustafa Dikmen<sup>b</sup>, Emily Rose Ciscato<sup>c</sup>, Carolina Blanch<sup>d</sup>, Michelle Plusquin<sup>e</sup>, Unni C. Nygaard<sup>f</sup>, Camilla Sandal Sejbæk<sup>g</sup>, Karin S. Hougaard<sup>h,i</sup>, Peter HM. Hoet<sup>a</sup>, On behalf of the EXIMIOUS consortium

<sup>a</sup> Centre for Environment and Health, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

<sup>b</sup> Support for Research Data Management (RDM), KU Leuven, Leuven, Belgium

<sup>c</sup> accelopment Schweiz AG, Zurich, Switzerland

<sup>d</sup> Imec (Interuniversitair Microelectronica Centrum), Leuven, Belgium

<sup>e</sup> Centre for Environmental Sciences, University of Hasselt, Hasselt, Belgium

<sup>f</sup> Section for Immunology, Division of Infection Control, Norwegian Institute of Public Health, Oslo, Norway

<sup>g</sup> Copenhagen University Hospital - Bispebjerg and Frederiksberg Hospital, Department of Occupational and Environmental Medicine, Copenhagen, Denmark

<sup>h</sup> National Research Centre for the Working Environment, Copenhagen, Denmark

<sup>i</sup> Department of Public Health, University of Copenhagen, Copenhagen, Denmark

### ARTICLE INFO

Handling editor: Jose L Domingo

#### Keywords:

Data management  
Data protection  
Ethics  
GDPR  
Exposome  
Sensitive information

### ABSTRACT

Within collaborative projects, such as the EU-funded Horizon 2020 EXIMIOUS project (Mapping Exposure-Induced Immune Effects: Connecting the Exposome and the Immunome), collection and analysis of large volumes of data pose challenges in the domain of data management, with regards to both ethical and legal aspects. However, researchers often lack the right tools and/or accurate understanding of the ethical/legal framework to independently address such challenges. With the guidance and support within and between the partner institutes (the researchers and the ethical and legal teams) in the EXIMIOUS project, we have been able to understand and solve most challenges during the first two project years. This has fed into the development of a Data Management Plan and the establishment of data management platforms in accordance with the ethical and legal framework laid down by the EU and the different national regulations of the partners involved. Through this elaborate exercise, we have acquired tools which allow us to make our research data FAIR (Findable, Accessible, Interoperable, and Reusable), while at the same time ensuring data privacy and security (GDPR compliant). Herein we share our experience of creating and managing the data workflow through an open research communication, with the aim of helping other researchers build their data management framework in their own projects. Based on the measures adopted in EXIMIOUS to ensure FAIR data management, we also put together a checklist “DMP CHECK” containing a series of recommendations based on our experience.

### 1. Background

Open data can maximize the impact of research findings and

promote innovation. Open data is broadly defined as “*data that can be accessed and reused by anyone without technical or legal restrictions*” (“[Enhanced Access to Publicly Funded Data for Science, Technology and](#)

**Abbreviations:** CA, Consortium Agreement; CPT, Cell preparation tube; DEXA, Dual-energy X-ray absorptiometry; DMP, Data Management Plan; DTA, Data Transfer Agreement; EC, European Commission; EDTA, Ethylenediaminetetraacetic acid; EHEN, European Human Exposome Network; FAIR, Findable, Accessible, Interoperable, and Reusable; GA, Grant Agreement; GC-FID, Gas Chromatography with flame ionization detection; GC-MS, Gas chromatography–mass spectrometry; GDPR, General Data Protection Regulation; GWAS, Genome wide association study; miRNA, micro RNA; MTA, Material Transfer Agreement; ORD, Open Research Data; PBMC, peripheral blood mononuclear cells; SOP, Standard Operating Procedure; VOC, Volatile organic compounds; WPs, Work Packages; H2020, Horizon 2020.

\* Corresponding author.

E-mail addresses: [gmanosij@gmail.com](mailto:gmanosij@gmail.com), [manosij.ghosh@kuleuven.be](mailto:manosij.ghosh@kuleuven.be) (M. Ghosh), [peter.hoet@kuleuven.be](mailto:peter.hoet@kuleuven.be) (P.HM. Hoet).

<https://doi.org/10.1016/j.envres.2023.116886>

Received 22 November 2022; Received in revised form 25 May 2023; Accepted 12 August 2023

Available online 18 August 2023

0013-9351/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

Innovation," 2020). However, to balance the openness of data and address privacy and security concerns, the European Commission (EC) only requires the data to be "as open as possible, as closed as necessary".<sup>1</sup> A key step towards making data open was established by the open research data (ORD) pilot of the EC, which enables open access and reuse of data from H2020 projects. The main pillar of such pilot includes the development (and updating) of a Data Management Plan (DMP) and provision of (restricted) access to data within a safe environment.

The first key challenge faced in developing a robust Data Management Plan for large multi-disciplinary projects such as EXIMIOUS "Mapping Exposure-Induced Immune Effects: Connecting the Exposome and the Immunome" () is to get a good overview of the large amount of data (sources and types) and the data flow. Additional challenges exist in the domain of data management and sharing with regards to ethical and legal aspects. Researchers are most often not equipped with the right tools and/or accurate understanding of the ethical/legal framework. Within the EXIMIOUS project, we were able to address many of these challenges through discussion between legal and ethical experts, data managers and researchers. Through this elaborate and complex exercise, we acquired tools which allow us to make our research data FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016), while at the same time ensuring data privacy and security (GDPR [General Data Protection Regulation] compliant) ("General Data Protection Regulation (GDPR) – Official Legal Text," n.d.). This paper aims to summarize the process followed during the first two years of the project towards developing a robust Data Management Plan and making data open. By sharing our experience of creating and managing data workflow through this open research communication, we hope to provide help to other researchers in building their own data management framework.

## 2. The EXIMIOUS project

The overall objective of the EXIMIOUS project (Ronsmans et al., 2022) is to establish new ways of assessing the human exposome. Specifically, innovative ways of characterizing and quantifying multiple environmental exposures (exposomics) mapped against immune effects (immunomics), to identify exposure related components of immune mediated disease development. As part of the project, new bioinformatics tools will be developed using systems biology, artificial intelligence and machine learning, first combining and then analyzing the huge datasets that link the exposome, immunome, and other omics data with clinical and socio-economic data (Ronsmans et al., 2022). By exploring the entire pathway from the exposome, over immune fingerprints, to disease throughout the life course (including prenatal exposures), we aim to better understand the factors that lead to exposure-related immune effects at different stages of people's lives. This can potentially help to pinpoint the most critical exposures and the groups most at risk and help delineate the right preventative actions and policies at the individual, group and population levels.

The EXIMIOUS project consists of ten Work Packages (WPs). Of these, eight form the core of the project while the other two are dedicated to ethical and cluster activities from the EC. Data are collected and analyzed in the following WPs and a schematic overview of the workplan is presented in Fig. 1:

- WP2: Population studies
- WP3: Exposome identification - exposure assessment through inhalation and skin
- WP4: Immunomics - single cell and soluble protein level
- WP5: Multi-omics for in-depth exposome investigation

<sup>1</sup> [https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management\\_en.htm](https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm).

- WP6: Statistics and system Immunology

## 3. EXIMIOUS data management, legal and ethical workflow

In accordance with Article 5 EU GDPR principles relating to the processing of personal data, i.e. the "Data protection by design and by default" and the principle of data minimization, the EXIMIOUS project intends to only collect information that is "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (Article 5.1c) ("EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex," n.d.). This has been additionally described in section 5 of the manuscript.

The project's first step towards data management was to prepare a DMP, while consulting several relevant resources, including the EC's H2020 online manual ("Data management - H2020 Online Manual," n. d.) and publications available in the open literature (Castelli et al., 2021; Michener, 2015; Schiermeier, 2018). The DMP will serve as a roadmap to the EXIMIOUS project, in line with the EC's definition of a DMP as a key element in support of good data management during the life cycle of a research project. The EXIMIOUS DMP is considered a 'living' document, laid down at the start of the project and with updates foreseen throughout the lifetime of the project, when needed. The DMP describes data types, sources, origins, processes of collection and formats for all data sets within the project, thereby addressing important aspects of data provenance and traceability. It describes how our data will be handled according to the FAIR principles, as discussed in the subsequent sections (Section 7). Furthermore, it presents how resources are allocated and managed, data are stored, and security is assured, and how ethical and legal issues are considered, relative to the major issues of data access, usage and access audit (Section 5-7). Fig. 2 presents a brief overview of the data management, legal and ethical workflow in the EXIMIOUS project.

The data management, legal and ethical workflow of the EXIMIOUS project outlined in Fig. 2 includes.

- Preparation of the Grant Agreement (GA) - the overarching contract made between the project and the EC, containing the legal framework of the project, defining the aim and scope of the project, rights and obligations, project duration, overall budget, etc.
- After the preparation of GA, the Consortium agreement (CA) was prepared following the DESCAs model Consortium Agreement ("DESCA Model Consortium Agreement - DESCAs Model Consortium Agreement, Model Consortium Agreement," n.d.) and signed by all partners outlining among others – the purpose, responsibilities, GDPR compliance, governance structure, ownership, transfer and dissemination of results. These agreements also form the basis for data sharing and FAIRness of the EXIMIOUS data (described below in section 7).
- Based on the WP descriptions and objectives set in the project, detailed documents were prepared and submitted for review to the EC:
  - o Draft Data Management Plan (DMP V1) - outlining the aspects of data collection, type, sources, and security measures
  - o Draft standard operating procedure (SOP V1) for biological sample collection, processing, and storage
- Ethical approvals - EXIMIOUS brings together 15 partner institutes from 7 countries (Belgium, Norway, Denmark, UK, Spain, Romania, Switzerland) with different national and institutional ethical and biobanking regulations. As a next step, each participating cohort/study population registered their respective studies in accordance with their national ethical and legal regulations and EU regulations. A complete summary of the ethical approval(s) for each participating center was submitted as a deliverable to the EC. The individual (national) ethical committees provided feedback and requested additional information/clarification which fed into revisions of the draft DMP and draft SOP (DMP V1 and SOP V1, submitted to the EC on month 6 and 10 respectively). These were then submitted for

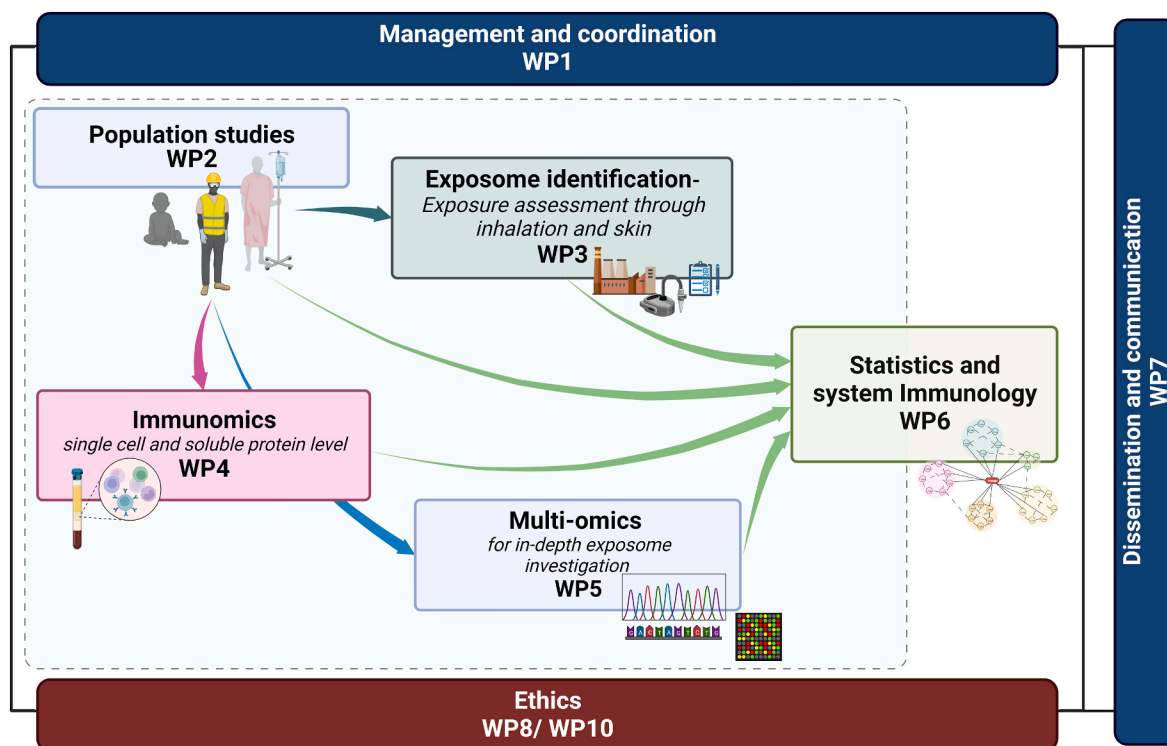


Fig. 1. Workplan of the EXIMIOUS project showing work packages (WP) involved in active data collection and analysis. WP7 is dedicated towards dissemination and communication. WP9 is not shown as it deals with cluster activities of the European Human Exposome Network and therefore is not of concern for the data management of the EXIMIOUS project. Created with BioRender.com.

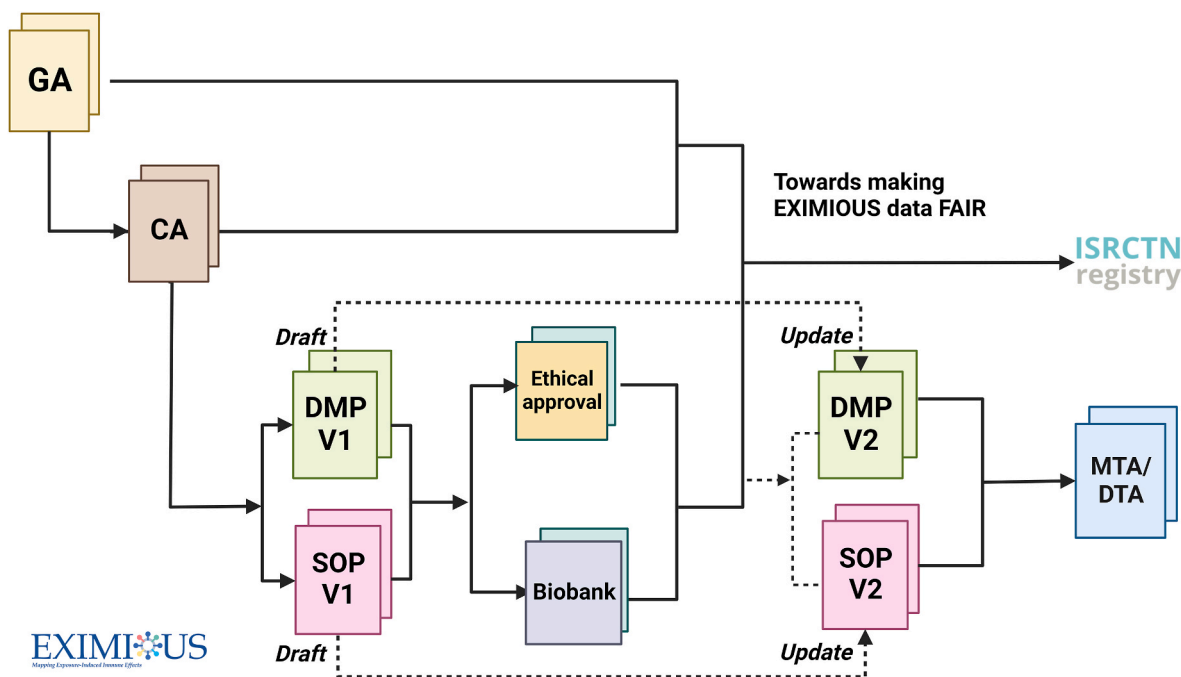


Fig. 2. Schematic representation of the data management, legal and ethical workflow of the EXIMIOUS project; Grant agreement (GA), Consortium agreement (CA), Draft Data Management Plan (DMP V1, draft submitted to the EC on month 6), Draft standard operating procedure (SOP V1, draft submitted to the EC on month 10), Revised Data Management Plan (DMP V2, draft submitted to the EC on month 18), Revised standard operating procedure (SOP V2, draft submitted to the EC on month 24), Biobank registration (Biobank), Material transfer agreement (MTA), Data transfer agreement (DTA), Clinical trial registry (ISRCTN registry); Created with BioRender.com.

review to the EC. The SOP was additionally revised based on feedback from cohorts involved in sample collection. Considerations such as ease of sampling in the field, handling and processing time

available, and most importantly quality of the samples (piloted in the project for multiple centres) were taken into account to produce the updated version of SOP.

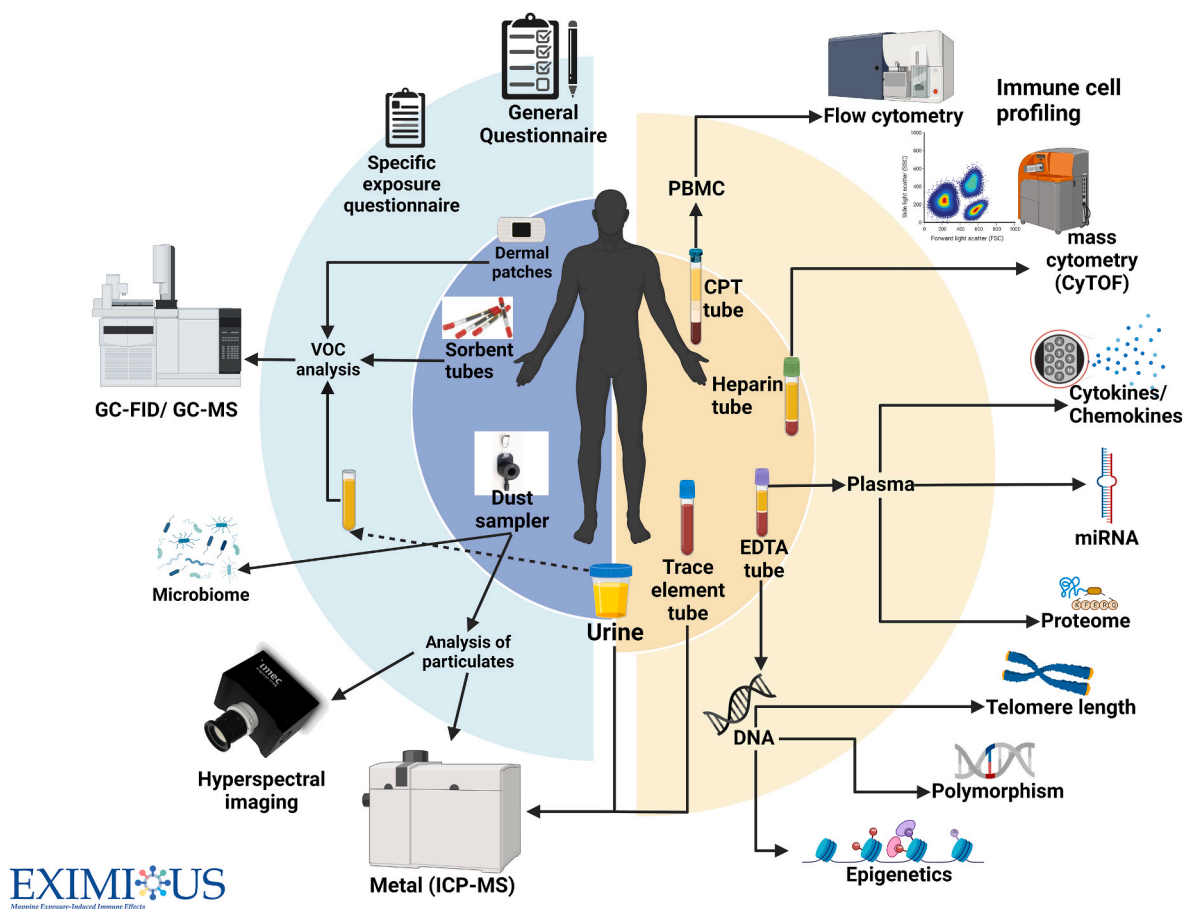
- These revised documents (DMP V2 and SOP V2, submitted to the EC on month 18 and 24 respectively) then formed basis for preparation of the common overarching “EXIMIOUS Agreement on SOP – study obligations and template MTA”, which included key aspects of Material transfer agreement (MTA), Data transfer agreement (DTA) and joint controllership agreement (JCA). Since both the SOP and DMP are considered living documents, updates can still be made, in consensus with all partners. Changes are documented in an accompanying log. The final versions will be made openly available on the project website. The latest version of the SOP is available through the KU Leuven research data repository (RDR) (“EXIMIOUS SOP for collection, storage and shipment of biological samples (updated version 3) - KU Leuven RDR,” n.d.)
- Simultaneously, EXIMIOUS was registered as an observational case-control study - [ISRCTN24225493](https://doi.org/10.1186/ISRCTN24225493) (<https://doi.org/10.1186/ISRCTN24225493>) within the [ISRCTN registry](https://www.isrctn.com/), a clinical trial registry recognized by the WHO and ICMJE. The registry is also a living repository and information will be updated on the ISRCTN portal every 6 months and will contribute towards making data FAIR. The registry has also been linked to the SOP and will be supplemented with all versions of protocol and aggregate results/publications resulting from such data.

#### 4. The complexity of the EXIMIOUS data

##### 4.1. Purpose of the data collection/generation

An overall description of the EXIMIOUS project, its study populations, collected data and planned biological measurements are presented in the EXIMIOUS profile paper (Ronsmans et al., 2022). Through the collected data and their integration (Fig. 3), we aim to achieve the following specific objectives:

- 1) **Delineating the exposome of individuals from selected cohorts (WP2 & WP3):** This first objective is the comprehensive mapping of the exposome in a broad range of cohorts (“ISRCTN - ISRCTN24225493: EXIMIOUS: Investigating the impact of exposures (exposome) on the immune system (immunome),” n.d.; Ronsmans et al., 2022), encompassing the entire lifespan: general population, including pregnant women and their children, occupational and disease cohorts. The mapping involves a combination of measurements, exposure models, and exposome data already available for the cohorts, e.g., via national registers. These are complemented with internal exposome data generated in the EXIMIOUS project, from exposure biomarkers (WP3), immune fingerprints (WP4) and other omics data (WP5, also see Objective 2). The combination of these data will form the basis for a generation of new exposome models that enable attribution of multiple exposures at the individual level (Objective 3).
- 2) **Mapping the immunome and other omics (WP4 & WP5)**



**Fig. 3.** Types of newly collected data within the EXIMIOUS project; GC-FID: Gas Chromatography with flame ionization detection; GC-MS: Gas chromatography–mass spectrometry; VOC: Volatile organic compounds; EDTA: Ethylenediaminetetraacetic acid; miRNA: micro-RNA; PBMC: peripheral blood mononuclear cells; CyTOF: Cytometry by time of flight; CPT: Cell preparation tube; In addition register-based data will be used for different tasks in EXIMIOUS project (Ronsmans et al., 2022). Created with [BioRender.com](https://www.biorender.com/).

- a **Immunomics** (WP4): We use various high-dimensional platforms (Multiplex cytokine and inflammatory protein analyses, high parameter flow cytometry, mass cytometry- CyTOF) to map the individual's immune system by assessing both function and phenotype, i.e., "immunomics". An insight into the immunome will provide additional, complementary information about the exposome, as we can consider the immunome an early biomarker of the effect of the exposome.
- b **Multi-omics for in-depth investigation** (WP5): This WP explores the role of multi-omics (Genome wide association study, epigenomics, circulatory miRNA, circulatory proteome) (Ronsmans et al., 2022) to fully describe the immune signature associated with exposure on selected cohorts.
- 3) **Combining the exposome, immunome, clinical data and societal impact via bioinformatical methods** (WP2, WP3 & WP6): Through the integration of the experimental platforms measuring both exposures (WP3) and immunological outcomes (WP4-WP5), we seek to determine the immunological consequences of environmental exposures.
- 4) **Creating a toolbox for researchers, policy makers and the general public** (WP7): We will create tools that meet the needs of stakeholders, to give them benefit from the project. The toolbox will take shape as the project evolves and will be documented in the follow-up versions of the DMP.

**4.2. Sources and types of data:** To achieve the defined objectives, a huge amount of data from various sources will be collected in the EXIMIOUS project. These are briefly presented below and summarized in Table 1.

- Questionnaire data
- Clinical data
- Environmental sampling of current exposures in different cohorts
- Spectral data and the machine learning models generated
- Modelling of environmental exposure (all cohorts)
- Biomonitoring data and immune fingerprints
- Neural network analyses outcomes

## 5. Applied principle of data minimization

The EXIMIOUS project will only collect information that is "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (Article 5 GDPR - Article 5.1c) ("EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex," n.d.). However, given the nature of this project, the relevance of each of the many collected data types is not easily evaluated in advance. We know from the scientific literature that some specific environmental exposures might be associated (at least in part) with some immune-mediated diseases, but these relationships are not well explored. Moreover, it is generally accepted that non-explored exposures, as well as some genetic variations, can play a role in the development of immune-mediated diseases (Cooper et al., 1999; Costenbader et al., 2012; Handel et al., 2010; Waubant et al., 2019). Therefore, it is our goal to explore previously uninvestigated factors, to pinpoint the most critical forms of exposure that lead to exposure-related immune effects at different stages of people's lives (young/old), as well as identify the groups of people most at risk. This will allow implementation of the right preventative actions and policies at the individual, group and population level.

The relevant/adequate amount of data per individual participant is therefore difficult to define in such complex studies of exposome or systems immunology (Davis et al., 2017). The choice for EXIMIOUS is to limit collection/generation/processing to data that have a high likelihood of importance for the project's objectives. In the collection, generation and processing of the broad set of data for the purpose of the project objectives, we conclude that the processing of personal data is:

- A. **Adequate**, in view of using both conventional and novel techniques that combine multiple fields of expertise.
- The high-dimensional single-cell and omics techniques allow for collection of a broad spectrum of data, which can be analyzed with conventional bio-statistical techniques in search for associations/correlations.
  - Using systems biology, artificial intelligence and machine learning, new bio-informatics tools (non-biased search for associations) will be developed.
- B. **Relevant**, in view of the importance of detailed and broad data collection (omics data from the immune system, the exposome and the genome). We collect the following data on the participants:
- Environmental and occupational exposure by environmental sampling, measuring biomarkers of exposure, and modelling exposure based on location (air pollution models) or based on job title (job exposure matrices – JEMs) (exposome).
  - Immune health status and function (immunomics), such as blood cells and plasma samples and clinical files.
  - Other omics: genetic variations, epi-genetic changes, proteome.
- C. **Limited**, in view of existing literature and power calculations (where possible) as well as by EU, national and ethical guidelines. It is probably even more difficult in this kind of studies to define the limits to data collection.
- The type of sample/laboratory and data analyses and the search for unknown associations demands sufficiently large datasets from various study populations (cohorts), including workers, patients, children and healthy volunteers of both sexes.
  - In order to estimate the number of volunteers/patients in epidemiological studies a power analysis can be performed. In our project, in which several 'unknowns' are explored, it is practically impossible to run a science-based power calculation. Nevertheless, we made an attempt to estimate the number of samples needed, two examples of which are presented here:
    - o Samples requested from the biobank LifeLines: a formal power calculation for the planned nested case-control studies was not possible. Therefore, we ran numerous scenarios of plausible exposures, detectable differences between groups to estimate necessary sample sizes: 300 patients with rheumatoid arthritis, 175 persons diagnosed with systemic lupus erythematosus and 350 patients with type I diabetes and 500 matched controls.
    - o GWAS analysis: A minimum of 720 participants will be genotyped for 700,078 variants. Power is 80% to detect variants explaining 5.5% of variation in an immune trait at genome-wide significance ( $P = 5 \times 10^{-8}$ ).

## 6. Secure workspace in the EXIMIOUS project

In a project like EXIMIOUS, several types of documents and data will be considered confidential or personal. In accordance with Article 5.1f of the GDPR (EU), personal data has to be "processed in a manner that ensures appropriate security ..., including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures ('integrity and confidentiality')" ("EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex," n.d.).

To preserve the confidentiality of personal data and other sensitive information, the workspace in the EXIMIOUS project has been divided into two independent secure platforms (Fig. 4) - a "collaborative platform" and a "data storage platform".

By definition, "personal data" includes any information that relates to an identifiable individual. Therefore, documents such as signed agreements, contact lists, minutes and recordings of meetings, confidential study protocols and deliverables, manuscripts or documents with embargo are stored in the secure and password protected collaborative SharePoint platform managed for the project by KU Leuven. As for all secure data platforms, access requires a two-factor authentication and is only granted upon explicit approval of partners. All documents that are

**Table 1**

Summary table showing the origin and type of data in the EXIMIOUS project (Ronsmans et al., 2022). PB: Peripheral blood; CB: Cord Blood; BAL: bronchoalveolar lavage; KU Leuven-Katholieke Universiteit Leuven, Belgium; UCL- Université catholique de Louvain, Belgium; VHIR- Vall d'Hebron Institute of Research, Spain; U Hasselt- Hasselt University, Belgium; NRCWE- The National Research Centre for the Working Environment, Denmark; RegionH- Region Hovedstaden, Denmark; UMFST- The University of Medicine, Pharmacy, Science, and Technology of Târgu Mureş, Romania.

SOURCE OF DATA (PRIMARY / SECONDARY)	TYPE OF DATA														
	Primary Biological Samples			Questionnaire Data				Clinical Data		Environmental Samples and Exposure Data					
	Blood	Urine	Other: BAL, ...	Health & medication	Sociodemographics	Lifestyle habits	Current and past exposures	Electronic patient files	Clinical assessment workers	Air	Mineral dust	Dermal patches	Risk assessment	Safety data sheets	Exposure profile
<b>DISEASE COHORTS</b>															
Sarcoidosis and systemic sclerosis (KU Leuven)	X (PB)	X	(x)	X	X	X	X	X						(x)	X
Systemic sclerosis, SLE and RA (UCL)	X (PB)	X	(x)	X	X	X	X	X						(x)	X
Hypersensitivity Pneumonitis (VHIR)	X (PB)	X	(x)	X	X	X	X	X						(x)	X
<b>GENERAL AND BIRTH COHORTS</b>															
Lifelines (U Hasselt)	X (PB)	X		X	X	X	X								
ENVIRONAGE (U Hasselt)	X (PB+CB)	X		X	X	X	X								
DOC*X(Generation) (NRCWE/RegionH)				X	X	X	X								
<b>OCCUPATIONAL COHORTS</b>															
Waste workers (NRCWE)	X (PB)	X		X	X	X	X			X	X	X	X	X	X
Park workers (VHIR)	X (PB)	X		X	X	X	X			X	X	X	X	X	X
Workers exposed to dust and solvents (UMFST)	X (PB)	X	(x)	X	X	X	X		(x)	X	X	X	X	X	X
<b>Objective 1: Delineating the exposome</b>	<b>EXPOSOME</b>														
<b>Objective 2A: Immunomics</b>	<b>IMMUNOME</b>		<b>IMMUNOME</b>				<b>IMMUNOME</b>								
<b>Objective 2B: Multi-omics</b>	<b>OMICS</b>		<b>OMICS</b>												
<b>Objective 3: Statistics and systems immunology</b>	<b>COMBINED ANALYSIS</b>														

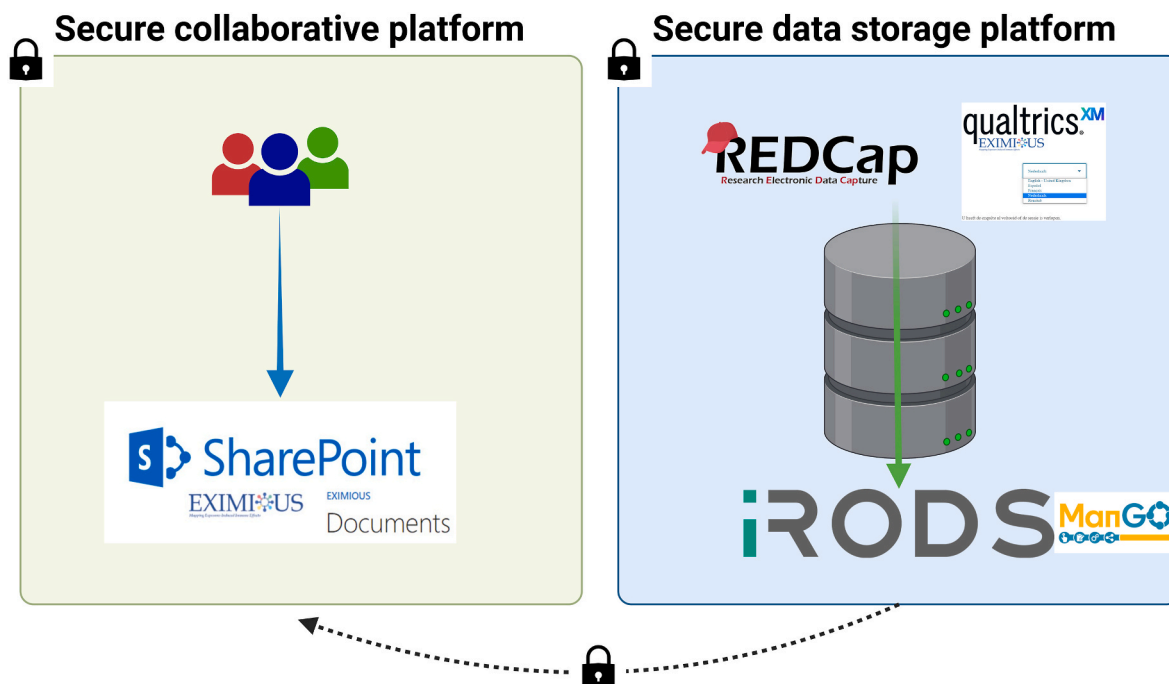


Fig. 4. Separation of secure workspace in the EXIMIOUS project. SharePoint-web-based collaborative platform; REDCap-Research Electronic Data Capture; Qualtrics-web-based survey tool; ManGO-web application designed to work alongside iRODS; Created with BioRender.com.

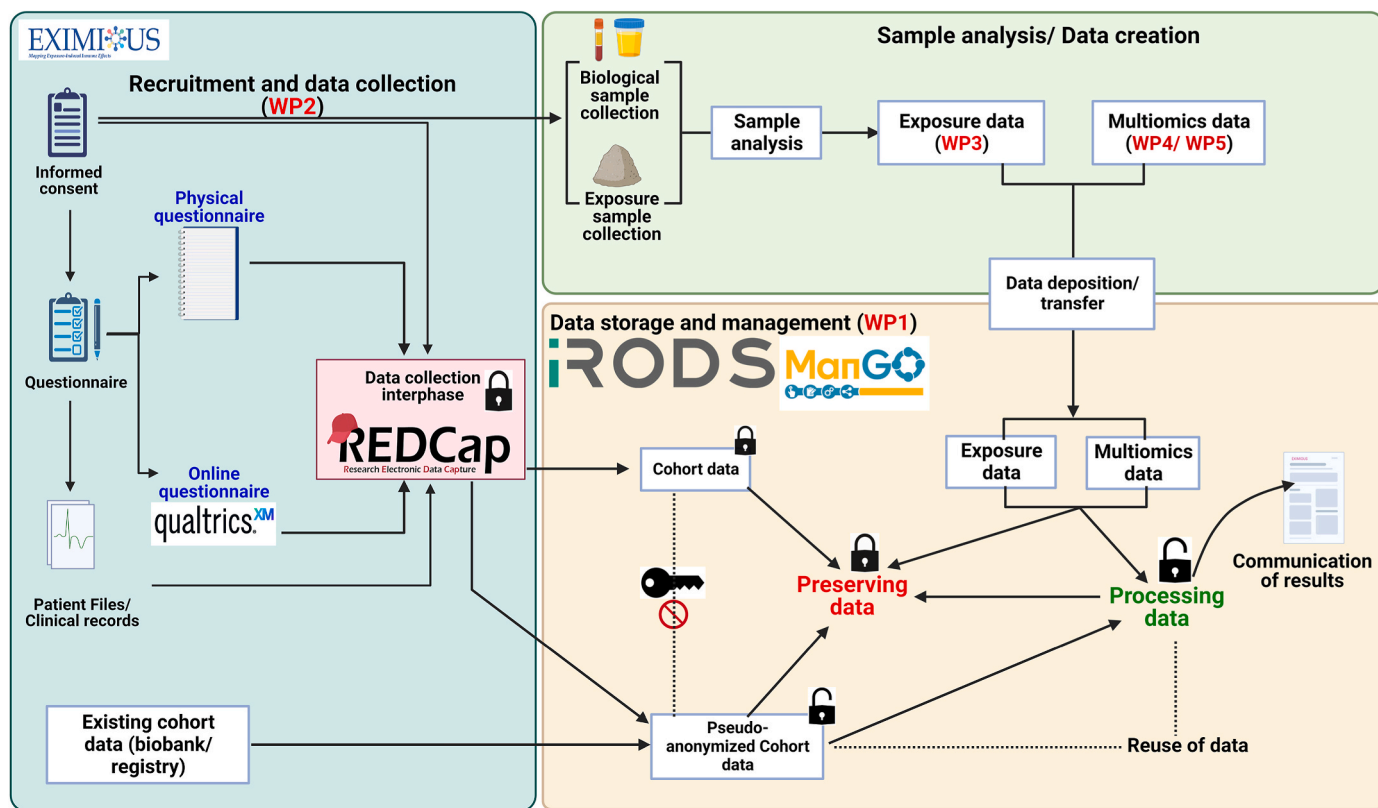


Fig. 5. Data collection workflow and management in the EXIMIOUS project; WP2 is responsible for data collection (clinical and questionnaire based) through interviews and using tools such as physical questionnaire or questionnaire administered through REDCap (Research Electronic Data Capture) or Qualtrics (web-based survey tool). WP2 also uses existing data from biobank/registry. Simultaneously environmental/exposure samples and biological samples are collected for analysis of external exposome characterization (WP3), and immune and omics markers (WP4 and WP5). All data collected from these WPs would be deposited through the secure ManGO-interphase of iRODS; ManGO interface allow setting access permissions at different levels depending on the tasks of a user (indicated by lock and key); Created with BioRender.com.

shared within the consortium's SharePoint site – as a means of collaboration – have a built-in versioning system where all versions are saved, and previous versions can be restored if necessary.

The secure data storage platform is characterized by a much higher level of complexity, with several data types arising from a variety of data sources that are subject to various national and EU GDPR regulations and is briefly described in the next section. Fig. 5 illustrates the flow of data in EXIMIOUS. We utilize several systems to aid data management, keeping in mind the FAIR principles. The two main, interconnected systems used for data collection are Qualtrics (“Online Survey Software - Digital Survey Management Tool,” n.d.) and REDCap (Harris et al., 2019; “REDCap,” n.d.). Qualtrics is a web-based survey tool, which we use to administer the participant questionnaire. REDCap acts as the collection hub where also Qualtrics data gets uploaded. Additionally, in a limited number of cases, data will be collected using physical questionnaires with subsequent loading of data directly into REDCap. All data collected (through questionnaires or the specific analyses described above) will thereafter be securely transferred and stored in iRODS (<https://irods.org/>), our main data management platform.

### 6.1. Additional technical and organizational measures to safeguard data and the rights and freedoms of the data subjects

#### 6.1.1. Specific measures to safeguard data and the rights and privacy of the study participants are briefly described below

- All partners, including the members of the scientific advisory committee, have signed a confidentiality statement.
- With regard to data storage, we have several technical and organizational measures in place to safeguard the rights and freedoms of the data subjects, some of which have been described above.
  - o We will store data/samples in (a) the biobank cluster of KU Leuven, (b) the secured REDCap platform for clinical data, and (c) all data are securely transferred and stored in iRODS.
  - o An exception to our general storage policy is the data from the Danish DOC\*X and the DOC\*X-Generation (DOC\*X-G) cohorts. The cohorts are based on national nationwide registers and remain at the servers of Statistics Denmark in accordance with national legal and ethical agreements.

#### 6.1.2. EXIMIOUS data managers

KU Leuven is the project coordinator of EXIMIOUS and, hence, the main responsible organization for the data management and implementation of the DMP. The coordinating institute is supported by the (data) management team. These data managers have access to the data management platforms with the sole purpose of good data management. The access rights of the data managers have been approved by the project steering committee and the managers have agreed to and signed a code of conduct. The data managers will be responsible for regular audits of data access. This has been organized as follows:

- a **SharePoint**: All partners, including data managers, have access to all documents
- b **REDCap**: In REDCap, three roles have been defined with specific access rights, namely “Project manager” – read-write access, “Project designer” – read-write access, “Read only right” – read only
- c **iRODS** (ManGO interface): In iRODS (ManGO interface supported by the KU Leuven Research data management support: <https://www.kuleuven.be/rdm/en/mango>), two roles have been defined with specific access rights, namely “Project owner” – read only, “Data manager” – read only. The use of iRODS data is monitored and auditing reports (data audit and data access audit) will be made available upon request or at least every 6 months.

#### 6.1.3. Data access groups

The REDCap platform and the iRODS-ManGO interface allow setting

access permissions at different levels depending on the tasks of a user. We have set specific user rights and created Data Access Groups. The Data Access Groups determine the data visibility of a user group: users that do not belong to a Data Access Group will not be able to see data from that group (e.g., cohorts).

#### 6.1.4. Data governance

The data managers, together with WP and task leads, will also be involved in data governance, usually defining policies, standards, and procedures to ensure quality of data and compliance with ethics, security and FAIR principles.

#### 6.1.5. EXIMIOUS works in accordance with the EU GDPR regulation described under Article 5.1e of ‘storage limitation’ (“EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex,” n.d.)“

*personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject”.*

Additionally, participants are informed in the consent letter of the following information:

- Participation in this study is voluntary. Participants have the right to cancel their participation at any time, without the need to give a reason, and cancellation will not have any personal consequences.
- The names of the participants will not be published. The anonymity and confidentiality of the data are guaranteed at every stage of the investigation. Anonymized data will be made available through scientific publications, or on the website of the project.
- If participants would like to be kept informed of the results of this research, they may contact the researchers. Participants may contact the cohort owner/principal investigator (PI) if they have questions, such as for exercising the right of data access, including corrections etc.

#### 6.1.6. Pseudonymization

The collected data will be pseudonymized as defined in Article 4 Nr. 5 (“EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex,” n.d.) “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person”. The key – linking ID to a specific person will be stored in a digital vault only accessible to the cohort holders and the PI of EXIMIOUS. The process of assigning unique ID to a participant is described in the subsequent section of FAIR data. Data published in publicly available publications (e.g., scientific publications, reports, data on the project website) will be fully anonymized.

## 7. Making EXIMIOUS data FAIR

In EXIMIOUS, our research data will be handled following the FAIR principles. The checklist presented in Fig. 6 provides a summary of the actions that make EXIMIOUS data FAIR. We assign specific attention to the FAIR principles, as EXIMIOUS is a collaboration of multiple research institutions in several countries and involves different stakeholders. While making data FAIR, we also apply the principle of “as open as possible, as closed as necessary”. In preparation of FAIR data, we referred to the EC report “Turning FAIR into reality-Final Report and Action Plan from the European Commission Expert Group on FAIR Data” (Commission, n.d.) and other publications (Boeckhout et al., 2018; “Open innovation,



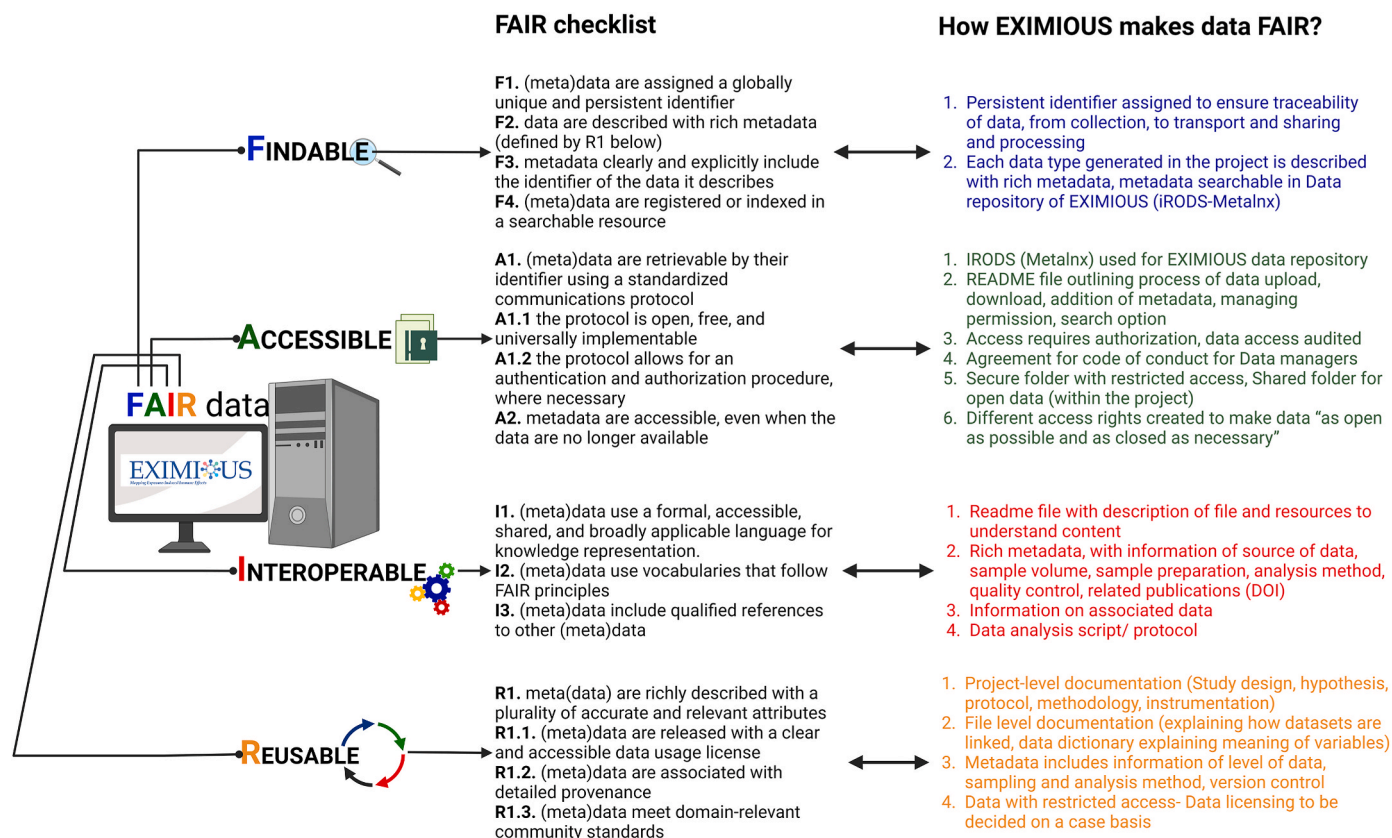


Fig. 6. Schematic representation, summarizing aspects of how the EXIMIOUS project makes data FAIR following the checklist proposed by Wilkinson et al., (Wilkinson et al., 2016); Created with BioRender.com.

open science, open to the world - Publications Office of the EU," n.d.; Wilkinson et al., 2016; Zare Jeddi et al., 2021).

## 7.1. Findable

### 7.1.1. Making data discoverable with metadata

Metadata is defined as "data that defines and describes other data and processes" ("OECD Glossary of Statistical Terms - Metadata Definition," n.d.). Data discovery by means of metadata is a major strength of iRODS (for more information: <https://irods.org/>). Researchers can upload their data in ManGO, which is a graphical user interface to access iRODS, and at the same time fill in the associated metadata, thus meeting FAIR principal F (= Data are described with rich metadata). In the ManGO interface (iRODS) metadata is customizable, and different disciplines can define differing fields if necessary. Additionally, within EXIMIOUS, through peer consultation, a draft metadata template has been developed, with inspiration from the IPChem (Information Platform for Chemical Monitoring) ("IPChem Portal," n.d.) metadata template and the RDA metadata standards catalog ("Metadata Standards Catalog," n.d.) [Supplementary file 1; subject to further modification]. To clarify the terms used in each metadata file, a dictionary/glossary will be created and saved as a README file together with the metadata file. Additionally, in the ManGO interface (iRODS), each dataset is assigned a permanent unique digital identifier. Moreover, in the context of European Human Exposome Network (EHEN), EXIMIOUS participates in the initiative of developing a harmonized metadata template across the nine EHEN projects.

### 7.1.2. All records will be linked using a unique ID

Each participant will have a unique ID with format: [EXIM]\_ [cohort abbreviation]\_[sub cohort]\_[unique number].

**Example.** EXIM\_ENV\_10\_00001 (= the first 10 years old individual in the ENVIRONAGE cohort)

**Example.** EXIM\_KUL\_SARC\_00001 (= the first individual in the KU Leuven disease cohort of Sarcoidosis patients)

Further, each sample collected, and any related analysis results will be linked to the unique participant ID, alongside the analysis protocols of the specific assay, through metadata. In case of already established cohorts, such as LifeLines and ENVIRONAGE, the unique EXIMIOUS ID are linked to the cohorts' internal (existing) individual ID. While pseudonymization of data has been described earlier, publicly available results (e.g., scientific publications, reports, data on the project website) will be fully anonymized. Additional steps may be required towards (pseudo)anonymization during data sharing for reuse, and will be done by use of an additional ID/key.

### 7.1.3. Provenance/traceability of samples and data - linking participant ID to analysis

A large and central part of the work in EXIMIOUS consists of analyses relating to constituents of peripheral blood, urine and environmental samples. Some individuals also contribute samples of other tissues. Samples from WP2 originate from different parts of Belgium, Spain, Denmark, and Romania (WP2) and will be shipped to partners in Belgium and Norway for analyses described in WP3-5. In making data FAIR, considerable effort was dedicated towards biobanking of materials and traceability of samples from collection to analysis and beyond.

As the coordinating center the main biobank is situated at KU Leuven, and therefore EXIMIOUS uses the Belgian Law on the procurement and use of human biological material of December 19, 2008 as guidance document ("Belgian Advisory Committee on Bioethics from a human genetics centre on the use of DNA banks," n.d.; Lalova et al., 2021). Human Bodily Material (HBM) is defined as "any biological

material, including human tissues and cells, gametes, embryos, fetuses, as well as the substances derived thereof, regardless of their degree of processing/transformation (DNA, RNA, proteins, ...). The Human Bodily Material Law further states that all samples “... must be registered within a notified Belgian biobank with which a contract or framework agreement ...”. Additionally, article 22.2 “requires the biobank to keep a register on the nature of the human biological material the storage and provision of which it assures, as well as on its origin and its destination” (“Belgian Advisory Committee on Bioethics from a human genetics centre on the use of DNA banks,” n.d.; Lalova et al., 2021). Biological samples that are shipped to Belgium, will be stored at the University Hospital Leuven/Katholieke Universiteit Leuven Biobank until analyzed. After the project is finished, any samples left and available will be stored for an additional five years to repeat parts of analyses and/or address additional scientific questions generated from the project. Thereafter, samples will be destroyed or returned to the providing laboratory. This will be done within the legal and ethical framework of both Belgium and the regulation of the nation of origin of the tissue.

As an EEA (European Economic Area) country, Norway has implemented most EU regulation in Norwegian law. Consequently, Norway adheres to the EU regulations relevant for the shipment of samples from the EU to Norway and the subsequent handling of samples and data. Both the GDPR (EU) 2016/679 (“EUR-Lex - 32016R0679 - EN - EUR-Lex,” n.d.) and Directive 2004/23/EC (“EUR-Lex - 32004L0023 - EN - EUR-Lex,” n.d.) on setting standards of quality and safety for the donation, procurement, testing, processing, preservation, storage and distribution of human tissues and cells are implemented in Norwegian law. Therefore, all research activities in EXIMIOUS will be performed in accordance with both Norwegian and EU law. The biological samples will be stored at Norwegian Institute of Public Health until analyzed. If agreed with the providing laboratory and if included in the consent documents, tissues will be stored and handled in agreement with Norwegian regulations and biobank requirements at the institute. Otherwise, samples will be destroyed or returned to the providing laboratory after the project is finished or as agreed between partners and described in the MTA and ethical approval.

Additionally, in accordance with the Belgian Law of HBM and the KU Leuven biobank agreement, in the EXIMIOUS project we will document all relevant information on human cells and tissues in the biobank registry. This ensures traceability of the cells, tissues and their derivatives.

Use of the registry is mandatory for all partners. The measures described below briefly summarize the steps taken to establish the provenance of all forms of data collected in the project:

- Each collected exposure sample, biological sample, sample aliquot and sample derived thereof (e.g., DNA, RNA, protein) will be linked to the unique participant ID as described above.

Example of biological sample linked to its participant ID: [Participant ID]\_[Sample information]\_[aliquot number]: EXIM\_ENV\_10\_00001\_CPT\_PBMC\_1 (= Aliquot 1 of PBMC isolated from CPT tube of the first 10 years old individual in the ENVIRONAGE cohort).

This will allow traceability of samples from collection to analysis and beyond, linked through metadata in the EXIMIOUS storage platform - ManGO interface (iRODS).

- All partners collecting biological samples are expected to enter the number and ID of the biological samples collected for each participant in REDCap. This acts as the first step of traceability, linking collected biological samples to individual participants and their questionnaire records.
- In the next step, samples are processed in the laboratory and labelled accordingly, and records of sample and storage condition are recorded by the cohort owner.
- All partners sending or receiving biological samples are required to complete a sample tracking sheet stored in the collaborative

SharePoint folder. All partners provide information on the shipped samples including sample ID, number of each type of sample, shipment conditions and a signed version of the MTA. The receiving partner is responsible for assigning the location of storage (freezer/biobank) and updating the information on the tracking sheet after receiving the samples.

- Prior to analysis, samples must be checked out of the system.
- After completion of analysis, the results must be linked to the sample ID and deposited in iRODS.

These steps are expected to ensure complete traceability of samples from collection to analysis and can be audited at the end of the project. The procedure also provides information of remaining biobanked samples, and hence, informs the choice to be made about the samples at the end of the project.

#### 7.1.4. Findability in the ManGO interface (iRODS)

A folder system has been set up in the ManGO interface (iRODS) for storage of all sample data and results of all technical analyses. The data from REDCap is automatically exported to the ManGO interface (iRODS) and organized in folders (i.e., Data Collections in the ManGO interface (iRODS)). Other data can be directly deposited in the correct folder on the ManGO interface (iRODS). A detailed guidance document has been shared with all collaborators on the organization and naming (convention) of folders.

#### 7.1.5. Version control of data

Version control will be handled by using the file name + v [version number]. The version number starts at 01. A new version is saved after each modification. Due to continuous recruitment of volunteers/patients, the data stored in the REDCap database and ManGO interface (iRODS) will grow continuously. The data collection instrument REDCap has a Project Revision History table, which includes all changes made to the project. REDCap also allows the user to download back-ups of the data and metadata at any given stage. All data from REDCap is automatically exported, on a daily basis, to the ManGO interface (iRODS).

Data stored will also be backed up at a regular interval in iRODS. For instance, the questionnaire data imported from REDCap will be backed up daily. The daily copies from the last day of the week will be stored as a ‘weekly back-up’. ‘Weekly back-ups’ will be bundled and stored as a ‘monthly back-up’ and eventually as a ‘yearly back-up’. This version control system allows us to revert to older copies of the data, if needed. A similar exercise will be done for all other data types deposited to the ManGO interface (iRODS), the frequency of backup of which will be decided on a case-by-case basis. The raw data will be protected strictly against changes (data may be added but raw data will not be edited). This is done by a clear folder structure in the ManGO interface (iRODS), where folders containing the raw data are Read Only (RO). An automated copy of the raw data is made in a different folder that has Read-Write (RW) permissions. In the metadata file, we will keep track of all the analyses performed in the analytical database containing research question-criteria-records included and model used.

## 7.2. Accessible

### 7.2.1. EXIMIOUS data to be made openly available as default?

The fact that we work with personal data warrants caution. We will therefore completely (pseudo)anonymize all data before making the data accessible outside of the consortium.

Within the consortium: All participants will give consent to participate in the research, as well as consent that their personal data may be processed in the context of the research within the scope of the research objectives. Personal data will be handled restrictively, even within the consortium. Only the data managers from the cohort owners will have access to personal data from participants of their own cohort. Other partners will have access to the pseudoanonymized data (no identifiers).

These restrictions are set in both REDCap (by use of Data Access Groups) and in the ManGO interface (iRODS) (by setting access restrictions and permissions on the level of Collections described earlier).

One exception is the geographical data (i.e., addresses), which is needed for exposure modelling in WP3. Addresses are obtained from the participants in the various cohorts. It cannot be fully anonymized since the address could serve as an identifier. For modelling purposes, only the 'Address information dataset' will be available for analysis and only for the WP3 partners. Thereafter, the processed and de-identified results will be made openly accessible within the consortium.

In EXIMIOUS, secondary data is retrieved from the DOC\*X, DOC\*X-G and LifeLines cohorts. In case of DOC\*X and the DOC\*-G, the reuse of data is described in the Consortium Agreement. In the case of LifeLines (an external biobank), LifeLines offer documents, and Data and Material Transfer Agreement (DMTA) are signed by all parties using the data.

**Beyond the consortium:** The data situated under objective 4 ("A toolbox for researchers, policy makers and the general public") is open and shared externally. Some of the anonymized results will be made available and shared through the toolboxes.

### 7.2.2. How will the shared data be made accessible?

**Within the consortium:** We will use REDCap and the ManGO interface (iRODS) to store and manage all data. Access is restricted but can be granted upon (written) request and is subject to "the acceptance of specific conditions aimed at assuring that these rights will be used only for the intended purpose" (Detailed in the EXIMIOUS CA). The access request procedure is described in the CA: "Access Rights". This will be handled in the ManGO interface (iRODS) using iRODS Permissions as described earlier. We note that one of the core components of the ManGO interface (iRODS) is 'secure collaboration'. ManGO interface (iRODS) permissions are analogous to UNIX file system permissions. The owner of a Data Object (i.e., a file) or Collection (i.e., a folder) can assign read or write access for any number of defined ManGO interface (iRODS) Users and Groups. Group membership is managed by the group administrator(s).

The sharing of files and other documents within the consortium, or between the consortium and the EC, will be handled securely. Via Belnet, KU Leuven offers the file sharing tool Filesender, through which (large) files of up to 2 GB can be shared (<http://filesender.belnet.be>). The files are stored only temporarily and can be accessed only by authorized users.

**Beyond the consortium:** This can only be considered for data for which the volunteer has consented to reutilization and approved for pre-determined research uses. Depending on the granted permission (pseudonymized or anonymized), on a case basis for each cohort, and the location of the requesting researcher (Belgium, EU including Norway, rest of world), the permission will allow different levels of controlled access to parts of EXIMIOUS data. To protect EXIMIOUS data, users will need to accept and agree to the terms and conditions of the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International Public License ("Public License"). This is encouraged by the European Commission as a useful licensing solutions ("European Research Council (ERC) Guidelines on Implementation of Open Access to Scientific Publications and Research Data," 2017). More information can be found on their website: <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>. When access is granted, the necessary keys for access will be transferred through secure channels either using the ManGO interface (iRODS) system or through the Globus platform. The owner of a Data Object or Collection can create a Ticket and share it with external users to grant them read or write access. Tickets can be revoked, and they can be set to automatically expire upon a specified date and time or a specified number of reads or writes.

### 7.2.3. Where will the data and associated metadata, documentation and code be deposited?

Most participant data and associated metadata is - firstly - captured in the REDCap database, from where it is automatically exported to the

ManGO interface (iRODS) database at KU Leuven. Everything else (raw data from experiments, analysis codes and results) will be deposited directly onto the ManGO interface (iRODS). At both locations, matching metadata files (TXT) will be stored. Due to the dynamic data collection during the project and restrictions concerning EU-legislation on free access to data, data access needs to be restricted, as described above. After finalization of the project, a secondary repository can be considered, such as Dryad. As patient data cannot be published, only fully anonymized data can be deposited.

### 7.2.4. Dissemination of results

Additionally, as mandated by Article 29.2 of the EXIMIOUS GA and in the H2020 model grant agreement ("H2020 Programme AGA-Annotated Model Grant Agreement," 2019), all academic partners will disseminate their specific scientific findings through publications in peer-reviewed journals. The majority will be written in English language, to ensure broad reach. All publications will be tracked and monitored within the context of WP7 to ensure compliance to open access. Additionally, a toolbox will be developed and made available through the project website. The toolbox will contain tools for the Patient, Scientific and Policy Target Groups, to enable such stakeholders to make use of the EXIMIOUS results and in certain cases also generate new data and information—also after the EXIMIOUS project—on the exposome and immunome, but also on clinical and societal implications, which are necessary for designing new policies.

### 7.3. Interoperable

#### 7.3.1. Are the data produced technically interoperable, i.e., enabling data exchange and re-use between researchers, institutions, organizations, countries, etc.?

The data files generated will be in standardized formats. To ensure the data is interoperable, the project coordinator responsible for data management will seek feedback and assistance from the KU Leuven Research Data Management Support Desk.

All data in the ManGO interface (iRODS) will be stored in standard formats, some of which are presented below:

- Flow cytometry and mass cytometry (CyTOF) raw data: FCS
- Sequencing raw data: FASTQ("FASTQ Format," n.d.)
- Figures: Jpeg, PNG (raster graphics), PDF (vector graphics)
- Tabular data: CSV
- MATLAB("MATLAB - MathWorks - MATLAB & Simulink," n.d.): M-files. MAT-files. MEX-files
- R code: R-files (TXT)
- Hyperspectral images: raw file and header file (ENVI format)("ENVI Header Files," n.d.)

Additional data and file types may be used/generated in the project and will be updated in the updated versions of the DMP.

#### 7.3.2. What data and metadata vocabularies, standards or methodologies will be followed to make your data interoperable?

In general, commonly used (biomedical) MeSH terms will be used. For assay metadata we will use:

- Flow cytometry: MIFlowCyt (Lee et al., 2008; Spidlen et al., 2012)
- Sequencing: MIAME (Brazma et al., 2001)

The metadata for specific instruments will include brand name, serial number, year of manufacture.

For environmental measurements we will follow the ISO 45001 standard.

### 7.3.3. Using standard vocabularies for all data types present in the data set, to allow inter-disciplinary interoperability?

Since this is a multidisciplinary project, it cannot be limited to only one standard vocabulary. MeSH terms will be used as a common ground, but, if necessary, the more technical glossary for workplace safety as defined in the OSHwiki ("Main Page - OSHWiki," n.d.), can be used. It is a collaborative online encyclopedia of accurate and reliable information on occupational safety and health ([http://oshwiki.eu/wiki/Main\\_Page](http://oshwiki.eu/wiki/Main_Page)).

### 7.3.4. If it is unavoidable that uncommon ontologies or vocabularies are used, will mappings to more commonly used ontologies be provided?

If the standardized terms are not precise enough, commonly used terms will be used with a descriptive adjective added to ensure precision. These 'new' terms will be included in our glossary in a README file (README.TXT).

## 7.4. Reusable

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

Data is collected from different cohorts, so re-use will depend on the consent given and national legislation. EXIMIOUS strives to open its data for re-use, by direct access under appropriate restrictions and license terms (i.e., a Creative Commons « Public license») or by deposition in public repositories (when applicable).

### 7.4.2. When will the data be made available for re-use?

After the end of the project, an embargo of 12 months will be applied to allow project members to finalize tasks that might have been delayed or postponed. At the end of this period, we foresee a re-evaluation of the status of the project before release of the embargo. Only certain part of the aggregated data can be made available openly beyond the consortium. Other data can be made available upon request after evaluating both the legal and ethical considerations.

### 7.4.3. Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?

Other than researchers, the data generated in EXIMIOUS is of interest to patient organizations, medical practitioners (occupational medicine), industrial hygienists and policymakers. During the project, specific tools for stakeholders will be developed to share this information, available at the project website (<https://www.eximious-h2020.eu/>). Access to other data is restricted but can be granted upon request (as described above).

### 7.4.4. How long is it intended that the data remains re-useable?

The ManGO interface (iRODS) database will remain active for at least five years after project termination. Then the data will be transferred to a secured central server at KU Leuven (Research Data Repository- RDR) or stored on iRODS for the remaining period, subject to availability of budget. While iRODS has been briefly described, KU Leuven's Research Data Repository (pronounced "Radar") is built on Dataverse (<https://dataverse.org/>). The website <https://www.eximious-h2020.eu/> will be maintained for at least five years after the end of the project, and tools and results developed during the project will remain accessible for all stakeholders through the coordinating center for at least fifteen years after the project is completed.

### 7.4.5. Are data quality assurance processes described?

Data quality will be controlled at different levels:

1. Collection of samples: Internally validated SOPs have been developed and used both for sample collection and analysis. In the first two years of the project, attention has been given to development of SOP for sample collection, labeling, secure storage and transport of physical samples (e.g., never sending all samples in one package to prevent loss).

2. Protocols for analysis of samples have also been made available by all partners. Additional efforts are being made to minimize batch variations in omics datasets, which otherwise would be a source of technical noise.
3. The project in its current phase is working towards harmonized protocol for analysis of sample and subsequent data analysis and reporting.
4. Field samples/measurements: We will follow ISO45001 standards, additional overview of environmental sampling strategies is described in the EXIMIOUS profile paper (Ronsmans et al., 2022).
5. Digital data: There will be version control and audit, raw files will be stored in read only folder and secure workspace will be provided for analysis of data.

## 8. Partner survey on data management

While the DMP is a guidance document providing a roadmap for the entire duration of the project, the process of data management is a continuous procedure. EU GDPR guidelines and feedback from different partners were considered in preparation of the DMP and the data management platform. In the third year of the project, as data collection had been initiated, we wanted to know whether the concerns expressed at the beginning of the project of all involved partners have been addressed.

To this end an anonymous Qualtrics survey (Supplementary file) was sent out to all 15 partners. We asked questions related to their area of expertise, their role in EXIMIOUS or other similar projects. As part of the survey, participants were asked to rank a minimum of five and a maximum of ten points as "key challenges" from a list of 25 possible options. The list was randomized for each participant as an effort to reduce bias.


The survey was completed by 27 participants, mostly researchers (67%), but also ethical, legal and privacy experts and project advisors (Supplementary Fig. 1). Of the 27 participants, 14 (52%) were actively involved in coordinating/drafting the legal/ethical documents for EXIMIOUS (Supplementary Fig. 1 a, b). Of the remaining 13 participants, 4 (31%) completed the form based on their experience with other exposome projects within the EHEN network (Supplementary Figure 1 c). Based on the results of the survey (Supplementary Figure 1 d) a ranked list of key challenges was obtained, and data and material transfer and sharing were among the highest ranked options. Among the ranked list most if not all these concerns have been addressed through:

1. Data and material transfer agreement - addressed in DMP, MTA/DTA
2. Data Sharing - addressed in CA, DMP, MTA/DTA, joint controllership
3. Data ownership - addressed in CA, DMP, MTA/DTA, joint controllership
4. Consent - addressed in DMP, ethics deliverables (informed consent)
5. Privacy - addressed in CA, DMP, MTA/DTA
6. Data access - addressed in CA, DMP
7. Reuse of data - addressed in DMP
8. Anonymity - addressed in DMP and ethics deliverables
9. Biobank - addressed in DMP and ethics deliverables
10. Open data - addressed in CA, DMP, MTA/DTA

While many of these challenges have been addressed, it is important to acknowledge that the process of research data management is complex and dynamic. And as a result, through various stages of the data lifecycle the process of organizing and maintaining the data may evolve accordingly. This includes (re)organization of folder (and sub-folder) structure, use of different version of iRODS interface (other than ManGO), added security and authentication measures, creation of different data access groups based on the analysis to be performed etc.


**Table 2**

The EXIMIOUS checklist for data management and recommendations based on the challenges addressed and measures followed by the project.

		Task	Means to achieve task (EXIMIOUS examples)	EXIMIOUS timeline (months)	Resources (Uploaded version of the documents might be available in course of time)	Check list
<b>Project pre-award</b>						
<b>Defining data collection objective</b>		Defining how all of the data is relevant and limited to the purposes of the project ('data minimization' principle)	Define clearly the research objectives, objectives of data collection, collection of personal information and defining how it is directly relevant and necessary to accomplish a specified task/objective. <i>(Additional data minimization steps in EXIMIOUS are described in the manuscript)</i>	Pre-application	<a href="https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725">https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725</a>	<input type="checkbox"/>
	<b>Ethics and Integrity</b>	Identify ethical considerations	Consult the EC self-assessment and the local ethics office of the applicant's institute	Pre-application		<input type="checkbox"/>
		Complete the ethics self-assessment	Consult the EC self-assessment and the local ethics office of the applicant's institute	Pre-application	<a href="https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf">https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf</a>	<input type="checkbox"/>
	Describe the clinical study <i>(For calls that involve clinical studies, participants might be required to add additional document to the application as annex to the proposal)</i>	Consult the EC self-assessment and the local ethics office of the applicant's institute	Pre-application	<a href="https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/af/information-on-clinical-studies_he_en.docx">https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/af/information-on-clinical-studies_he_en.docx</a>	<input type="checkbox"/>	
	Prepare a brief outline describing data management and personal data protection as requested by the funding agency	Consult the Research Data Management Helpdesk at applicants host institute	Pre-application	<a href="https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf">https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf</a>	<input type="checkbox"/>	
<b>Project Post-Award</b>						
<b>Contracts and agreements</b>	<b>Ethics review, Grant signature</b>	Ethics review by funder	Reviewed by ethics officer (funding agency) multiple 'ethics requirements' become contractual obligations and are implemented in the grant agreement before the agreement could be signed. Ethics requirements due after project start are included in the grant agreement in the form of 'ethics deliverables' and are submitted to the EC.	0	<a href="https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf">https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf</a>	<input type="checkbox"/>
		Grant Agreement (GA)	GA signed between funding agency and beneficiaries.	0	<a href="https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf">https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf</a>	<input type="checkbox"/>
	<b>Agreement between consortium members</b>	Consortium Agreement (CA) between the consortium members	Drafted by legal department(s) of partner institute(s) following DESCA model Consortium Agreement and signed by representatives of each partner within the consortium.	1	<a href="https://www.desca-agreement.eu/desca-model-consortium-agreement/">https://www.desca-agreement.eu/desca-model-consortium-agreement/</a> <i>(DESCA for Horizon Europe)</i>	<input type="checkbox"/>
<b>Data Management Workflow</b>	<b>Data summary</b>	Define what types and formats of data the project will generate or re-use	By project partners, preparing an exhaustive list of all data types, sources, origins, processes of collection and formats for all data sets.	3	The latest version of the template will be downloadable from the EC Funding and Tenders Portal.	<input type="checkbox"/>
		Define the purpose of the data generation or re-use and its relation to the objectives of the project			<a href="https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm">https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm</a>	<input type="checkbox"/>
		Identify the expected size of the data that you intend to generate or re-use			(ERC info document-Open Research Data and Data Management Plans)	<input type="checkbox"/>
	<b>DMP V1</b>	Identify the origin/provenance of the data, either generated or re-used			(Template horizon 2020 data management plan- DMP)	<input type="checkbox"/>
		Identify to whom might the data be useful ('data utility'), outside the project				<input type="checkbox"/>
		Prepare and submit a first version of your DMP (as a deliverable) within the first 6 months of the project.	Based on the data summary generated, prepare a data management plan describing the security measures to be	6		<input type="checkbox"/>


(continued on next page)

Table 2 (continued)

	Task	Means to achieve task (EXIMIOUS examples)	EXIMIOUS timeline (months)	Resources (Uploaded version of the documents might be available in course of time)	Check list
SOP V1	SOP for collection, storage and shipment of biological samples	<p>implemented. This document is a draft version which would be revised during the entire duration of the project, incorporating information as the project progresses and FAIR provisions are implemented.</p> <p>Based on the study objectives defined in the project, analysis to be performed and technical know-how of the analyzing partners, clear protocols were drafted for each of the methods to ensure good quality samples. This protocol was then organized for optimal collection during field work- clearly defining -hygiene practices; required material for collection, processing and storage; the order of sample collection; instructions for sample handling; time allowed from collection to processing and storage among others. At every step of the process, feedback was collected from cohort owners, lab technicians about ease and practicality of sampling and the protocol was updated for clarity.</p>	8		□
Ethics	Submit application(s) for ethical approval or amendments (as might be applicable) for each included study population	Based on the DMP V1 and SOP V1, cohort owners submitted their application to their respective institutional ethical committee for ethical approval keeping both National and EU regulations in mind.	8	<a href="https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm">https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm</a>	□
Data management platforms	Select suitable data management platforms and define security measures	Consult the data management helpdesk at the coordinating institute and reach consensus among partners. Some of the selected platforms include: SharePoint- web-based collaborative platform; REDCap- Research Electronic Data Capture; Qualtrics- web-based survey tool; ManGO - web application designed to work alongside iRODS	Continuous		□
FINDABLE data provisions	<p>F1. (Meta)data are assigned a globally unique and persistent identifier</p> <p>F2. Data are described with rich metadata (defined by R1 below)</p> <p>F3. Metadata clearly and explicitly include the identifier of the data it describes</p> <p>F4. (Meta)data are registered or indexed in a searchable resource</p>	<p>1. Persistent identifier assigned to ensure traceability of data, from collection, to transport and sharing and processing</p> <p>2. Each data type generated in the project is described with rich metadata, metadata searchable in data repository of EXIMIOUS (iRODS-ManGO)</p> <p><i>(Additional provisions are described in the manuscript)</i></p>	Continuous	<p><a href="https://doi.org/10.1038/sdata.2016.18">https://doi.org/10.1038/sdata.2016.18</a></p> <p><a href="https://force11.org/info/the-fair-data-principles/">https://force11.org/info/the-fair-data-principles/</a></p> <p><a href="https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/">https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/</a></p> <p><i>(FAIR Data Management in Horizon, 2020)</i></p>	□ □ □ □
ACCESSIBLE data provisions	<p>A1. (Meta)data are retrievable by their identifier using a standardized communications protocol</p> <p>A1.1 The protocol is open, free, and universally implementable</p> <p>A1.2 The protocol allows for an authentication and authorization procedure, where necessary</p> <p>A2. Metadata are accessible, even when the data are no longer available</p>	<p>1. iRODS (ManGO) used for EXIMIOUS data repository</p> <p>2. README file outlining process of data upload, download, addition of metadata, managing permission, search option</p> <p>3. Access requires authorization, data access audited</p> <p>4. Agreement for code of conduct for Data managers</p> <p>5. Secure folder with restricted access, Shared folder for open data (within the project)</p> <p>6. Different access rights created to make data “as open as possible and as closed as necessary”</p> <p><i>(Additional provisions are described in the manuscript)</i></p>	Continuous	<p><a href="https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf">https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf</a></p> <p><a href="https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm">https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm</a></p> <p><i>(European Research Council (ERC) Guidelines on Implementation of Open Access to Scientific Publications and Research Data)</i></p> <p><a href="https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf">https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf</a></p>	□ □ □ □
INTEROPERABLE data provisions	I1. (Meta)data use a formal, accessible, shared, and broadly	1. Readme file with description of file and resources to understand content	Continuous	<p><a href="https://rdamsc.bath.ac.uk/">https://rdamsc.bath.ac.uk/</a></p>	□

(continued on next page)

Table 2 (continued)

	Task	Means to achieve task (EXIMIOUS examples)	EXIMIOUS timeline (months)	Resources (Uploaded version of the documents might be available in course of time)	Check list
	applicable language for knowledge representation. I2. (Meta)data use vocabularies that follow FAIR principles I3. (Meta)data include qualified references to other (meta)data	2. Rich metadata, with information of source of data, sample volume, sample preparation, analysis method, quality control, related publications (DOI) 3. Information on associated data 4. Data analysis script/protocol <i>(Additional provisions are described in the manuscript)</i>			<input type="checkbox"/> <input type="checkbox"/>
<b>REUSABLE data provisions</b>	R1. Meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (Meta)data are released with a clear and accessible data usage license R1.2. (Meta)data are associated with detailed provenance R1.3. (Meta)data meet domain-relevant community standards	1. Project-level documentation (Study design, hypothesis, protocol, methodology, instrumentation) 2. File level documentation (explaining how datasets are linked, data dictionary explaining meaning of variables) 3. Metadata includes information of level of data, sampling and analysis method, version control 4. Data with restricted access- Data licensing to be decided on a case basis <i>(Additional provisions are described in the manuscript)</i>	Continuous		<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<b>Data security</b>	Define the provisions that are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data) Define how the data will be safely stored in trusted repositories for long term preservation and curation	In brief, <i>SharePoint</i> collaboration site has an automatically set backup procedure (daily). <i>REDCap</i> : backups can be generated at any moment; since data is automatically exported to iRODS (ManGO interface) on a daily basis, our back-up procedures situate mainly on iRODS. In iRODS (ManGO interface), the central database, we can set up different levels of secured storage: 1. The EXIMIOUS project on iRODS (ManGO interface) is secured with a two-factor authentication linked to a KU Leuven account (also for external users). An iRODS (ManGO interface) administrator needs to grant access. 2. Different user roles are set, defining the user permissions and rights. 3. Furthermore, users will only have access to Data Collections they need to access, thereby restricting access to all data. A version of the raw data is stored in both a Read Only (RO) folder and a Read-Write (RW) folder. No changes can be made to the RO folder, thus allowing us to restore the raw data if needed. Automated back-ups are made at regular intervals: daily, weekly, monthly and yearly. <i>(Additional provisions are described in the manuscript)</i>	Continuous		<input type="checkbox"/> <input type="checkbox"/>
<b>DMP V2</b>	Review and revise the DMP through the course of the project. Submit a revised version of the DMP when recommended by the funding agency	DMP is a living document and therefore it was revised several times during the first 2 years period based on feedback from data managers, ethical committee and project partners. This includes more detailed information on the type of data to be analyzed, implementation of the FAIR data management structure by defining metadata, folder structure and implementation of the security measures. The revised document (DMP V2) was submitted to the EC as updated deliverable of the DMP V1.	18		<input type="checkbox"/>

(continued on next page)

Table 2 (continued)

		Task	Means to achieve task (EXIMIOUS examples)	EXIMIOUS timeline (months)	Resources (Uploaded version of the documents might be available in course of time)	Check list
<b>SOP V2</b>		Revise SOP for clarity and submit as a revised deliverable	SOP for collection, storage and shipment of biological samples were revised based on feedback from partners (cohort owners and lab technicians) on the ease of following the protocol, and a pilot to check the quality of samples.	24		<input type="checkbox"/>
<b>Biobank registry</b>		Register study with biobank at centers that will be working with human bodily material	To assure traceability of human bodily material according to National and EU legislation, all HBM that is/ will be obtained has been registered with the respective biobanks (for instance EXIMIOUS has been registered with the UZ/KU Leuven biobank for the coordinator center and at NIPH, Norway).	24		<input type="checkbox"/>
<b>Ethics</b>		Receive ethical approvals	Ethical approval was obtained by all cohort owners, after revision and clarification of comments provided by the Institutional ethics committee.	10–24		<input type="checkbox"/>
<b>Clinical study registry</b>		Registration of clinical study in a WHO- or ICMJE- approved registry (when applicable) that also allows later posting of study results.	While EXIMIOUS is not a clinical study in traditional sense, to make the project findings FAIR, EXIMIOUS was registered with the ISRCTN registry as an observational case-control study (ISRCTN24225493; <a href="https://doi.org/10.1186/ISRCTN24225493">https://doi.org/10.1186/ISRCTN24225493</a> )	24		<input type="checkbox"/>
<b>Contracts and agreements</b>	<b>Agreement between consortium members</b>	Material transfer agreement Data transfer agreement Joint controllership agreement	In EXIMIOUS, this was done through an overarching agreement covering all partners who signed a common "Agreement on Standard Operation Procedure (SOP), Study and template MTA".  This legal document was drafted for EXIMIOUS by the coordinating institute- in this case KU Leuven (Research data management and KU Leuven Research & Development) and reviewed by the legal departments of all partner institute.	24		<input type="checkbox"/>

9. The EXIMIOUS “DMP CHECK” tool for FAIR data management

Based on the measures adopted in EXIMIOUS to ensure FAIR data management, we put together a checklist and a series of recommendations from our experience to address each of the aspects included in the checklist (Table 2). This checklist will be made available through the EXIMIOUS project website as part of the toolbox. The checklist “DMP CHECK” (Preliminary visual identity: Fig. 7) will be updated till the end of the project with any new information and will be available on the project website.

10. Conclusions

EXIMIOUS is one of the nine H2020 exposome projects which is part of the EHEN. As part of the EXIMIOUS project mostly constituted of researchers, we navigated the complex landscape of data protection and security, ethics and making data FAIR through our DMP. While all projects (H2020, Horizon Europe and beyond) are obligated to conform to many of these regulations, researchers often lack the right tools and/or accurate understanding of the ethical/legal framework to independently address such challenges. More than two years into the project, we

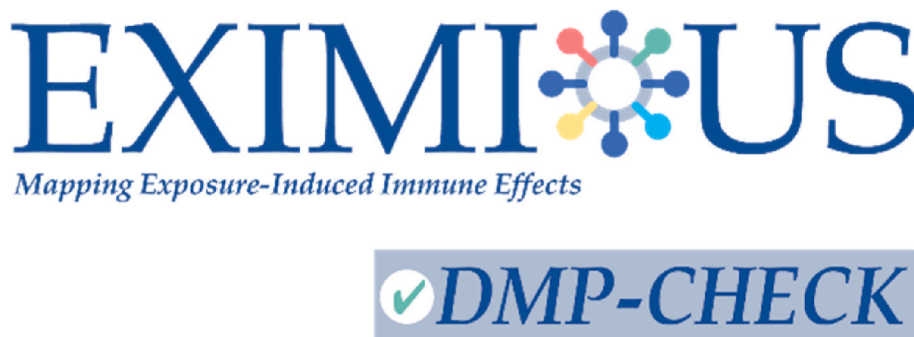


Fig. 7. Preliminary visual identity of the EXIMIOUS “DMP CHECK” tool for FAIR data management.





- Lambrechts, A., Lavery, A., Lenaerts, C., Lepecuchel, L., Lindeman, B., Liston, A., Loeb, L., Madsen, A.M., Mertens, G., Flachs, E.M., Moldovan, H., Mueller, J., Muñoz, X., Nawrot, T.S., Ndaya-Kabundi, A., Nygaard, U.C., Olah, P., Plusquin, M., Roca, C., Ronsmans, S., Sejbæk, C.S., Smythe, C., Hougaard, K.S., Storvik, N., Suci, N., Tang, A., Tunney, M., Rasmussen, P.U., Vanoirbeek, J., Vermeir, P., Vernimme, P., Verpaele, S., Vogel, U., Winther, N., Yserbyt, J., 2022. The EXIMIOUS project-Mapping exposure-induced immune effects: Connecting the exposome and the immunome. *Environ. Epidemiol.* 6, E193. <https://doi.org/10.1097/EE9.000000000000193>.
- Schiermeier, Q., 2018. Data management made simple. *Nature* 555, 403–405. <https://doi.org/10.1038/D41586-018-03071-1>.
- Spidlen, J., Breuer, K., Brinkman, R., 2012. Preparing a minimum information about a flow cytometry experiment (MIFlowCyt) compliant manuscript using the international society for advancement of cytometry (ISAC) FCS file repository (FlowRepository.org). *Curr. Protoc. Cytom.* <https://doi.org/10.1002/0471142956.CY1018S61/FULL>.
- Waubant, E., Lucas, R., Mowry, E., Graves, J., Olsson, T., Alfredsson, L., Langer-Gould, A., 2019. Environmental and genetic risk factors for MS: an integrated review. *Ann. Clin. Transl. Neurol.* 6, 1905–1922. <https://doi.org/10.1002/ACN3.50862>.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hoof, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3 (1 3), 1–9. <https://doi.org/10.1038/sdata.2016.18>, 2016.
- Zare Jeddi, M., Virgolino, A., Fantke, P., Hopf, N.B., Galea, K.S., Remy, S., Viegas, S., Mustieles, V., Fernandez, M.F., von Goetz, N., Vicente, J.L., Slobodnik, J., Rambaud, L., Denys, S., St-Amand, A., Nakayama, S.F., Santonen, T., Barouki, R., Pasanen-Kase, R., Mol, H.G.J., Vermeire, T., Jones, K., Silva, M.J., Louro, H., van der Voet, H., Duca, R.C., Verhagen, H., Canova, C., van Klaveren, J., Kolossa-Gehring, M., Bessems, J., 2021. A human biomonitoring (HBM) Global Registry Framework: further advancement of HBM research following the FAIR principles. *Int. J. Hyg Environ. Health* 238, 113826. <https://doi.org/10.1016/j.IJHEH.2021.113826>.