Project number: 874662
Project Acronym: HEAP

Project Title: Human Exposome Assessment Platform
Project website URL:
Project Coordinator: Joakim Dillner
Organization: KI
E-mail: Joakim.Dillner@ki.se

Work Package 7
Data interoperability and data sharing

Work package Leader: Heimo Müller
Organization: Medical University of Graz
E-Mail: heimo.mueller@medunigraz.at

Project Deliverable

D7.1: Data Management Plan

Deliverable due date: 2020-05-31
Deliverable due month: 05

Document history

Version	Date	Changes	By	Reviewed
0.1	2020-01-14	First draft	Heimo Müller Roxana Merino	Joakim Diller
0.2	2020-03-12	HEAP specific requirements	Heimo Müller	Roxana Merino
0.3	2020-04-27	Add all information on data sources	Roxana Merino all leads of research tasks	Heimo Müller
0.4	2020-04-27	Alignment with system architecture	Heimo Müller Stefan Negru	Roxana Merino
0.5	2020-05-05	Ethical and Legal Aspects	Martin Boeckhout Evert-Neb van Veen	WP Heimo Müller
0.6	2020-05-15	Input from all HEAP WPs	All WP leads	WP leads
0.7	2020-05-27	Final Review	Heimo Müller Roxana Merino	full consortium
1.0	2020-05-29	FINAL	Heimo Müller Bettina Kipperer	Joakim Dillner Roxana Merino
1.1	2022-04-29	Update Lifestyle cohort draft	Patrick Nitsche	Catarina Dias

Executive Summary

This deliverable is the Data Management Plan (DMP) of the Human Exposome Assessment Platform (HEAP), which aims to create a technical research platform to assess the impact of the internal and external exposome in human health. HEAP will collect a significant amount of data from 6 research projects:

HEAP - Cervical screening cohort

HEAP - Maternity cohort

HEAP - HPV vaccination cohort

HEAP - Consumer cohort

HEAP - Wearable data collection study

HEAP - Lifestyle cohort

All data objects from human patients, pathogenic organisms, as well as global genetic resources will be stored securely, with assurance and documentation that any processing and analysis will be in line with HEAP ethical and legal framework. The technical infrastructure to manage all data and metadata in the HEAP's Information Commons (IC) will be developed by WP 10 (Secure Infrastructure for big data).

The dedicated work package WP 7 (Data interoperability and data sharing) is committed to make the HEAP Information Commons (IC) FAIR – findable, accessible, interoperable and re-usable – including what data the project will generate, whether and how it will be made accessible for verification and re-use, and how it will be curated and preserved.

We will connect the HEAP Information Commons (IC) to the European Human Exposome Network and to the EOSC via consistent, FAIR annotation practices including provenance information. Based on the ethicolegal analysis of the data, value chain access to data and cloud services will be underpinned by a shared federated authentication and authorization infrastructure (Life Science AAI) supported by a formal governance model developed by the WP2 (Ethics and Regulations).

Table of Contents

Introduction

The Human Exposome Assessment Platform (HEAP) aims to create a technical research platform to assess the impact of the internal and external exposome in human health. The final contribution will be a platform that can be deployed and scaled at any research institution and will make available high-quality, well curated exposome data from five different cohorts.

In order to achieve those objectives, a significant amount of data will be collected, processed, generated and deposited in the HEAP Information Commons (IC). In addition to HEAP deliverables, scientific publications (e.g. peer-reviewed research articles) will be produced. According to the European Commission (EC), "research data is information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion, or calculation".

HEAP Information Commons (IC) is the data repository infrastructure that provides on-demand data accessibility and availability following the HEAP ethical and regulatory framework. It includes the physical resources to store and archive HEAP data and the interfaces to make this data available for analysis and future research through HEAP PaaS (Data & Feature Warehouse and Knowledge Engine).

In general terms, the HEAP Information Commons (IC) will follow the "FAIR" principles, meaning "Findable, Accessible, Interoperable and Re-usable"). The FAIR principles will ensure soundly managed data, leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse. The data will be made findable and accessible within the Consortium, and to the broader research community, stakeholders and policy makers. Also, data must be compliant with national and European ethic-legal framework, such as the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679).

Within Horizon 2020, the Data Management Plans (DMPs) describe the data management life cycle for all data to be collected, processed and/or generated by a project. Former, it contains information on the nature of the data, on the handling and processing of research, and standards / methodology. An essential information of the DMP is whether and how data will be shared or made open access, and how the data will be curated and preserved.

This document contains information on the general HEAP strategy concerning data management within the form of an overarching data management plan. It agrees on a strategy for managing HEAP data in compliance with the requirements of Horizon 2020.

The Consortium Agreement will determine strict compliance to the DMP. Because of the heterogeneity of the data that will be collected, processed or generated within HEAP, and as a result of the level of detail needed, each specific research project will also have to compile project specific DMP's, using as basis the present overarching DMP. Since HEAP is strongly

based on already existing cohorts and data collections, a common consensus between partners and stakeholders is necessary to collect, process and use the data.

WP 7 is dedicated to data interoperability and data sharing. It starts in task 7.1 with the DMP, which addresses all relevant aspects of making data FAIR – findable, accessible, interoperable and re-usable – including what data the project will generate, whether and how it will be made accessible for verification and re-use, and how it will be curated and preserved. The DMP is the "bible" for all further tasks in WP7, dedicated to Standards, Recommendations, Interoperability Training, development of a FAIR toolbox, organization of Bring Your Own Data (BYOD) workshops, coordination of the liaisons to ESFRI research infrastructures (e.g. ELIXIR, BBMRI), the European Open Science Cloud (EOSC) and large-scale coordination projects as EOSC-Life, EJP-RD, GA4GH, GO FAIR, Global BioImaging, Galaxy, Bioconda and RDA, knowledge exchange with all other projects in the Human Exposome cluster, and finally the generation and publishing of the HEAP catalogue and assessment of quality of service, data quality and interoperability, availability of FAIR data and services, technical sustainability and accessibility of data.

The overarching DMP is structured according to the H2020 templates for Data Management Plans and provided in the next chapters:

1. Data Summary
2. FAIR data
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other issues

The last section provides an action plan, which presents important topics requiring progress and/or update in future versions of the DMP.

1 Data Summary

The overall goal of HEAP is to enable global collaborative exposome research towards cost-effective health interventions. This will be provided by a platform enabling combination of i) advanced, joint information technology, ii) advanced exposome measurement technology and iii) uniquely large longitudinally followed population-based cohorts that will provide valuable actionable knowledge about women´s health and healthy child-bearing.

HEAP will provide a system for obtaining, managing, sharing and analysing exposome data. Large- scale, comprehensive and population-based systems will be exploited to build sustainable systems for measuring the human exposome and its impact on health. Innovative wearable exposure sensors will monitor exposures to pregnant women in relation to healthy childbearing. We foresee that all the data and generated knowledge from that constellation of female & childbearing exposures outcomes can lead to precision health of women and children towards enhancing effective actions from healthcare and society.

In order to achieve this, the HEAP consortium will launch innovative scientific analysis methods, beyond state- of-the-science technologies, high throughput epigenomics and metagenomics methods, computation resources, wearable exposome sensors, and applied Artificial Intelligence.

Expertise and technical components will be integrated into a global research resource that enables ethical and efficient management and processing of massive data from geographically distributed large-scale population cohorts. HEAP will create a standardized, integrated and generic informatics platform that will enable the creation of collaborative networks towards the joint production of consistent and actionable knowledge to tackle the effects of exposome on health and society.

Data will be generated in 6 research sub-projects, all contributing to the HEAP information commons (IC). The research sub-projects are the

1. HEAP - Cervical screening cohort
2. HEAP - Maternity cohort
3. HEAP - HPV vaccination cohort
4. HEAP - Consumer cohort
5. HEAP - Wearable data collection study
6. HEAP - Lifestyle cohort

For all the sub-projects the DMP provides information according to the DMP template describing the

a) Purpose and relation to the objectives of the project,
b) Types and formats of data generated and collected,

c) Origin of the data and usage of legacy data items,
d) Expected size of the data collection and handling of big data objects,
e) Data utility, to whom data will be useful during and after the project.

1.1 HEAP - Cervical screening cohort

Leading partner (PI): KI, Joakim Dillner

a) Purpose and relation to the objectives of the project

The main purpose of cervical screening is to screen all women between 23 -70, to prevent the occurrence of invasive cervical cancer by detecting and removing precancerous lesions, which if left untreated, could develop into invasive cervical cancer. As a result, the cervical screening program has provided to the National Clinical Cytology Biobank, around 800 000 samples corresponding to 450 000 unique women. Every year, 140 000 new women are enrolled. The use of the Swedish Cervical Cancer screening cohort aims to evaluate internal exposome factors (virus, bacteria, others) in women as measured in cervical exfoliated cells.

The cervical screening cohort will provide samples from healthy tissue, pre-neoplastic lesions and cancer. The sequencing analysis of specimens together with metadata aims to identify and characterize risk factors that may impact risk for having disease or being healthy. The data will be provided to the central HEAP database for systematic analysis of the exposome in relation to human disease.

b) Types and formats of data generated and collected

The data collected consists of

- Consent Information
- Register data (SAS, MySQL)
- Sample data (LIMS, CSV format)
- Exposome data (sequencing data following standards as in TCGA database)

c) Origin of the data and usage of legacy data items

Data is collected from the samples taken in the organized cervical screening program in Stockholm region, capital region of Sweden, and national health registers, and quality register databases. In addition, the following legacy data will be used:

- National Cervical Screening Register data (MySQL format)
- National health registers at National Board of Health and Welfare

d) Expected size of the data collection and handling of big data objects

Terabyte scale (>500 Tb). The cohort comprises >800 000 samples from 450 000 women. Although we cannot sequence every specimen, data on every patient will be listed.

e) Data utility, to whom data will be useful

Researchers: Data will elucidate important mechanisms for human health, including screening programs. Data will ultimately be rendered available to the community through open access in a GDPR-compatible format.

Citizens: Results will help citizens understand risk factors increasing the risk of cancer and other diseases.

Society including national environmental, health and screening authorities, non-governmental organizations, European agencies, and the European Commission.

1.2 HEAP - Maternity cohort

Leading partner (PI): UOULU, Heljä-Marja Surcel; KI/FICAN-MID, Matti Lehtinen.

a) Purpose and relation to the objectives of the project

The aim of the HEAP Maternity cohort is to establish a population-based cohort biobank. The data collected is of calendar time and over generations, hence the relation to HEAP is to understand the exposome in this context, over time and over generation.

b) Types and formats of data generated and collected

The baseline cohort data for approx. 1.000.000 healthy participants consist of consented participation data. Registry linkage will be done by the original data provider only and anonymous data (not pseudo-anonymised) will be delivered.

- Individual Consent Date
- Randomized and encrypted participation data (e.g. serum sampling dates. sample volumes, sample series)(Encrypted CSV)
- Health Information (Word/PDF/HTML)

c) Origin of the data and usage of legacy data items

The origin of the data is the Finnish National prenatal screening, with additional FMC screening results (HIV, syphilis, hepatitis B, rubella). Data will be provided by CSC through controlled access after approval of data owner(s).

d) Expected size of the data collection and handling of big data objects

We plan to use a subset from 10 x 2.000.000 participants for microbe-omics and epidemiological data. The expected size for the baseline data is in gigabyte range, for the sequence data in the terabytes range.

e) Data utility, to whom data will be useful during and after the project

Researchers: Purpose of multi-faceted study designs for epidemiological analysis (including serology and in this context also microbe-omics analysis).

Citizens: Results will help citizens understand risk factors increasing the risk of cancer and other diseases.

Society including national environmental, health and screening authorities, non-governmental organizations, European agencies, and the European Commission.

1.3 HEAP - HPV vaccination cohort

Leading partner (PI): KI/FICAN-MID, Ville Pimenoff and Matti Lehtinen.

a) Purpose and relation to the objectives of the project

The aim of the HEAP HPV vaccination cohort is to establish a randomised stratified intervention cohort. The exposome in this context is related to the impact of exposure to health interventions over time.

b) Types and formats of data generated and collected

The baseline data for 60.000 healthy participants consists of participating cohort data. Registry linkage will be done by the original data provider only and anonymous data (not pseudo-anonymised) will be delivered.

- Individual Consent Data
- Randomized and encrypted participation data (Encrypted CSV)
  intervention:
  - vaccination and/or screening status
  - sampling dates for serum
  - oral gargle
  - cervical DNA samples
  - sample series) (Encrypted CSV)
- Documentation (Word/PDF/HTML)
- Participant personal data: HPV centre internal DB
- Sample data: HPV centre LIMS
- Sequencing data: HPV centre data storage (server and cluster)

- The data will be formalised in HEAP PaaS

c) Origin of the data and usage of legacy data items

All data collected come from an Intervention trial follow-up with additional STD follow-up results (cytology, HPV and chlamydia) available. Data will be provided by CSC through controlled access after approval of data owner(s).

d) Expected size of the data collection and handling of big data objects

We plan to use a subset from up to 50 HPV and chlamydia types x 60.000 participants x (5 to 10) for microbiome metagenomics data. The expected size for the baseline data is in gigabyte range, for the sequence data in the terabytes range.

e) Data utility, to whom data will be useful during and after the project

Researchers: Purpose of microbe-omics, epidemiological and impact analysis (including serological and nucleic acid samples from different anatomical sites)

Citizens: Results will help citizens understand risk factors increasing the risk of cancer and other diseases.

Society including national environmental, health and screening authorities, non-governmental organizations, European agencies, and the European Commission.

1.4 HEAP - Consumer cohort

Leading partner (PI): SSI, Frederik Trier Møller

a) Purpose and relation to the objectives of the project

The aim of the HEAP Consumer cohort is to use consumer purchase data to continuously model and assess the household consumed exposome's impact on health. The aim is to identify purchases that may impact health and to provide input to an efficient platform for handling and analysing large amounts of data on environmental exposures and their health effects.

b) Types and formats of data generated and collected

The data collected consists of:

- Consumer purchase data: JSON (received format) saved in MS SQL DB
- Scientific reports Word/PDF/HTML
- Data from registers and other research projects SQL DB, CSV

c) Origin of the data and usage of legacy data items

Data is collected from commercial digital receipt providers (such as start-up company Storebox) and if possible, loyalty programmes (such as COOP, pending approval). The data is re-used from other sources such as Danish registers, quality databases, or research projects, as well as digital receipt solutions and loyalty programs.

d) Expected size of the data collection and handling of big data objects

Consumer purchase data: Depends on the final number of participants and the number of years data are collected; probably hundreds GB.

Report and papers, presentations etc. few MB.

e) Data utility, to whom data will be useful during and after the project

Researchers: Methods will help facilitate more research in the field. Data may be reused for all projects covered by the purpose of the data collection within the duration of the project.

Citizens: Results will help citizens identify and avoid products increasing the risk of one or more diseases.

Society including national environmental, health and food authorities, researchers interested in public health, European agencies such as EFSA, ECDC.

European Commission: As results may lead to lifestyle communication models that may support lasting habit change improving the individual's health, and epidemiologic surveillance systems, facilitating rapid responses to emerging threats.

1.5 HEAP - wearable data collection study

Leading partner (PI): KI, Michael Snyder; KI/FICAN-Mid, Matti Lehtinen; OU, Heljä-Marja Surcel.

a) Purpose and relation to the objectives of the project

The aim of the HEAP wearable data collection is to establish a proof-of-concept for environmental exposome with a seasonal follow-up data. The data collected is environmental data created for the purpose of HEAP, and to particularly understand human exposome over calendar year time. Pilot study to monitor exposures on pregnant women in relation to healthy childbearing and produce continuous personal exposome profiling of the participants.

b) Types and formats of data generated and collected

The data collected consists of

- Consent Information
- Geo-location data
- Environmental parameters (temperature, humility, airflow rate, etc.)
- Sample acquisition metadata
- Analysis of airborne particulate matters and toxic substances
- Metabolomics analysis data (high-resolution mass spectrometry coupled with liquid chromatography)
- Sequence Analysis (Illumina NovaSeq platform)

c) Origin of the data and usage of legacy data items

All wearable device data will be collected for the purpose of HEAP. We will follow 100 pregnancies to pilot investigating how the exposome influences human health. The pilot will follow 100 pregnant women participating in the Finnish Maternity Cohort to investigate how the exposome influences the health of the mother and the baby. Participants will be either asked to carry the device with them or use the device at their home to monitor their personal exposures. Each location would be matched by one sample, e.g. one sample for workdays (up to 5 days) and one for weekend trips to a national park. Samples will be collected by the participants (at least twice per week) and stored temporarily in a -20 degrees freezer holder before being delivered to the lab and stored in -80 degrees freezer monthly. The samples will be collected and sent to the lab monthly. When special circumstances such as flu happen, more sampling may be requested.

d) Expected size of the data collection and handling of big data objects

We plan to use full metabolome and microbiome data from 100 participants for comprehensive microbe-omics and epidemiological analysis using collect data from 100 devices. The expected size for the baseline data is in 10-50 gigabyte range, and in the terabytes range for the sequence data and metabolomics analysis data.

e) Data utility, to whom data will be useful during and after the project

We aim to pilot monitoring of multiple exposures and health measurements from wearable sensors that collect airborne particulate matters and toxic substances as well as capturing particular matters, potentially consisting of virus particles, bacteria, fungal spores, animal debris, and plant pollens. We will provide the entire process from device manufacturing, recruitment of participants, data acquisition, management, analysis and sharing of data and results. This solution will provide progress towards personal exposome profiling and agnostic identification of environmental risk factors.

1.6 HEAP - Lifestyle cohort

Leading partner (PI): UIBK, Martin Widschwendter

a) Purpose and relation to the objectives of the project

The main aim of the HEAP Lifestyle cohort within the TirolGESUND study (NCT05678426) is to investigate dynamics in DNA methylation risk signatures associated with ageing and women's cancers in the presence of non-pharmacological disease-preventing interventions. Specifically, the effect of two lifestyle changes, smoking cessation or interval fasting with/without ketogenic dietary supplementation, in addition to guided, targeted exercise, will be evaluated over an intervention period of 6 months. Several biological samples that may act as surrogate samples for tissues at risk of (age-related) disease will be collected at baseline and every two months for 6 months. Optionally, biological samples will be collected at 12 and 18 months after initiation of the intervention. The results of the study in the HEAP Lifestyle cohort will inform future preventive intervention studies.

b) Types and formats of data generated and collected

The data collected consists of

- Consent information
- Fitness Data (e.g. Garmin Smart Watch)
- Methylation data from matched longitudinal samples (cervical, buccal, blood)
- Clinical data, including
    - Haemogram values
    - Sports medicine measurements (ergonometry; baseline and 6 months only)
    - Vascular endothelial thickness (baseline and 6 months only)

c) Origin of the data and usage of legacy data items

All data originate from measurements derived from biological specimens collected, clinical observations made, or fitness tracker devices worn by participants, within the TirolGESUND study. Following initial publication of the study data, DNA methylation data and other clinical data will be deposited in a secure controlled access repository (European Genome-Phenome Archive). Data cannot be made available without restrictions due to the nature of the informed consent given by participants. Data also can only be shared with third countries that are GDPR compliant.

d) Expected size of the data collection and handling of big data objects

Cohort comprises methylation data from > 1,800 samples from 156 women (3 samples per visitation, 4 visitations), and mutation data from a subset.
Analysis data: GB-1TB.

e) Data utility, to whom data will be useful during and after the project

Researchers: Data will provide an initial insight into DNA methylation dynamics in response to lifestyle changes and will inform future large scale disease-prevention studies. Moreover, researchers investigating the effect of interventions on age markers and DNA methylation dynamics more generally will benefit from this well-characterised cohort.
Citizens: The results from this study will help citizens understand factors decreasing the risk of cancer or other age-related disease, that individual factors play a key role in cancer development, and that individual risk prediction may be an option supporting cancer prevention in the future.

Society: Evaluation of novel markers to monitor effectiveness of preventive measures.

2 FAIR data

HEAP will provide a technical infrastructure to manage all data and metadata in the HEAP's Information Commons (IC). All data objects from human patients, pathogenic organisms, as well as global genetic resources will be stored securely, with assurance and documentation that any processing and analysis will be in line with HEAP ethical and legal framework.

HEAP will connect the Information Commons to the European Human Exposome Network and to the EOSC via consistent, FAIR annotation practices including provenance information. HEAP's mission is to contribute data to the exposome cluster to collaboratively generate a knowledge base that promotes better health decisions and better care. The exposome research community goes beyond European patients, researchers and health professional including also healthy citizens (maternity is not a disease).

Access to data and cloud services will be underpinned by a shared federated authentication and authorization infrastructure (Life Science AAI). A common provenance information model will enable meaningful reuse of data, as it allows to trace history of the data from the biological material acquisition, through data generation, to data processing and meet requirements from GDPR and the Nagoya protocol on access and benefit sharing.

In order to minimize risks of semantic differences, we will use available de facto standards and have employed key expertise in ontology in the project. The data upload engine will detect if there are semantic differences. Also, the metadata engine for public search will be extracting data for public display of aggregated information and semantic differences that may hinder later retrievals will therefore be detected at an early stage.

HEAP's Information Commons will enable efficient retrieval by its catalogues if appropriate permissions from data controllers and ethical review boards are in place. In the main HEAP data catalogue, all HEAP trusted data repositories and their metadata will be listed. This "catalogue of catalogues" acts as entry point to the HEAP's Information Commons and will (a) provide a federated search functionality for all of the HEAP trusted data repositories and (b) information on how to access a specific catalogue and data entries in the HEAP's Information Commons.

All catalogues in the HEAP's Information Commons will have a persistent identifier (PID) and will be listed in registries of repositories, e.g. in EOSC-Life registries. Each catalogue and trusted data repository in HEAP's Information Commons will support

- Explicit roles and responsibilities and an explicit data deletion policy
- Different access policies for different versions of the data
- Technical support for predefined file formats
- Reuse of community standards and ontologies from public registries
- Use of PIDs as the manifestation of a data policy

- Explicit data policies (like versioning and dynamic data) and PID policies in human and machine interoperable way
- Documentation of interfaces and APIs

2.1 Making data findable, including provisions for metadata

2.1.1 Outline the discoverability of data (metadata provision)

Catalogues in the HEAP's Information Commons (IC) will fulfil the following technical requirements:

- All catalogues will follow the same PID policy
- Catalogues will provide metadata in different formats, which can be harvested by different search engines.
- Metadata will include human-friendly presentations and visualizations
- Catalogues will provide metadata at the level of databases, files, variables, attributes, where the granularity will be decided according to demands of data users (fit for purpose).
- Catalogues will provide landing pages and will be machine-interpretable or implement content negotiation
- Catalogues will provide machine-readable and interpretable metadata about repository / cohort itself and expose a Meta Data Model in machine-readable form
- Catalogues will provide a machine-readable license
- Catalogues will provide a search interface and be linked to aggregating services

The assignment and management of persistent identifiers to the data will be assessed in the course of the project and will be described in the project DMPs. It is recommended to use Uniform Resource Identifier (URI) to facilitate links between different data.

All deliverables will be listed on the HEAP website and the ways by which HEAP outputs can be accessed will be communicated via social media and other suitable channels to increase visibility of HEAP work. For public deliverables, a link will be available between the HEAP website and the appropriate open repositories where the data is submitted.

Each HEAP cohort will use metadata standards or metadata models appropriate to their own data, which will be described in the individual project SOPs. The DMP team will provide an inventory of metadata standards or metadata models related to HEAP data.

2.1.2 Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

All sample information will be harmonized according to the MIABIS standard. The Minimum Information About Biobank data Sharing (MIABIS) aims to standardize data elements used to describe biobanks, research on samples and associated data. The MIABIS Community Standards work on several granularity levels, with the aim to support interoperability

between biobanks sharing their data. General attributes to describe biobanks, sample collections and studies at an aggregated/metadata level are defined in MIABIS Core 2.0. MIABIS Core 2.0 represents the minimum information required to initiate collaborations between biobanks and to enable the exchange of biological samples and data. The aim is to facilitate the reuse of bio-resources and associated data by harmonizing biobanking and biomedical research. The attributes are defined in accordance with epidemiological literature and terminology.

MIABIS Core 2.0 consists of three main components: "Biobank", "Samples Collection" and "Study". Further components standardize the description of Samples, Sample Donors, SOPs and Image collections. MIABIS Core is currently being updated to version 3.0. The update will include addition of a 'network' -component, and overall additional updates will set up a liaison to the BBMRI-ERIC CS-IT MIABIS working group.

In the metagenomics data analysis we will collect metadata for the study and collection context and information about the sampling process, including pre-analytic quality parameter and sample storage information, description of the sequencing process, analysis metadata and information about how results are archived, stored and archived. We will use the following metadata standards to make the collected information interoperable:

- Minimum information about a Genome Sequence (MIGS)
- Minimum information about Metagenomics Sequence (MIMS)
- Minimum information about a marker gene sequence (MIMARKS)
- Minimum information about any (x) sequence (MIxS)
- Controlled vocabularies for library- source, strategy, selection, layout and quality scores
- Information how the datasets are analysed, described e.g. with the Common Workflow Language (CWL)
- For archiving, DOI and handles according to the EGA/ENA metadata standards

When collecting and describing data in the consumer cohort, we will make use of data from commercial digital receipt providers (such as start-up company Storebox) and if possible, loyalty programmes (such as COOP, pending approval). The collected metadata will reflect the data provided from retailers. By the end of the project, it is expected to have metadata following the following formats where applicable

- EFSA food classification system FoodEx2
- Global Data Synchronisation Network (GDSN)
- Global Product Classification (GPC)

2.1.3 Outline the identifiability of data and refer to standard identification mechanisms

Personal data included in the HEAP information commons may be anonymous as well as indirectly identifiable. This will depend both on the specifics of data management in various sub-projects as well as on the granularity of data and pseudonymisation and anonymisation

measures in the data flows. All data will be stripped of direct identifiers, including national civic registration numbers, and be assigned unique research code numbers. Specific measures and policies will be developed further in the course of the project in the context of deliverable D2.1 (legal mapping).

In Sweden, the register-based research is protected by law, and all projects should be approved by the Ethical Review Authority (former Ethical Review Boards). Register-based research is commonly not deviated from clinical routine nor does it involve direct contact with the study participants, and therefore, informed consent could be waived. Registries link the data based on the Swedish PIN, which was established in 1947 and serves as a unique identifier in Swedish healthcare system as well as many other areas of the society. The PIN enables linkages between population registers and large-scale medical registers allowing for virtually 100% complete coverage in Swedish healthcare system. Once the linkage is done, registries substitute the PIN by a corresponding sequential running number before data is delivered for research purposes. Data usage is under strict regulation and delivered data cannot be used for purposes other than those approved by the Ethical Review Authority. Re-use of data is possible if the analysis is within the field covered by approval, or an amendment of ethical approval for that specific research purpose is granted. In the case that samples need to be retrieved from different biobanks, and PIN is needed, a single database administrator working in a highly integrity-assured environment at Karolinska Institutet will obtain an encrypted file with the personal identification numbers of patients and will contact the Pathology archives to request the specimens from the identified patients. Before the specimens are sent for analysis, the administrator will assign code running numbers to the specimens. All analyses will be performed on coded specimens and no research analysis will be linked back to any individually identifiable patient.

2.1.4 Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

CSC will provide the service public and permanent identifiers as described in the reference architecture. The selection of the actual identifier system will be established as part of the implementation and will be aligned with “sister projects” in the Human Exposome Network and with EOSC-Life.

To maximize interoperability the following persistent and unique identifiers and controlled vocabularies will be considered:

DOI:	https://en.wikipedia.org/wiki/Digital_object_identifier Digital Object Identifier System
Handles:	https://en.wikipedia.org/wiki/Handle_System http://www.handle.net
XRI/IRI/URI:	Extensible Resource Identifier Internationalized Resource Identifier Uniform Resource Identifier

Identifiers:	Identifiers.org
Schema:	https://schema.org/docs/schemas.html
BioSchemas:	https://bioschemas.org https://bioschemas.org/groups/Samples/
BioPortal:	http://bioportal.bioontology.org/ontologies
FHIR:	https://www.hl7.org/fhir/diagnostics-module.html https://www.hl7.org/fhir/terminologies-systems.html

The naming convention for all HEAP deliverables will follow the HEAP Grant Agreement which is in the format: "D.x.y Name of deliverables".

2.1.5 Outline the approach towards search keyword

Metadata elements need to be aligned across all partners in order to simplify the inquiries using keywords. Consequently, the metadata elements must contain the term "HEAP", to facilitate finding of HEAP data. The variety of the appropriate repository for the HEAP deliverables and data has to allocate a filtering system based on the metadata elements, e.g. SPARQL system, which is a standardised language for querying RDF data, also capable to search for linked data.

2.1.6 Outline the approach for clear versioning

Data originated from sequencing analysis will specify the software used, their version, and the parameters chosen for the analysis. Same applies for reference databases used for alignment or mapping of reads that will state the date of accession and downloading. Alerts will be set, and scripts will be written to assure downloading updates and performing corresponding changes in documentation to assure running the last updated versions.

2.2 Making data openly accessible

Standardly, European Commission services and European Agencies, EU National Bodies, HEAP consortium, and the general public will be able to have free access to data and metadata of HEAP.

Within the framework of research and innovation, 'scientific information' includes the meaning of peer-reviewed scientific research articles (published in scholarly journals), or research data (data underlying publications, curated data and/or raw data). Open access to scientific publications guarantees free online access for any user. As stated in the Grant Agreement, the costs of open access publishing are eligible. Open access to research data ensures the right to access and reuse digital research data under the terms and conditions established in the Grant Agreement.

2.2.1 *Specify which data will be made openly available? If some data is kept closed provide rationale for doing so*

All catalogues describing data records in the HEAP Information Commons will be openly available and provide metadata information about data access rules. The basic premise is that data will be as open as possible and as closed as necessary, especially considering the data protection issues when opening data without restrictions. We will build on the FAIR data stewardship guidelines and European Open Science Cloud recommendations in this regard.

Sequencing data, code relating to bioinformatics pipelines and bioinformatics analysis will initially be restricted to HEAP project partners, and later released to the public. All data that may be linked to identifiable persons will be highly restricted. The rationale for keeping data closed might include:

- Open access is incompatible with rules on protecting personal data: protection of the personal right needs to be ascertained either by avoiding open access to sensitive and personal data, or by anonymizing the data if relevant and feasible.
- Open access is incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited
- Open access is incompatible with the need for confidentiality in connection with data from external owners/providers: Because of the co-funding setup of HEAP, partners might use data collected or generated by or with co-funders. If relevant for other research partners, agreements with co-funders will be discussed to make those data accessible to other HEAP partners, while respecting compliance with European and national ethic-legal framework.
- Open access is incompatible with the need for confidentiality in connection with security issues
- Open access would mean that the project's main aim might not be achieved.

Data, including deliverables, produced in the course of the project should be made openly available as the default, while respecting compliance with European and national ethic-legal framework on personal data protection.

In order to help partners in their decision to use open access, restricted access or keeping data closed, WP 2 and WP 7 will develop access rules and a clear governance structure how data and under what conditions data will be available for research (tasks 2.2 and 2.3). So, the access to research data will be data specific. The final decision to select a specific type of access (open, restricted or close) will be under the responsibility of the data custodians which collected, processed or generated the data, yet must still be explainable in FAIR terms. The rationale to keep data restricted or closed will be described in the HEAP deliverables.

Personal data included in the HEAP Information Commons (IC) may be anonymous as well as indirectly identifiable. This will depend both on the specifics of data management in various sub-projects as well as on the granularity of data and pseudonymisation and anonymisation measures in the data flows. All data will be stripped of direct identifiers, including national

civic registration numbers, and be assigned unique research code numbers. Specific measures and policies will be developed further in the course of the project in the context of deliverable D2.1 (legal mapping).

All data from women's history and their sequencing data will be open access with the exception of the PIN numbers in order to restrict identification of people. Running code numbers will substitute the PIN numbers.

2.2.2 Specify how the data will be made available

All research data will be made available through the HEAP Information Commons. All technical details will be specified in the HEAP reference architecture.

HEAP Information Commons (the knowledge as a sum of data, metadata or features etc.) is represented in the Metadata Warehouse and is available for sharing and reuse along with associated metadata, thus making it also findable - ultimately enabling the platform to be FAIR compliant.

Regarding the users of the Human Exposome Assessment Platform we have identified three primary users:

- Researcher
- Data Custodian
- Data Access Controller/Committee

A Researcher would be the primary type of user that makes use of the HEAP Metadata Warehouse to discover data and its features and utilise the HEAP Knowledge Engine in order to run analysis pipeline(s) for e.g. large scale assembly inference from multiple data sources.

A Researcher can request the Data Access Controller for access to specific data sources. The Data Custodian can make use of the HEAP Submission Engine and ETL jobs to provide the data or a subset of it depending on the request of the Researcher. The Data Access Controller will make use of the HEAP Entitlements Management System to grant permissions to access the data. Once permission has been granted the Researcher will have access to the data sources contained in the Information Commons layer via the HEAP Knowledge Engine.

The Data Custodian and Data Access Controller can be represented by the same person(s), but this is not a requirement.

The HEAP Warehouse acts as an online catalogue that makes metadata about the data stored in Information Commons and Knowledge Engine Data and Feature Store available to stakeholders and the general public. The HEAP Metadata Warehouse will leverage repositories and metadata specifications in order to index data resources in a catalogue and make it available for access and use by researchers or other interested parties.

Considering that such a catalogue will contain metadata, a connection to the HEAP Entitlements Management System might be appropriate, either in the form of providing information on how to access data sets or linking to the HEAP Entitlements Management System in order to obtain access.

The HEAP platform guarantees access of the deliverables. Even though a small portion of deliverables will be under confidentiality, the majority will be open and accessible publicly. Furthermore, public deliverables will be connected to the open repository filed in a machine-readable format.

2.2.3 Specify what methods or software tools are needed to access the data?

In most cases, only standard software, e.g. web browsers, pdf-file readers, and text readers, will be required. The exception here is certain data, i.e. genomic data, which requires specialized tools and languages accessing the data.

2.2.4 Specify where the data and associated metadata, documentation and code are deposited

All data and associated metadata will be documented and deposited at the HEAP Information Commons. The bioinformatic pipelines, code and supporting databases will be deposited in the KI cluster which has implemented the Hopsworks platform, which will be fully integrated into the HEAP Information Commons.

The HEAP Information Commons will provide an appropriate data submission workflow and data stewardship tools. All data repositories in the HEAP Information Commons will:

- Ensure long-term persistence and preservation of datasets
- Provide expert curation
- Provide stable identifiers for submitted datasets
- Allow public access to data without unnecessary restrictions

2.2.5 Specify how access will be provided in case there are restrictions

By default, data generated with HEAP co-fund and accompanying metadata are directly accessible for use within HEAP. For sensitive data, the data custodian shall finally agree to access and/or the transfer of the data at high level of granularity to a HEAP defined repository, using appropriate measures to anonymise or pseudonymise data. Tasks 2.2 and 2.3 will further define the conditions under which data will be made available.

HEAP deliverables and data can either be public or confidential. Some results might be restricted in their use. Sensitive and personal data can be made accessible only following the GDPR requirements. The aim is to reach an appropriate level of GDPR compliance, amongst others by:

- Relying on the EOSC Life Science AAI and security protocols for data sharing.

- Applying a strict policy in granting and revoking access to the data.
- Logging of user identity during data access, download, and upload, including version control. As several repositories will be used to store data, the policy on how to grant access to restricted results will be provided by the catalogue metadata.

Prior to generation of the data, the data custodian shall confirm ethico-legal compliance of the study in which new data are generated. For existing data, not generated with HEAP co-funding, the data custodian specifies the level of granularity that data will be stored and/or transferred: anonymised or coded single measurement data; pseudonymised or coded single measurement data; or aggregated data. The data custodian indicates for each level of granularity whether the data are directly accessible for use within the HEAP. In case the data custodian indicates that the data are not directly accessible for use within HEAP by default, the data custodian will be asked approval when consortium members request access to the data to meet the goals of a particular objective.

2.3 Making data interoperable

HEAP partners are involved in major European standardisation and interoperability initiatives and will align with, and contribute towards efforts in e.g. EOSC-Life, EJP-RD, GA4GH, GO FAIR, and RDA. In order to achieve the highest possible level of interoperability, we will adopt European and international metadata and data standards. For interoperability topics related to biobanking, we will follow the recommendation of BBMRI-ERIC and for all bioinformatic topics we will rely on the ELIXIRs knowledge base.

2.3.1 Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

We will base the HEAP Metadata Warehouse and the HEAP Knowledge Engine (Hopsworks) fully on standard vocabulary as described in section 2.1.2 and align the selection of controlled vocabularies across all projects in the Human Exposome Network.

2.3.2 Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

The main standards for metadata creation are described in section 2.1.2 of the DMP. In addition, we will enable secondary reuse of clinical and eHealth data in HEAP by the integration of observational data and harmonization of with a common metadata model. The HEAP metadata warehouse can be used for cohort identification in clinical trials, retrospective clinical studies or public health studies. In order to integrate data from different sites we will support a Clinical Common Data Model (CDM), which is used to provide a standardized and logically unified way to represent EHR data from distributed research networks. CDMs ensure that clinical research methods are consistent and reusable across the networks for producing meaningful, comparable and reproducible results according the FAIR data principles. Several CDMs exists, e.g. PCORnet, the I2B2 CDM and OMOP, in the following described shortly:

The **PCORnet** common data model (PCORnet CDM) is supported by all networks in the Patient Centered Outcomes Research Institute, and thus has a wide base of existing support. Over 80 institutions have already transformed their data into this model. It was derived from the Mini-Sentinel data model, which has increasing uptake in claims data analysis. Technically, PCORnet CDM (v3.1) is a traditional relational database design, in which each of fifteen tables correspond to a clinical domain (e.g., diagnoses, labs, medications, etc.). The tables have many columns including both the table key (patient identifier, encounter identifier, etc.) and additional details (e.g., medication frequency). New releases of the data model have added new clinical elements or format–for example, new domains (e.g., lab values) and changes in data representation (e.g., smoking status).

Informatics for integrating biology in the bedside (**i2b2**) was first developed by the National Institutes of Health (NIH) grant and has over 200 sites world-wide, and it is used in several large-scale networks. Technically, i2b2 uses a star-schema format, pioneered by General Mills in the 1970s and widely used in retail data warehouses. The i2b2 star-schema uses one large "fact" table containing individual observations. This is a narrow table with many rows per patient encounter. Ontology tables (hierarchical arrangements of concepts) provide a window into the data; these are often developed by local implementers. Consequently, the data model is only modified when core features are added to the platform.

The Observational Medical Outcomes Partnership (**OMOP**) CDM was developed to be a shared analytics model from the beginning, and it has been adopted by the Observational Health Data Sciences and Informatics (OHDSI) Consortium Technically, OMOP is a hybrid model that provides domain tables in the vein of PCORnet, as well as a "fact" table containing individual atomic observations similar to i2b2. The OMOP schema is significantly more complicated than PCORnet, and some domain tables are derived values for specific analytical purposes. Unlike PCORnet (but similar to i2b2's ontology system) OMOP provides metadata tables providing information on terminology and concept relationships. WP 7 will establish a liaison to OHDSI Europe with its coordinating centre at the Erasmus University Medical Center in Rotterdam and to the German medical informatics platform MIRACOLIX, which currently evaluates an OHDSI based I2B2/TransMart solution. We will, thus, actively contribute to the further extension of the OMOP-CDM and provenance tool development for exposome related data.

2.4 Increase data re-use (through clarifying licenses)

2.4.1 Specify how the data will be licensed to permit the widest reuse possible

The open repositories provide the data available for reuse. Additionally, the project HEAP will establish software and specific tools, disseminated as open source software assuring their widest reuse. For access policies concerning the re-use of personal data, see to section 2.2.

2.4.2 *Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed*

The HEAP partners are responsible coming to a clear decision on an embargo for the research data. Scientific research articles should be available publicly with the latest publication in an Open Access Journal, or within half a year after its launch. Concerning research data, free access should be granted by default when the related issue is available in open access. See also in section 2.2.

2.4.3 *Specify whether the data produced and/or used in the project is useable by third parties, in particular, after the end of the project? If the re-use of some data is restricted, explain why*

Public data will be guaranteed from open repositories, and therefore reusable by third parties, even after the end of the project. Concerning confidential data, access to personal data will be consistent with GDPR, whereas data relating to intellectual property must be of discussion between relevant partners, and an agreement will be achieved following the governance procedures developed in the course of the project (task 2.3, see also section 2.2), in accordance with European and national legislation and ethics requirements

2.4.4 *Specify the length of time for which the data will remain re-usable*

Even after the end of the project all files will be stored within the HEAP Information Commons (IC) to meet the requirements of good scientific practice. A detailed approach concerning long term sustainability of data and metadata will be developed by WP7.

Concerning secondary data which is stored on other repositories, researchers, institutions, journals and data repositories share a common responsibility to warrant long-term data preservation. At least five years after publication, the datasets must be preserved by the partners on their own institutional servers. If, within this timeframe, the repository to which the data were originally submitted disappears or experiences data loss, the partners will be held responsible and need to upload the data to another repository and provide a public correction or update to the original persistent identifier, if required.

2.4.5 *Describe data quality assurance processes*

The HEAP consortium commits to provide good quality data, which is assured through various approaches: Firstly, each and every HEAP partner applies to existing data quality assurance processes, described in their quality manual; Secondly, similarly to publications, which will be distributed in peer-reviewed journals, research data will be applicable on repositories offering a curation system suitable to the data.

Furthermore, developing guidance documents to evaluate data quality is part of some project objectives. These guidance documents will be examined and optimised during these specific projects and will be verified using suitable approaches.

Data providers from the projects described above (Cervical screening cohort, Maternity cohort, HPV vaccination cohort, Consumer cohort, Wearable data collection study) are key

players for the development of the HEAP's Information Commons. To ensure an efficient and agile co-development, "Bring your own data – BYOD" workshops will be jointly organized by WP7 and WP11 during the first 2 years for the project, and targeting in priority partners from WP 3, 4 and 5. Hands-on training modalities will be privileged for these workshops, on the model of hackathons, which will allow providers to examine their data in view of the HEAP DMP, as well as improve and evaluate the interoperability and FAIRness of their data sets.

BYOD workshops will be planned as (virtual) 3-4 day events, combining (1) relatively short live sessions through web conferencing tools, supported by specific online software for interaction and brainstorming (e.g. Q&A, polls, mind mapping etc.), with (2) asynchronous activities (e.g. working on own data assignments), and (3) communication channels (e.g. slack) for support and longer term follow-up. BYOD's participants will have access to a dedicated online space on the HEAP learning platform, which will provide the frame and the tools for the workshop.

Once the model of the workshop established through the first events, further BYOD workshops will also include future users of the HEAP platform, to contribute to dissemination and sustainability activities.

3 Allocation of resources

Tasks related to making HEAP data FAIR are in WP 2 (Ethics, data protection and governance) and WP 7 (Data interoperability and data sharing) and for the implementation in WP 10 (Secure Infrastructure for big data). The responsibilities are clearly assigned in the WP description and all tasks will be managed by the dedicated WP leaders. All costs for FAIR data management are covered by WP2, WP7 and WP10 resources.

Long term preservation will be done by resources provided by the participating Biobanks, research institution and by the CSC. Metadata catalogues will be published to the catalogues of ESFRI RI, e.g. BBMRI-ERIC for sample catalogues and ELIXIR for bioinformatics data collections.

4 Data security

WP10 will develop a secure platform suitable for managing sensitive data for the consortium. The secure IaaS platform includes secure data storage and secure cloud environment, as well as a dataset authorization management system. This forms the core of the Information Commons for the HEAP informatics platform. Generally accepted norms and certification schemes concerning personal data security will be implemented in accordance with the requirements of the GDPR (article 42).

HEAP's Information Commons can be deployed across locations in a distributed manner, that allows streaming of remote data on secure cloud on-demand for processing. This setup allows the data owners to store data in their local repository while making it available for research in a controlled environment through the Information Commons.

Access to the secure data platform could be provided through

- dedicated network connection (MPLS technology) to ePouta cloud. The connection is already implemented between CSC and KI; and is straightforward to open for organizations connected with National research and education networks,
- remote desktop connection to ePouta cloud that works through a web browser and standard encrypted internet connections.

The requirements for the HEAP sensitive data platform are derived from research use cases, especially those utilising register data in WP3. The secure data platform will adopt the recommendations on legal and ethical items from WP2. In addition, this WP collaborates with WP6 to develop and test integration of the IaaS and PaaS platforms in the consortium, to enable seamless data use for different user scenarios of register research and big data analysis.

5 Ethical aspects

Exposomics research is another form of big data analysis and poses new challenges for confidentiality, consent and public trust. These relate to the whole chain from the uptake of data in big data analytics to the intended outcomes. Most ethical and legal research has been done on the uptake process, leading to various proposals for data subjects' control which will be further explored in the context of HEAP. Data subjects' control is strengthened by EU General Data Protection Regulation (2016/679, GDPR) which leaves certain exemptions for research and public health according to EU or member states law. Outcomes of data research can also have adverse consequences for those whose data have not been used at all in the analytics, but rightfully or wrongly are subsumed under the same group whose data have been used. This requires ethical assessment of the intended output as well. HEAP will address both aspects (input and output) through an appropriate governance system as the linking pin. Stakeholder involvement from patient organisations and the public at large will play a pivotal role here.

From a regulatory point of view, the set-up where the data holder is the data controller and the HEAP platform is the data processor helps to maintain the former in control according to their legal basis and agreements with the data subjects to submit data for further research. Therefore, the platform will not create one large European 'data lake' but will combine the advantages of a federated and a central approach. Big data and purpose limitation seem contradictory. Yet the HEAP platform will analyse data only based on specific research questions with specific permissions. This will, together with the security measures (as

described in WP 9), also significantly reduce the risks as e.g. described in an earlier ENISA report on data-related threats.

HEAP places much emphasis on veracity. This is an ethical requirement in the context of health research where data must be verifiable given the possible huge impact of this research which might lead to new treatments or preventive strategies for large groups of the population. Veracity means that fully anonymous data cannot always be used in the Information Commons. Apart from the granularity needed for the methodologically correct analyses (which would make them indirectly identifiable), they must often be two-way pseudonymised for data source control and individual feedback of actionable findings to the data subjects. Within these methodological constraints' privacy by design and by default is used and the highest norms regarding data security will be implemented.

The HEAP platform must lead to outcomes which will further an equitable healthcare system and public health. Involvement of all relevant stakeholders is essential to that and will be incorporated in the governance structure. It should also be trustworthy to society at large and applicable to other data sources from other European member states other than the current partners in HEAP. This requires a governance framework which enables an interaction with civic society. Finally, transparency and accountability - essential elements for trust - about data processing through HEAP is a prerequisite that will be applied - ranging from the data sources, the HEAP platform, via the governance to the output.

The table below provides a short overview of the main ethical and legal issues related to data sharing within and through HEAP, focusing on ethics and data protection. The table provides some basic clarification of terms, as well as an assessment of current status: whether information about these items is provided elsewhere in this DMP, can already be provided by HEAP partners, or whether characteristics still need to be determined (TBD) in the course of HEAP.

Contributing data sources ('research projects')	*Clarification*	*Current status*
Is data contributed by the data sources personal data or anonymous data?	That will depend, see section 2	TBD, partly provided in cohort data summary
Have participants provided consent for data sharing and reuse for HEAP research? If not, what is the (national) legal ground for data sharing and reuse?	This will also depend on the cohort used	TBD, partly provided in cohort data summary

What is the scope of research purposes to which participants have consented or permitted by the legal ground?	This is one of the issues which will be determined by each data custodian according to its applicable procedures before data can actually be used for research	TBD, partly provided in cohort data summary
Is transfer of data to non-EU countries allowed? If so, how?		TBD, partly provided in cohort data summary
Are there other restrictions and safeguards which need to be met, such as?	Yes, which can differ for each cohort/data custodian, see task 2.1	TBD, partly provided in cohort data summary

HEAP infrastructure (repositories, catalogue, computing platforms)	*Clarification*	*Current status*
Data transfer to and within HEAP infrastructure		
Data federation between contributing sources and infrastructure	One-time uploads from contributing data sources (‘data lake’) or case-by-case temporary transfers (a more federated system)?	TBD
‘Data protection by design’ specifications	Pseudonymisation, data minimisation, procedures for linking individual-level data between cohorts	TBD
Storage, archiving, computing		
Legal responsibility/custody for storage, archiving, computing, analysis	E.g. joint controllership, representation	TBD
Location of and jurisdiction over data	Each data custodian will remain the controller of the data. Hence, the legal regime does not change because of transfer of data.	

Involvement of data processors (in the legal data protection sense)	Yes, amongst others CSC	TBD
Security measures and certificates		TBD
Catalogue		
Does the catalogue only contain metadata?		TBD
Measures to prevent reidentification through metadata	Whether such measures are needed; what measures have been taken (e.g. k-anonymity)	TBD

Access and use	*Clarification*	*Current status*
Access		
Access policy, terms and procedure	(FAIR) access policy	TBD in task 2.2 and 2.3
Access governance	Organizing decision-making on access and use. E.g. HEAP Data Access Committee to be installed; are actual use (first case) or transfers (second case) always subject to data sources' individual approval, only subject to approval by HEAP, or a combination (e.g. approval through HEAP with possibility for data source to opt-out)	TBD in task 2.2 and 2.3
Use		

Roles, monitoring, logging	Assuming that after approval, only access to the data will be given, no transfer of the data (data can be analysed on the commons Platform).	TBD
Additional personal data for joint analysis by users		TBD
Data protection for software and code: vetting of software and code for data protection issues	The platform will allow users to install their own code and statistical programs	TBD
Licensing terms on data and results generated through the platform		TBD in chapter 2.4 of DMP

6 Other issues

The following provides an action plan, which presents important topics requiring progress and/or update in future versions of the DMP.

DMP component	Issues to be addressed	Actions
1. Data summary	1. Explain the relation to the objectives of the project 2. Specify the types and formats of data generated/collected	Update the detailed data type in deliverables of the research projects.
	1. Specify if existing data is being re-used (if any) 2. Specify the origin of the data 3. State the expected size of the data (if known) 4. Outline the data utility: to whom will it be useful	List of data collected/generated as specification for the HEAP submission engine.
2. FAIR Data 2.1. Making data findable, including provisions for metadata	1. Outline the discoverability of data (metadata provision) 2. Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? 3. Outline naming conventions used 4. Outline the approach towards search keyword 5. Outline the approach for clear versioning 6. Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how	Provide a detailed metadata catalogue and update the inventory of relevant metadata standards and models in the HEAP Metadata Warehouse Provide training for the research projects, organise BOYD workshops.
2.2 Making data openly accessible	1. Specify which data will be made openly available? If some data is kept closed provide rationale for doing so 2. Specify how the data will be made available 3. Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? 4. Specify where the data and associated metadata, documentation and code are deposited 5. Specify how access will be provided in case there are any restrictions	Developing a decision tree to choose between data open access, restricted access to data or keeping data closed List of repositories with filtering system based on topics
2.3. Making data interoperable	1. Assess the interoperability of your data. Specify what data and metadata vocabularies,	Liaise with appropriate support to ensure sustainability?

	standards or methodologies you will follow to facilitate interoperability.	List of metadata standards useful to HEAP
	2. Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter- disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?	
2.4. Increase data re-use (through clarifying licenses)	1. Specify how the data will be licensed to permit the widest reuse possible 2. Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed 3. Specify whether the data produced and/or used in the project is useable by third parties, in particular, after the end of the project? If the re-use of some data is restricted, explain why 4. Describe data quality assurance processes 5. Specify the length of time for which the data will remain re-usable	Set up a curation system, supported by WP2 (governance model)
3. Allocation of resources	1. Estimate the costs for making your data FAIR. Describe how you intend to cover these costs 2. Clearly identify responsibilities for data management in your project 3. Describe costs and potential value of long-term preservation	Defined in the HEAP grant agreement
4. Data security	Address data recovery as well as secure storage and transfer of sensitive data	Will be implemented in WP10
5. Ethical aspects	To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former	Will be developed by WP2 in the HEAP governance model
6. Other	Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)	

7 References

Azab, H. Meling, E. Hovig and A. Pursula, Stroll Filesystem Client-Server for Seamless Job Management in Sensitive Data Cloud Federation, 2018 IEEE International Conference on Big Data (2018).

Bischof J, Harrison T, Paczian T, Glass E, Wilke A, Meyer F. Metazen - metadata capture for metagenomes. Stand Genomic Sci. 2014;9:18. Published 2014 Dec 8. doi:10.1186/1944-3277-9-18

Bowers RM, Kyrpides NC, Stepanauskas R, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea [published correction appears in Nat Biotechnol. 2018 Feb 6;36(2):196] [published correction appears in Nat Biotechnol. 2018 Jul 6;36(7):660]. Nat Biotechnol. 2017;35(8):725-731. doi:10.1038/nbt.3893

Genomic Standards Consortium (GSC) [Internet]. Available from: http://gensc.org

Kottmann R, Gray T, Murphy S, et al. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS. 2008;12(2):115121. doi:10.1089/omi.2008.0A10

Niina Eklund, Ny Haingo Andrianarisoa, Esther van Enckevort, Gabriele Anton, Annelies Debucquoy, Heimo Müller, Linda Zaharenko, Cäcilia Engels, Lars Ebert, Michael Neumann, Joachim Geeraert, Veronique T'Joen, Hans Demski, Élodie Caboux, Rumyana Proynova, Barbara Parodi, Sebastian Mate, Erik van Iperen, Roxana Merino-Martinez, Philip R. Quinlan, Petr Holub, and Kaisa Silander.

Holub, Petr, Florian Kohlmayer, Fabian Prasser, Michaela Th Mayrhofer, Irene Schlünder, Gillian M. Martin, Sara Casati et al. "Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health." Biopreservation and biobanking (2018).

Lappalainen I, Almeida-King J, KumanduriV, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, PaschallJ, Flicek P. The European Genome-phenome Archive of human data consented for biomedical researchNat Genet Volume 47 (2015) p.692-695 DOI: 10.1038/ng.3312

Ondřej Májek, Ahti Anttila, Marc Arbyn, Evert-Ben van Veen, Birgit Engesæter, Stefan Lönnberg; The legal framework for European cervical cancer screening programmes (European Journal of Public Health, Volume 29, Issue 2, April 2019, Pages 345–350, https://doi.org/10.1093/eurpub/cky200)

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the

automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;19;9:386.

Tsesmetzis N, Yilmaz P, Marks PC, Kyrpides NC, Head IM, Lomans BP. MIxS-HCR: a MIxS extension defining a minimal information standard for sequence data from environments pertaining to hydrocarbon resources. Stand Genomic Sci. 2016;11:78. Published 2016 Oct 12. doi:10.1186/s40793-016-0203-5

E.B. van Veen: Observational health research in Europe: understanding the GDPR and underlying debate, Eur J Cancer, 2018, DOI: https://doi.org/10.1016/j.ejca.2018.09.032

E.B. van Veen, Health care data for health research, The Hague: MedLawconsult 2011

E.B. van Veen: Observational health research in Europe: understanding the GDPR and underlying debate, Eur J Cancer, 2018, DOI: https://doi.org/10.1016/j.ejca.2018.09.032

Yilmaz P, Kottmann R, Field D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415-420. doi:10.1038/nbt.1823

Yilmaz, Pelin et al. “Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.” Nature biotechnology vol. 29,5 (2011): 415-20. doi:10.1038/nbt.1823

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B. W., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature biotechnology, 29(5), 415–420. https://doi.org/10.1038/nbt.1823