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 10
Secure Infrastructure for big data

Work package Leader: Stefan Negru
Organization: CSC IT Center For Science
E-Mail: stefan.negru@csc.fi

Project Deliverable

D10.1: Reference Architecture

Deliverable due date: 2020-06-30
Deliverable due month: M6

Document history

Version	Date	Changes	By	Reviewed
0.1	2020-03-19	First draft	Stefan Negru	Juha Törnroos
0.2	2020-04-21	HEAP requirements	Stefan Negru, Teemu Kataja	Juha Törnroos
0.3	2020-04-27	Alignment with Data Management Plan	Heimo Müller Stefan Negru	Roxana Merino
0.4	2020-04-30	WP 6 input	Alex Ormenisan	Stefan Negru
0.5	2020-05-15	Input from all HEAP WPs - RFC part 1	Stefan Negru, Juha Törnroos	Martin Boeckhout, Evert-Neb van Veen, Roxana Merino
0.6	2020-05-28	Input from all HEAP WPs - RFC part 2	Stefan Negru	Juha Törnroos, Roxana Merino
0.7	2020-06-16	Add summary, data flow diagram, Hopsworks metadata layer and other details	Stefan Negru, Alex Ormenisan	Roxana Merino
0.8	2020-06-22	Input from all of HEAP WPs	Stefan Negru	WP leads
1.0	202-06-23	Final	Stefan Negru, Juha Törnroos	Roxana Merino Joakim Dillner

Executive Summary

This document provides guidance for the development and use of the Reference Architecture within the Human Exposome Assessment Platform (HEAP). This document describes the technical architecture, data and metadata flow and means for accessing data in the platform. It informs of the tools required, the nature of work to be carried out and the required set of components or contributors to be involved in developing the platform.

The architecture presented in this document and its implementation are coordinated within the Secure Infrastructure for Big Data Work Package (WP10) and its aim is to develop an IaaS (Infrastructure as a Service) platform for the HEAP Information Commons with the main features storing and managing sensitive data. The platform will provide secure data storage in a cloud environment and allow streaming of remote data for processing through a secure data access mechanism.

Table of Contents

1 Introduction

1.1 Purpose

This document provides guidance for the development and use of the Reference Architecture within the Human Exposome Assessment Platform (HEAP). This document describes the technical architecture, data flow and associated metadata and means for accessing data in the platform. It informs of the tools required, the nature of work to be carried out and the required set of components or contributors to be involved in developing the platform.

This document does not address implementation details including in depth description of technical solutions or security standards. The interfaces between components are not specified in this document as this depends on the implementation of the Reference Architecture.

1.2 Overview

HEAP aims to provide a system for collecting, managing, sharing and ultimately analysing exposome and exposome related data on a large scale, as defined in the project proposal. In order to achieve these goals, the platform will enable collaborative exposome research by utilising data from longitudinally followed population-based cohorts in advanced exposome measurement technologies, all aided by distributed high-performance computational resources.

Key aspects facilitated by the platform are:

- providing distributed high-performance computational resources across institutions;
- software platform that integrates the computing resources and enables secure management of large population-based cohorts;
- analysis pipelines implemented in the platform for processing of harmonised and interoperable heterogenous data.

1.3 Key Requirements

The implementation of HEAP reference architecture should consider the following requirements and limitations. See RFC2119[1] for the definition of terms **MUST**, **SHOULD**, and **MAY**.

The implementation:

- **MUST** enable controlled access to datasets;
- **MUST** enable scientific analysis workflows as bioinformatics pipelines;
- **SHOULD** support working with encrypted data;
- **SHOULD** facilitate distributed analysis and processing of omics data;
- **SHOULD** provide means to make results publicly findable and provide means to access them
- **MAY** reuse existing standards.

[1] https://www.ietf.org/rfc/rfc2119.txt

1.4 Definitions, Acronyms and Abbreviations

The table below lists definitions for acronyms and abbreviations relevant for this document.

Term	Definition/Reference
HEAP	Human Exposome Assessment Platform
Hopsworks	Software developed by Logical Clocks https://www.logicalclocks.com/hopsworks
GA4GH	Global Alliance for Genomics & Health - https://www.ga4gh.org/
AAI	Authentication and Authorisation Infrastructure
ELIXIR	Elixir Europe https://elixir-europe.org/
EOSC	European Open Science Cloud https://www.eosc-portal.eu/
GDPR	General Data Protection Regulation - https://gdpr.eu/
DMP	Data Management Plan
ELSI	Ethical, Legal and Social Implications
oAuth	open standard for access delegation
PaaS	Platform as a Service
IaaS	Infrastructure as a Service
HPC	High performance Computing
KE	Knowledge Engine
SE	Submission Engine
IC	Information Commons
EMS	Entitlements Management System
FAIR	Findable Accessible Interoperable and Reusable
AI	Artificial Intelligence
API	Application Programming Interface
TCGA	The Cancer Genome Atlas Program
DAC	Data Access Controller/Committee
ETL	Extract, Transform and Load
MW	Metadata Warehouse
CSC	CSC IT Center for Science services for science https://www.csc.fi/
REMS	CSC Resource Entitlement Management System
JWT	JSON Web Tokens https://tools.ietf.org/html/rfc7519
ML	Machine Learning

Table 1. Acronyms and Abbreviations

1.5 Structure of the document

The structure of the document is as follows:

Chapter 1 (this chapter) introduces the document, its purpose and a list of acronyms and abbreviations.

Chapter 2 illustrates the Reference Architecture and details each of the components, and the rationale behind the design of the architecture.

Chapter 3 provides an outlook of the data sources and the flow and metadata standards taken into consideration.

Chapter 4 provides a requirements analysis of the identified use cases and the correspondence to each of the components of the platform.

Chapter 5 provides the conclusions and a summary on follow-up activities.

The following documents were used or referenced in the development of this report:

- D7.1 Data Management Plan.

2 Reference Architecture Overview

2.1 Reference Architecture Overview

The general technical architecture of HEAP encompasses PaaS and IaaS that enable on-demand data streaming from remote data repositories and on-demand processing from integrated infrastructure resources.

The proposed solution for the HEAP Reference Architecture consists of the following components:

- Submission Engine;
- Information Commons;
- Knowledge Engine;
- Metadata Warehouse;
- Entitlements Management System.

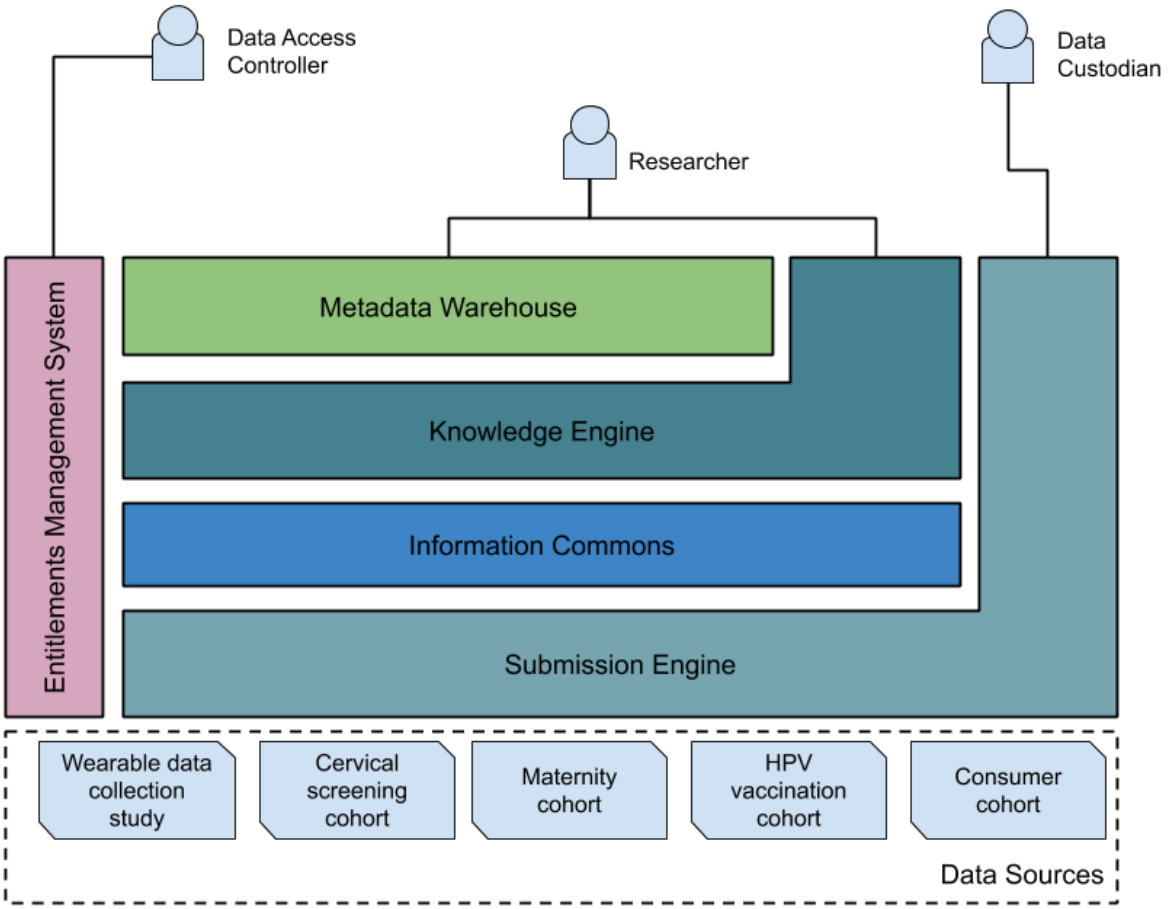

Figure 1. HEAP Reference Architecture Components

In Figure 1 we illustrate the HEAP Reference Architecture and its components as single instances, however considering the need for a distributed system, we mention that the

Information Commons, Knowledge Engine, Metadata Warehouse and Entitlements Management System components can be scaled to multiple instances. The Submission Engine layer will be able to submit data to any of the instances of the Information Commons.
Each of the layers utilises a different color in order to set them apart, or as in the case of the data sources and primary user groups to illustrate that they are representing similar concepts. The arrows illustrate a direct connection between the primary user groups and a certain layer.

As the system needs to handle heterogeneous data sources and, in some cases, even raw data, before the data could be part of the HEAP Information Commons component it needs to be harmonised. This is achieved through extraction, transformation and loaded (ETL) jobs created as part of the Submission Engine component. Therefore, this requires a dedicated layer and is positioned on top of the data sources.
The data can be on-demand stored in the IC. This implies that the data can remain at its source and is provided to the IC only when it is required for analysis. The data will be tailored to the analysis and can include pseudonymised data.

Once the data is made available in the IC, it is only accessible through the HEAP's Knowledge Engine (KE). The mechanism for accessing the data from the IC is via oAuth 2[2] and the apparatus for providing access is described in section 2.3 Authorisation and Authentication. Entitlements Management System is the software that registers the data access requirements and keeps track of the researchers that requested data access and the data authorisations.

The Knowledge Engine provides the tools and interfaces for the implementation and integration of scientific analysis workflows as bioinformatics pipelines. It provides means for applying AI (as Machine learning and Deep learning) to analyse, mine, produce and validate hypotheses and associations to extract actionable knowledge from the data. The resulting information (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[3] compliant.

Primary Users

With regard to 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 Metadata Warehouse to discover data and its features and utilise the Knowledge Engine in order to run analysis pipeline(s) for e.g. large-scale assembly inference from multiple data sources.

[2] https://tools.ietf.org/html/rfc6749
[3] https://www.force11.org/group/fairgroup/fairprinciples

A Researcher can request the Data Access Controller for access to specific data sources. The Data Custodian can make use of the Submission Engine and ETL jobs to provide the data or a subset of it depending on the request of the Researcher. The DAC will make use of the 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 KE.

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

Even though the current design of the Reference Architecture for HEAP is focused on these three user types it leaves rooms to integrate other types of users along with their needs.

2.2 Components and Component Design

Submission Engine

The IC aims to contain quality-assured and standardised data from a plethora of sources, such as registers, data from different omics experiments and analyses, data from the exposome monitoring system, as well as reference data and physical and social environments. In order to achieve the "quality" feature, or to enable researchers to use the data sources as a whole or even as a subset, the data needs to be submitted to IC via the Submission Engine.

As part of this component a Data Custodian can utilise ETL jobs to remove identifiable information from the data, create subsets or any necessary jobs in order to make data available to the Information Commons.
The SE will interface with the Information Commons via the Submission API, an API that facilitates storing sensitive data in an encrypted format.

Information Commons

Some of the main functions of the IC component are to preserve data and information for current and future research as well as provide the compute infrastructure to support the Knowledge Engine component data processing needs.

The data stored in IC will be stored encrypted utilising existing standards (recommended standards AES256[4], Crypt4GH[5] etc.) and via the *Data API* it can be provided either encrypted with a provided public key or decrypted. In both cases the HTSGET[6] protocol can be one of the standards that facilitates data transfers or streaming.

[4] https://competitions.cr.yp.to/aes.html
[5] https://www.ga4gh.org/news/crypt4gh-a-secure-method-for-sharing-human-genetic-data/
[6] http://samtools.github.io/hts-specs/htsget.html

The IC has the architecture of a vault with safety deposit boxes or private resources that only the data owner/data (custodians) of deposited data will have access and control over their respective datasets. However DAC can provide means to facilitate data access via the Entitlements Management System component.

The IC will provide two endpoints:

- Data Submission (Data In) - used by the ETL job processes (via Submission Engine component), researchers or Data Custodians to submit data;
- Data Access (or Data Out) - utilised by the KE to access data, in a controlled manner, following the Authorisation and Authentication oAuth 2.

Figure 2 illustrates how the two endpoints can be accessed via their provided interface/API.

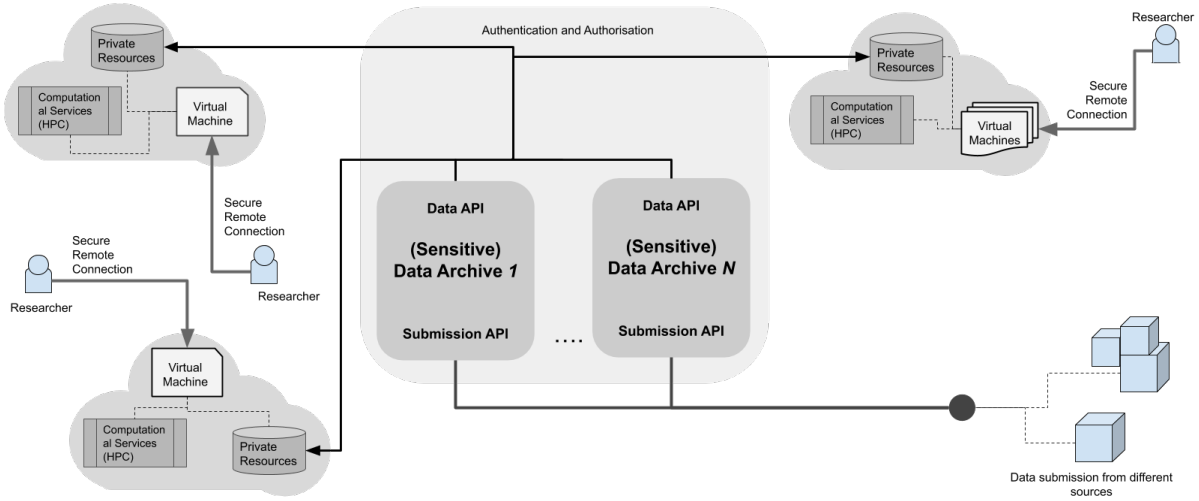

Figure 2. Information Commons ((Sensitive) Data Storage and Computing Resources)[7]

The custody over the data in the IC will not be transferred. The Information Commons (in legal terms: the hosting institutions) will simply be the processor of the data.

Knowledge Engine

The system for data-driven analysis, bioinformatics and applied AI to data in the Information Commons to produce actionable knowledge. The KE and its general requirements are satisfied by the Hopsworks[8] platform which brings great flexibility to build a continuous learning platform.

The KE will interface directly with the Researcher providing the needed functionality to run analysis pipelines, compare sequence data, visualise results and even run machine learning pipelines by integrating with the feature store - see Requirements Matrix (chapter 4). The

[7] Based on https://research.csc.fi/epouta

[8] https://hopsworks.readthedocs.io/en/latest/overview/introduction/what-hopsworks.html

Researcher will not interact directly with the Information Commons as its role is to facilitate access to the data - an instance of Hopsworks can be installed in a Virtual Machine (see Figure 2).

Among the components of the KE there is a feature store for machine learning, enabling researchers, data scientists to easily access and discover data, and features used in or to train machine learning models.

Figure 3 illustrates the overall design of the Hopsworks platform along with its components, also the data sources are provided through the Information Commons.

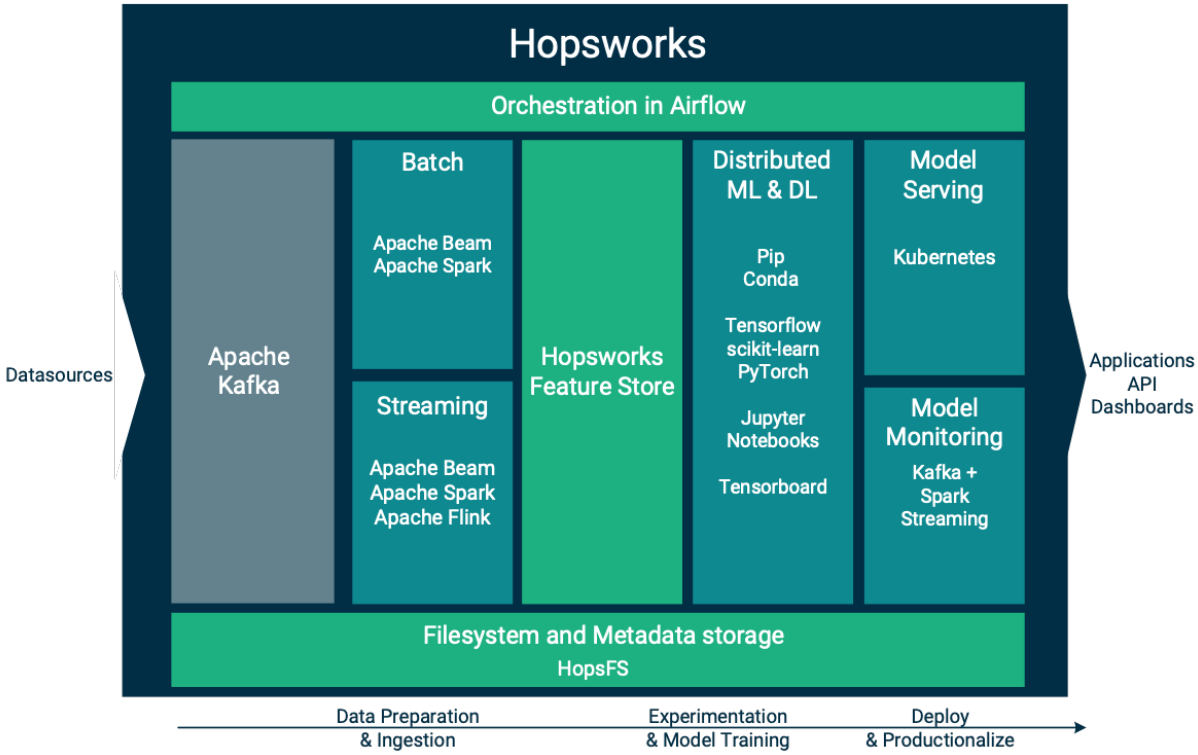

Figure 3. Hopsworks Platform

Metadata Warehouse

Metadata Warehouse acts as an online catalogue that makes metadata about the data stored in Information Commons and KE Data and Feature Store available to stakeholders and the general public. Researchers as the primary user group (or other stakeholders) will be able to find information about data and analyses in HEAP though this catalogue.

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

Another connection is with the Knowledge Engine as a collection mechanism of metadata through the Hopsworks platform, through its Metadata Layer - described in the next section, along with the collected artifacts.

The work on this component will be carried out by WP7.

Hopsworks Metadata Layer

The Hopsworks metadata layer provides the capability to share/search/reuse of the following artifacts:

- projects;
- datasets;
- feature stores - such as feature groups, features, training datasets;
- experiments;
- models;

A **project** can be thought of as an abstraction combining three entities: Data, Users, and Programs/Notebooks.

A **dataset** is a collection of files within Hopsworks. When creating a new project, a group of datasets are automatically created: Resources (programs and other artifacts required to run the programs), Jupyter (notebooks), Logs (logs from different programs/jobs run within the current project), Experiments (runs of experiments launched using the HopsML[9] API), Models (ML models), Hive[10] (HiveDB tables), Feature Store (ML features in HiveDB format), Training Datasets (curated data generated mainly from aggregating and saving in file format different features from within the Featurestore). In addition to these base datasets other datasets can be created.

Hopsworks provides a **Feature Store** to curate, store, and document features for use in ML pipelines. The feature store serves as the interface between data engineering and data science in HopsML pipelines and was designed with a focus on reuse/collaboration, versioning, automatic feature analysis etc.

The **feature** is an individual versioned and documented data column in the feature store, e.g the average rating of a customer.
The **feature group** is a documented and versioned group of features stored as a Hive table. The feature group can be seen as a context for a number of related features.

The **training dataset** is a versioned and managed dataset of features and labels (potentially from multiple different feature groups). Training datasets are stored in HopsFS as tfrecords, parquet, csv, or tsv files.

[9] https://hopsworks.readthedocs.io/en/0.10/hopsml/hopsML.html
[10] https://hive.apache.org/

HopsML provides an **experiment** API for Researchers to run their Machine Learning code, such as TensorFlow[11], Keras[12], PyTorch[13].

In the machine learning code ML **models** may be exported, typically as a result of an experiment run. HopsML facilitates versioning and attaching metadata to models to reflect the performance of a given model version.

All these artifacts contain a title, description, creator (user) and tags (user defined key-values) that are searchable and provide greater capabilities for sharing, discovering and reusing work already done within the platform. Hopsworks metadata layer also contains information to allow linkage and navigation between feature groups, training datasets, experiments and models generated.

Entitlements Management System

EMS is a tool that concentrates on two goals:

- Facilitate data access as security and rights management can make data access unnecessarily cumbersome;
- Manage access rights to resources and provide traceability of the data access rights granted.

Researchers can use their federated user IDs to log in, fill in the data access application and agree to the dataset's terms of use. The EMS system then circulates the application to the Data Access Controller or designated representative for approval. EMS also produces the necessary reports on the applications and the granted data access rights. The basis for such a system could be CSC Resource Entitlement Management System (REMS)[14].

2.3 Authorisation and Authentication

Access to Information Commons could be provided through:

- dedicated network connection (MPLS technology[15]) to ePouta cloud. The connection is already implemented between CSC and KI; and similar connections could be established for other organizations;
- remote desktop connection to ePouta cloud, that works through a web browser and standard encrypted internet connection, facilitates the interface for secure access to the data;

11 https://www.tensorflow.org/
12 https://keras.io/
13 https://pytorch.org/
14 https://www.csc.fi/en/rems-kayttovaltuuksien-hallintajarjestelma
15 https://tools.ietf.org/html/rfc3031

- Authentication and authorization for the system is handled through the EMS and validated secure standards for access delegation such as oAuth and OpenID Connect [16] and GA4GH passports.

"The GA4GH Passport uses the AAI access token to transport a researcher's digital identity and permissions across organizations, tools, and environments and then maps access to data across these" [4].

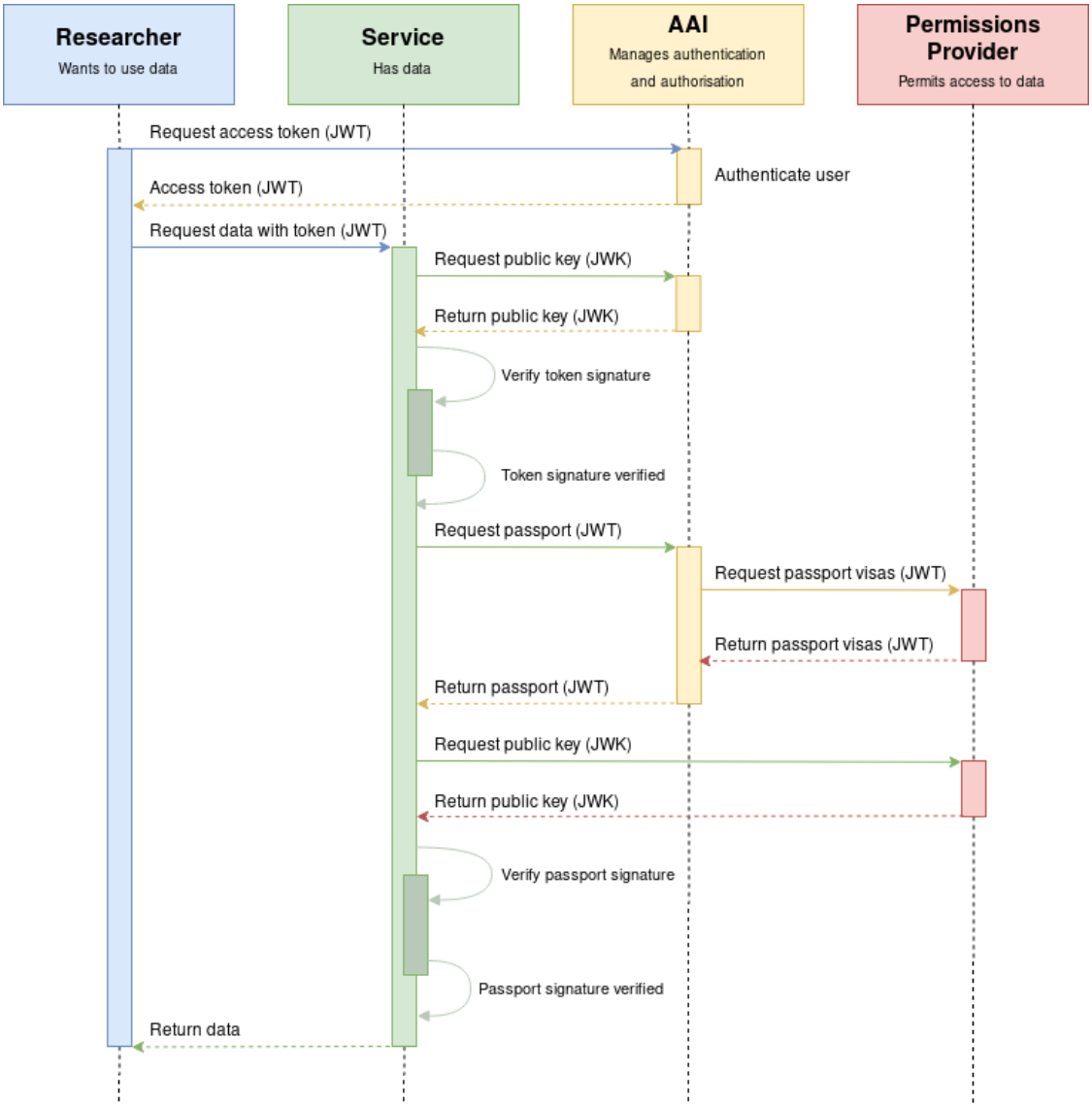

Figure 4. GA4GH Authentication sequence diagram [3] [8]

In Figure 4 we illustrate via a sequence diagram the Authentication and Authorization process utilising oAuth and GA4GH Passports. In the figure the *Researcher* represents one of the primary groups, the *Service* represents in our case the KE whilst the AAI represents the infrastructure that facilitates the authorization and authentication. The information relayed

[16] https://openid.net/connect/

to an AAI (e.g. Elixir AAI infrastructure[17] or EOSC Life Science AAI[18]) is required in order to authorise a researcher of the Knowledge Engine to access data in the Information Commons. The Entitlements Management System (EMS) component represents the *Permission Provider* that encodes in the GA4GH Passport standard [3] the information required to obtain registered or controlled access to datasets.

For public datasets access could be provided upon authentication, no authorization required. Data access KE and IC will be done aided by the use of JWT, the interface implies that KE will always ask AAI providers to facilitate the JSON Web Tokens required for data access. The GA4GH Passports specification defines within this JWT what data the user can use via the GA4GH Visa. This ensures the KE can check the user's identity and entitlements across locations (distributed installations of KE and IC).

2.4 Design Rationale

The reference architecture for the Human Exposome Assessment Platform presented in this document takes into consideration the requirements identified by the Reference Architecture Survey (Use Cases identified summarised in Chapter 4 - Requirements Matrix), the Data Sources described in the HEAP Data Management plan as well as the goals of the HEAP project as a whole.
Other design constraints are defined as key requirements presented in section 1.3.

With the proposed reference architecture, we took into consideration:

- addressing key requirements and use cases part of the HEAP project;
- reusing proven solutions and standards, thus minimising the cost to develop new components;
- minimise the complexity and at the same time provide a flexible architecture that can address future challenges throughout the HEAP project.

3 Data Overview

HEAP partners are involved in major European and international standardisation and interoperability initiatives and will align with and continue to contribute towards efforts in e.g. EOSC-Life, EJP-RD, GA4GH, GO FAIR, and RDA. Interoperability can be regarded as one of the main drivers of the platform, thus we consider where possible adopting European and international metadata and data standards.

[17] https://elixir-europe.org/services/compute/aai
[18] https://www.eosc-life.eu/aai/

3.1. Data Sources and Data Flow

As per HEAP Data Management Plan (DMP) - see WP7 D7.1 - and Reference Architecture survey (Use Cases identified summarised in Chapter 4 - Requirements Matrix), we can break down the data sources into the following types:

- HEAP - Cervical screening cohort;
- HEAP - Maternity cohort;
- HEAP - HPV vaccination cohort;
- HEAP - Consumer cohort;
- HEAP - Wearable data collection study.

The data sources and their full description and characteristics are detailed in the Data Summary (Section 1 of the HEAP DMP). The DMP gathers information from each data source considering its purpose and relation to the objectives of the platform (HEAP) and the data utility (to whom the data will be useful). Similar points have been addressed throughout the Reference Architecture document as Primary User types have been identified in section 2 and the purpose and requirements of some of the use cases are detailed in section 4.

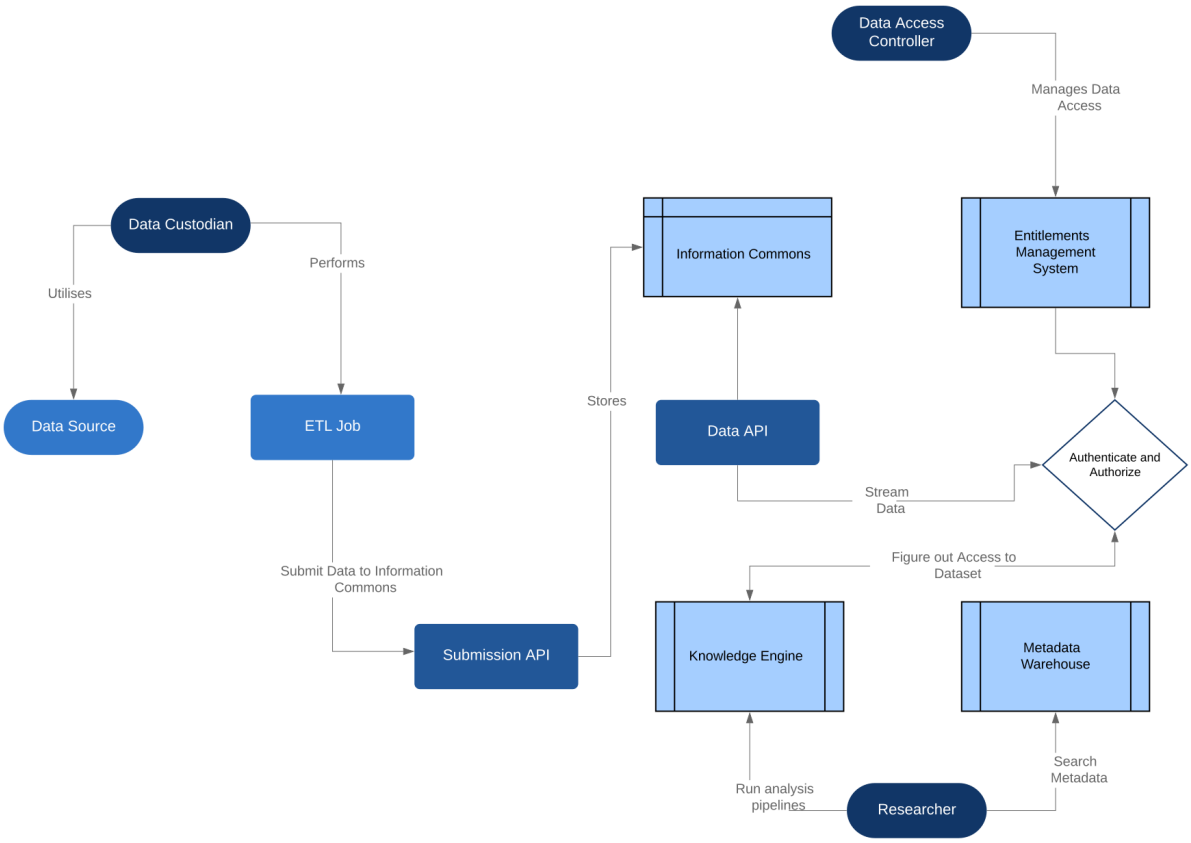

Figure 5. Data flow and the HEAP components involved.

Some of the current data types recorded in the DMP include:

- Register data (e.g. SAS, MySQL);
- Sample data (e.g. LIMS, CSV format);

- Exposome data (e.g. sequencing);
- Documentation: Word/PDF/HTML etc.

The main components that will handle any of the data types mentioned above directly are the Submission Engine, Information Commons and the Knowledge Engine, whilst the Metadata Warehouse will focus on making the data findable through metadata and providing information on how to access set data.

In a common pipeline data will flow from a data source (see Figure 5), processed in ETL jobs, and stored in the Information Commons. A Researcher will be using the processed data in the Knowledge Engine authenticating via oAuth in order to retrieve set data.

Note that in Figure 5 we depict only an instance of each of the components, however some of the components can be scaled to multiple instances and distributed across nodes as mentioned in section 2.

All approved research analyses will be conducted using the Knowledge Engine and only anonymised aggregated results will be able to be shared and published in HEAP and further referenced and described in the Metadata Warehouse. The responsibility for anonymising the data is on the *Data Custodian*.

3.2. Metadata

The main standards for metadata creation are described in section 2.1.2 of the DMP.
As specified by DMP (WP7 D7.1) sample information will be harmonized according to the MIABIS[19] standard.

For metagenomics data analysis we will collect store metadata for:

1. study and collection context and information, some proposed standards are:
    a. Minimum information about a Genome Sequence (MIGS) and Minimum information about Metagenomics Sequence (MIMS)[20];
    b. Minimum information about a marker gene sequence (MIMARKS)[21];
    c. Minimum information about any (x) sequence (MIxS)[22];
2. description of the sequencing process e.g. controlled vocabularies;
3. analysis metadata with standards such as Common Workflow Language (CWL)[23];
4. information archived datasets such as EGA/ENA metadata standards[24].

[19] http://www.bbmri-eric.eu/services/miabis/
[20] https://www.nature.com/articles/nbt1360
[21] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367316/
[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367316/
[23] https://www.commonwl.org
[24] https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

Other standards to be considered for reuse as part of the metadata are one proposed by regulatory bodies (SNOMED[25], ICTV[26], etc.) and naming convention such as the ones provided by TCGA[27] database. For embedding dataset permissions in the Entitlement Management System Data Use Ontology[28] could be utilised.

In order to reference datasets throughout the system, persistent identifiers need to be established. Information Commons supports the use of public and permanent identifiers.

To maximize interoperability the following persistent and unique identifiers and controlled vocabularies will be considered: such as DOI, XRI/IRI/URI, schema.org, bioschemas.org, and other similar standards. For a full list consult the D7.1 Data Management Plan.

4 Requirements Matrix

4.1 Use Cases

Identifying the use cases and the requirements has been done via two surveys, one performed within the WP7 and one in WP10, both aiming to understand the data sources, the needs of the researchers and Data Custodians.

	Use Case	**Use case description**
UC1	Maternity cohort and HPV vaccination cohort managed in a more efficient way and linked to a metadata search engine.	Make use of a metadata search engine in order to make data more accessible and easier to find via a metadata search engine.
UC2	Analyse available sequence data in Sweden (e.g. Cervical screening cohort) and from US public database	nalyse for the presence of all exposures (viral, bacteria, others), using existing bioinformatics pipelines developed, and make use of them in the HEAP platform installed at partner nodes (e.g. CSC, KI).
UC3	Consumer cohort data analysis	Make use of IC as a storage engine in order to work/combine with other data - e.g Analysis of consumer data determinants of health and

[25] http://www.snomed.org/
[26] https://talk.ictvonline.org/
[27] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[28] https://github.com/EBISPOT/DUO

		diseases.
UC4	Investigations on how to best explore the advantages of Hopsworks	Pilot retrieving data archived in IC for analysis and/or for processing from Knowledge Engine (Hopsworks).

Table 2. Use cases identified as per Reference Architecture survey.

The table below summarizes the Use cases identified via the Reference architecture survey, while in Table 3 we combine data from that survey with data gathered via the DMP survey in order to obtain a holistic view of some of the requirements per each use case.

Note: The requirements identified above do not represent the full set of requirements of the HEAP Reference Architecture and more requirements may follow as use cases are being implemented. The same is valid for use cases, as more use cases could be identified throughout the HEAP project.

4.2 Requirements

In the initial use case Use Case analysis for the HEAP project we identified the following requirements:

- **RQ1**: Compare sequence data to existing microbe sequence data and perform large scale metagenome analysis (such as assembly, classification and network analysis, etc);
- **RQ2**: Enable controlled data access, as some datasets require a traceability of the data usage purpose or if data can or cannot be transferred across country borders (or is behind a firewall);
- **RQ3**: Enable large scale metagenome analysis and inference in computational environments (HPC) that support parallelisation;
- **RQ4**: Provide support for plotting and visualising across datasets;
- **RQ5**: Work heterogeneous and distributed datasets;
- **RQ6**: Manage and make the data findable and accessible.

In Table 3 we will match the requirements to the HEAP components that could facilitate satisfying the requirements in part or as a whole.

	Requirement	**RQ1**	**RQ2**	**RQ3**	**RQ4**	**RQ5**	**RQ6**
Platform Component							

Submission Engine					x	x	
Information Commons			x	x	x	x	
Knowledge Engine		x		x	x	x	
Metadata Warehouse		x					x
Entitlements Management System			x			x	x

Table 3. Requirements Matrix

The requirements for the HEAP sensitive data platform are derived from research use cases, especially those utilising register or controlled data (WP3). The Information Commons will adopt the recommendations on legal and ethical items from WP2. In addition, WP10 collaborates with WP6 to develop and test integration of the IaaS and PaaS platforms in the consortium, to enable seamless data use for additional scenarios of register research and big data analysis, than the ones identified via the survey.

5 Summary

The purpose of this document is to provide guidance and direction for implementing and working HEAP components, by identifying technologies and standards necessary for the development of such a platform.

5.1 Collaboration with other Work Packages

WP10 collaborates with WP2 and WP6 to develop and integrate ethically sound, secure, and technically robust IaaS and PaaS that will enable all the components detailed in the document to function seamlessly in order to process and analyse exposome and biomedical data sources from WP3-5 organised under the semantics designed by WP7.

Reference Architecture Components in the Context of HEAP Project

In order to orchestrate the development of the components detailed in this Reference Architecture document we propose that the work could be divided in the following work packages:

- WP6 focuses on Knowledge Engine;
- WP7 targets the Metadata Warehouse;
- WP10 handles the Information Commons, Entitlements Management System and necessary interfaces for the Submission Engine.

5.2 Achievements and Future Plans

The current document encompasses the work of gathering use cases and requirements for the HEAP platform and WP10, alignment with HEAP Data Management Plan (WP7), establishing a terminology of different components of the platform, identifying standards as well as designing a Reference Architecture and its components, that other work packages can reference throughout the development of the platform.

As a follow up of this document, the following steps have been planned:

- Continuous update and refinement of the Reference Architecture leveraging the feedback and interaction with other WPs;
- Work towards finalizing the structure of the HEAP components, interfaces and standards to be utilised;
- Identifying interoperability endpoints between components through pilots.

Long term goals of the WP10 are grouped around implementation and deployment of the HEAP IC and EMS components and making the services operational, and can be further broken down into:

- **Development of the sensitive data management platform** - This task develops the secure data storage and secure cloud environment that form the core of the Information Commons for the HEAP proposal.
- **Dataset authorization and access** - This task develops and implements authorization and controlled access mechanisms for the data on the sensitive data platform, including secure remote desktop with restricted download of data. The data authorization management will be implemented using the REMS tool that is developed by CSC.
- **Secure data access from institutional repositories** - This task implements technology that makes data stored at local repositories available for processing at the secure infrastructure.
- **Support research use cases** - This task will implement and support the use of secure infrastructure for research use cases from the HEAP consortium utilising the data from project partners.

6 References

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

[2] CSC Cloud computing services
Available from: https://research.csc.fi/cloud-computing (Accessed June 2020)

[3] Data Use and Researcher Identity (DURI)
GA4GH Passport Specification. Available from:
https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md
(Accessed June 2020)

[4] GA4GH Passports Announcement. Available from:
https://www.ga4gh.org/news/ga4gh-passports-and-the-authorization-and-authentication-infrastructure/
(Accessed June 2020)

[5] Lappalainen I., Almeida-King J., Kumanduri V., Senf A., Spalding J.D., 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 A.C., de Argila J.R., Llobet O.M., Medina I., Puy M.S., Alberich M., de la Torre S., Navarro A., Paschall J. and 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

[6] Lappalainen I., Lopez J., Skipper L., Hefferon T., Spalding J.D., Garner J., Chen C., Maguire M., Corbett M., Zhou G., Paschall J., Ananiev V., Flicek P. and Church D.M.
DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res Volume 41 (2013) p.d936-41 DOI: 10.1093/nar/gks1213

[7] Linden M., NyrönenT. and Lappalainen I.
Resource Entitlement Management System, Selected papers of TNC2013 Conference http://www.terena.org/publications/tnc2013-proceedings/ (2013)

[8] Linden M. , Voisin C. and Bernick D.
GA4GH Authentication and Authorization Infrastructure (AAI) OpenID Connect Profile
Available from: https://github.com/ga4gh/data-security/blob/master/AAI/AAIConnectProfile.md
(Accessed June 2020)