D5.7

Design of the Data Integration Platform - v1

ABSTRACT

This document describes the fundamental software components---together referred to as *Data Integration Platform*---that are responsible for the conversion and translation of Electronic Health Records across languages and across local and national healthcare standards.

Delivery Date	3rd October 2019
Work Package	WP5
Task	T5.2
Dissemination Level	Public
Type of Deliverable	Report
Lead partner	UNITN

InteropEHRate project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106.

This document has been produced in the context of the InteropEHRate Project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106. All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.

CONTRIBUTORS

	Name	Partner
Contributors	Gábor Bella	UNITN
Contributors	Roberto Bona	UNITN
Contributors	Alessio Zamboni	UNITN
Contributors	Simone Bocca	UNITN
Reviewers	Theodora Zacharia	ISA
Reviewers	Sébastien Hannay	A7
Reviewers	Juan Fernandez	EFN

LOG TABLE

Version	Date	Change	Author	Partner
0.1	2019-08-20	First draft created	Gábor Bella	UNITN
0.2	2019-08-30	Contributed information on platform development progress and plans	Roberto Bona	UNITN
0.3	2019-09-10	Contributed information on platform implementation details	Alessio Zamboni	UNITN
0.4	2019-09-13	Finished first version, ready to be reviewed	Gábor Bella	UNITN
0.5	2019-09-13	Review with comments	Juan Fernandez	EFN
0.6	2019-09-17	Review with comments	Sébastien Hannay	A7
1.0	2019-10-02	Quality check	Argyro Mavrogiorgou	UPRC
VFinal	2019-10-03	Final check and submission	Laura Pucci	ENG

ACRONYMS

Acronym	Term and definition
CSV/TSV	Comma-Separated Values / Tab-Separated Values: simple machine-readable tabular file formats.
EHR	Electronic Health Record (e.g., as provided by a hospital).
IHS	InteropEHRate Health Services: a high-level software component (a collection of libraries) that provides high-level EHR translation and conversion services to end-user applications.
IHT	InteropEHRate Health Tools: a set of interactive helper tools that are used by hospital employees (data scientists) to set up and maintain the EHR data integration system. These tools are outside the scope of the project deliverables and thus are not fully specified in any deliverable document.
SEHR	Smart Electronic Health Record: the interoperable, multilingual, standard, FHIR-based representation of EHRs as defined and used by the InteropEHRate project.
XML	Extensible Markup Language: machine-readable tree-structured file format.

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1. Scope of the document 1
1.2. Intended audience 1
1.3. Structure of the document 1
1.4. Updates with respect to previous version (if any) 1
2. PLATFORM ARCHITECTURE 2
3. THE KNOWLEDGE LAYER 3
3.1. EHR Knowledge Storage 3
3.2. Knowledge Integration 3
3.3. Knowledge Query 4
4. THE EHR DATA LAYER 6
4.1. EHR Integration 6
4.2. EHR Data Cache 6
4.3. EHR Data Query 6
5. CONCLUSIONS AND NEXT STEPS 8

LIST OF TABLES

Table 1 - Principal endpoints of the Platform knowledge importing API 4
Table 2 - Principal endpoints of the Platform knowledge query API 4
Table 3 - Principal endpoints of the Platform EHR data query API 7

LIST OF FIGURES

Figure 1 - The Data Integration Platform and its use by outside components 2

1. INTRODUCTION

1.1. Scope of the document

This deliverable provides a first specification of the *InteropEHRate Health Data Integration Platform* (abbreviated as *Platform* in the rest of the document). The Platform is a fundamental system that provides low- and mid-level functionalities for a deep, *semantic*---i.e., meaning-level---integration, conversion, and translation of Electronic Health Records (EHRs in the following). Other InteropEHRate deliverables define how these functionalities are used by higher-level *conversion* and *translation services* to convert and translate EHRs [D5.9] and how such converted EHRs are presented to healthcare professionals [D5.4].

1.2. Intended audience

The Platform is a low-level software component used by higher-level health IT services rather than directly by end users. Consequently, this deliverable is primarily aimed at technical audiences, such as the IT managers/staff of a hospital or third-party developers of semantic services on top of EHR data, who wish to understand:

- how InteropEHRate Health Services work internally;
- how to build services on top of the Platform.

The deliverable focuses on the *architecture*, components, and interfaces of the Platform. It does *not* present in detail the process by which the Platform and the services built on top need to be set up and used by the hospital. In order to understand the detailed process, the reader is referred to [D5.9].

1.3. Structure of the document

Section 2 presents the high-level architecture of the Platform, its components, and its interfaces both on the input and the output side. Sections 3 and 4 present the two layers of the Platform: the Knowledge Layer and the EHR Data Layer, respectively. The two sections explain the role of each layer, this principal components, and the principal API endpoints that they provide to the services using the Platform. Section 5 provides conclusions and next steps.

1.4. Updates with respect to previous version (if any)

Not applicable.

2. PLATFORM ARCHITECTURE

The role of the Platform is to provide fundamental syntactic and semantic data integration and query functionalities to the *InteropEHRate Health Services* (IHS). Figure 1 below shows the high-level architecture of the Platform, as well as its relation to the IHS and to the users of the IHS: on the one side, *Hospital Information Systems*, and on the other side, applications that consume integrated EHRs such as the *HCP App* and the *SEHR Cloud*.

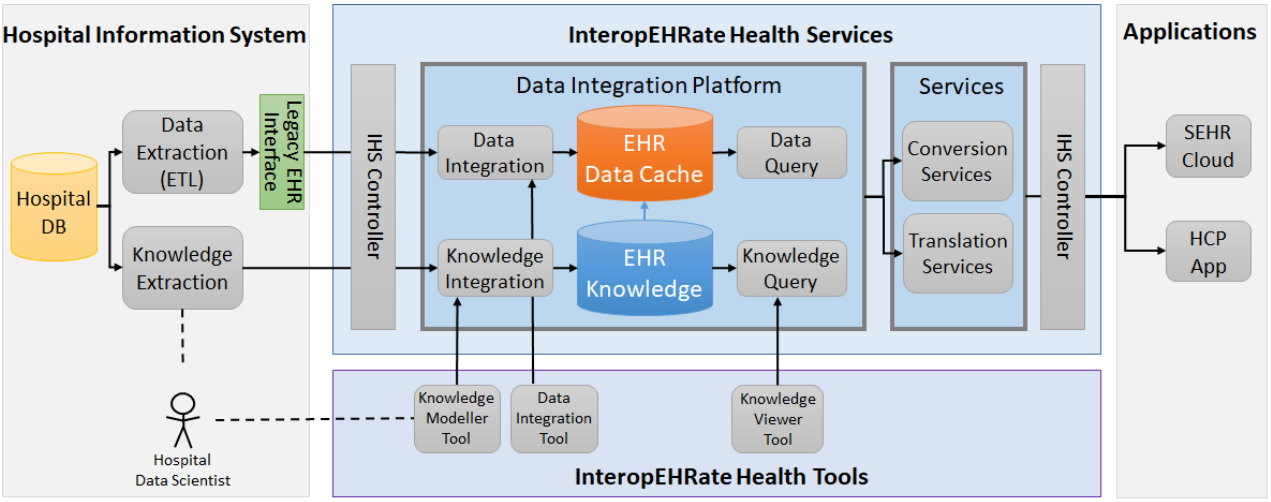

Figure 1 - The Data Integration Platform and its use by outside components

The Platform is composed of two horizontal layers:

1. the **Knowledge Layer** that is responsible of patient-independent healthcare knowledge, such as data structures, encodings, and terminologies;
2. the **Data Layer** that is responsible of patient-specific EHR data and that uses integrated knowledge from the Knowledge Layer.

Vertically, the Platform is divided into three phases:

1. the **Integration phase** that takes informally or semi-formally (i.e., non-semantically) defined knowledge and data from data providers (i.e., hospitals) and converts them to a formal semantic representation;
2. the **Storage phase** that stores the integrated semantic representations;
3. the **Query phase** that serves integrated knowledge and/or EHR data to consumers according to their requirements with respect to language (e.g., French or Italian), terminology (e.g., ICD-9 or ICD-10), and structure (e.g., FHIR).

The process through which the Platform is used---that is, filled in with knowledge and EHR data and subsequently queried---is described in detail in the deliverable [D5.9].

3. THE KNOWLEDGE LAYER

The Knowledge Layer of the Platform takes as input *informal* or *semi-formal health knowledge*, such as human-readable or machine-readable descriptions of:

- **EHR data structures** as provided by a given hospital;
- **coded values** as used within EHRs of the hospital;
- **terminology** and **natural language** words as used by the hospital.

The goal of the Knowledge Layer is to formalise such health knowledge into a machine-readable form that the Platform can use to automate the integration of EHRs. While the full health knowledge on which a hospital implicitly relies is vast, only a small subset of it needs to be formalised as required for the representation of EHRs.

3.1. EHR Knowledge Storage

The central storage component of the Knowledge Layer (depicted as the blue DB in the figure above) maintains formal knowledge that describes both local hospital-specific knowledge (terminology and coded values) and their mappings to international knowledge as used by InteropEHRate, such as international codes and FHIR data structures. For each hospital, the Storage component comes pre-loaded with formalised international knowledge:

- FHIR data structures;
- international terminology such as SNOMED CT, LOINC, or ICD-10;
- if available, the lexicalisations of all of the above in the local language, as well as in any other language that the local institution is willing and capable of supporting (e.g., in multilingual regions of Europe the support of more than one language may be required);
- international data instances ("entities") such as richly described pharmaceutical substances as provided by WHO ATC.

3.2. Knowledge Integration

The initial *Knowledge integration* phase consists of formalising local, hospital-specific knowledge, as well as its mapping to international knowledge, into a *formal knowledge representation* that the Platform can use in order to automate the conversion and translation of EHRs. Knowledge integration is always a manual or semi-automated process that is governed by persons provided by the hospital having the following roles:

- a **data scientist** whose role is to understand the human-readable descriptions of informal health knowledge, and subsequently to use interactive tools to convert these descriptions into formal knowledge. In particular this person has to know the format of the data in the hospital database as well as the medical knowledge to understand the data.
- a **software developer** who assists the data scientist in case the formalisation activity is automatable and would otherwise be too onerous to implement manually, such as in the case of integrating large terminologies with tens of thousands of entries.

It is possible that no full documentation of all aspects of knowledge above is available at a given hospital, but some of the knowledge is implicit within the hospital's information system (e.g., embedded within tools and local data structures). The manual *knowledge extraction* process, executed by the data scientist, is supposed to identify such implicit forms of knowledge and make them explicit, e.g., in the form of a report.

Knowledge integration produces the following results:

- **spreadsheets** (of simple Excel/CSV format) that formally describe the terminology and coded data values used by the hospital, as well as how they map to international terminology as used by

InteropEHRate. These spreadsheets may either be produced manually by the data scientist (if their contents are small) or automatically by scripts written by the software developer.

- **direct intervention on the EHR Knowledge Store** by the data scientist using the *Knowledge Modeller Tools*, for the purpose of smaller updates on knowledge.

For both of these knowledge integration modalities above, RESTful *importing APIs* are used, exposed by the EHR Knowledge Storage component.

Endpoint	Description
importLanguage	Imports an Excel sheet that contains concepts and their corresponding words (lexicalisations) in one or more languages
importTypes	Imports data structures (so-called *entity types*) in OWL format.

Table 1 - Principal endpoints of the Platform knowledge importing API

These endpoints are mainly used by *InteropEHRate Health Tools* for knowledge management.

3.3. Knowledge Query

Integrated knowledge can be queried through RESTful APIs for the following purposes:

- browsing and visualisation of formal hospital-specific knowledge, international IEHR knowledge, and their mappings;
- the automation of data mappings, conversions, and translations.

The following query API endpoints are exposed by the Platform (the list is not exhaustive and contains only the most important endpoints):

Endpoint	Description
vocabularies	Returns the languages supported (e.g., English, Italian, or French).
concepts	Allows querying the general and healthcare concept hierarchy. Concepts are language-independent units of meaning, represented as numerical IDs, that formalise healthcare terms such as diseases, procedures, or lab tests and their results.
words	Returns the lexicalisation ("translation") of a concept in a given language (vocabulary).
types, attributedefinitions	Returns the descriptions of the FHIR data structures (resources) supported and of their attributes.

Table 2 - Principal endpoints of the Platform knowledge query API

These endpoints are mainly used by the *Conversion Services* and *Translation Services* of the *InteropEHRate Health Services* for the purposes of converting EHRs to the FHIR-based representation, to map the coded values

contained within, and to translate their textual across languages. These services are described in the deliverable *D5.9 - Design of the Data Mapper and FHIR Converter*.

4. THE EHR DATA LAYER

The purpose of the Data Layer is:

1. to **integrate** local EHRs, i.e., convert them from their local representations into a formal, supra-lingual and international representation;
2. to **store** these representations temporarily in an *EHR Data Cache* (the time of storage being customisable by the hospital);
3. to **query and retrieve** these representations in standard serialised FHIR format and in a given language, ready to be used by FHIR- and SEHR-compliant applications such as the *HCP App*, the *SEHR App*, and the *SEHR Cloud*.

4.1. EHR Integration

The integration is a semi-automated process that is executed in two phases:

- **semi-automated setup phase:** in this manually executed phase a data scientist from the hospital defines the rules, based on a small- or medium-sized corpus of local EHRs, by which the local EHR data structures and data values will be mapped to FHIR: this set of rules or "recipe" is called the *data integration model*;
- **automated data integration phase:** in this fully automated phase the Platform converts EHRs automatically from their local representation into their FHIR-based international representation, using the "recipe" defined by the data scientist in the setup phase.

In the setup phase, the data scientist uses the interactive *StarLinker Data Integration Tool* from the *InteropEHRate Health Tools*. The output of the tool is the data integration model (recipe) that helps automate the integration process.

In the production phase, an automated data integration service uses the recipe from above to automate the conversion of local EHRs. The output of the production phase, namely the converted health record, is directly written into the *EHR Data Cache*.

In both phases, the input consists of one or more EHRs, each EHR physically represented as a set of files. Each file corresponds to a "data table" or data structure as provided by the hospital. The files must be expressed in a table-structured CSV/TSV (comma-separated/tab-separated) format or in a tree-structured XML or JSON format.

4.2. EHR Data Cache

The *EHR Data Cache* is logically a graph database that stores EHRs as knowledge graphs. The nodes of the knowledge graph are qualified using the *EHR Knowledge* defined in the Knowledge Layer. The graph uses a language-independent representation of the EHR that can subsequently be queried and retrieved in any language provided by the EHR Knowledge.

4.3. EHR Data Query

The EHR Data Cache provides a search and query API that allows the retrieval of both entire SEHRs and portions of SEHR data. The principal output formats provided are:

- for entire SEHRs or FHIR resources: a serialised FHIR format (XML or JSON);
- the Platform-internal and JSON-based *EML (Entity Markup Language)* format that preserves the original graph structure and can thus be used for the transfer of rich graph data across systems;
- JSON snippets for individual data values and fine-grained data.

The API is able to cater to the requirements of all three IEHR scenarios:

- scenarios 1 and 2 require the retrieval of entire (converted and translated) SEHRs;
- scenario 3 requires the retrieval of subsets of patient data specific to the research experiment, corresponding to a *research query* that may be formulated based on meaning as opposed to surface representations.

The table below provides a summary of the principal API endpoints that we foresee to be used within the project.

Endpoint	Input	Description
instances, attributes	instanceID, attributeID, lang	These low-level calls return data instances (corresponding to the FHIR resources of a SEHR) or individual attribute values, expressed in the language given as input. The output format is either the Platform's internal EML graph data format or RDF.
exportSEHR	ehrID, lang	This extension of the Platform query API allows the export of an entire SEHR in serialised FHIR format, in the language given as input.
simpleSearch	searchString	Returns data values corresponding to a conventional text-based search. This functionality can be used by the Research Scenario.
semanticSearch	searchString	Returns data values corresponding to a "semantic" search where the search is formulated not through words but through concepts. It returns data values that are equivalent or more specific than the query concept. For example, a search for the concept of "cerebrovascular disease" will return EHRs with all kinds of cerebrovascular diseases mentioned within, such as "cerebral infarction". This functionality can be used by the Research Scenario.
simpleQuery	queryString	Returns data according to a simple SQL-like query over integrated EHR graph data. This functionality can be used by the Research Scenario.
semanticQuery	queryString	Returns data values corresponding to a "semantic" query. Prior semantic analysis of natural-language data values allows semantic queries to retrieve EHRs based on the meaning of the text contained within. "Semantic" queries allow for semantic reasoning over data values, e.g., "SELECT * WHERE disease is_a *'cerebrovascular disease*' " would return all EHRs with a specific type of cerebrovascular disease, e.g., cerebral infarction. This functionality can be used by the Research Scenario.

Table 3 - Principal endpoints of the Platform EHR data query API

5. CONCLUSIONS AND NEXT STEPS

This document provided the first (v1) specifications for the IEHR Data Integration Platform. The specifications will continue to evolve as development progresses and the precise project needs are discovered, especially relating to the formats and challenges related to processing legacy hospital data from the hospital partners in the four participating countries.

In particular, an active collaboration and co-development with project partners will be necessary in the following areas:

- with hospitals: specify the input legacy EHR formats that the Platform can accept and process, and some minimum requirements for the hospitals' records to be processed;
- with all: specify more detailed data integration and query requirements based on the needs of the three scenarios.

REFERENCES

- **[D2.4]** InteropEHRate Consortium, InteropEHRate Architecture v1 (deliverable D2.4). https://www.interopehrate.eu/resources/
- **[D2.7]** InteropEHRate Consortium, Interoperability Profile v1 (deliverable D2.7). https://www.interopehrate.eu/resources/
- **[D5.9]** InteropEHRate Consortium, Design of the Data Mapper and Converter to FHIR (deliverable D5.9). https://www.interopehrate.eu/resources/
- **[D5.11]** InteropEHRate Consortium, Design of information extractor and natural language translator v1 (deliverable D5.11). https://www.interopehrate.eu/resources/
- **[D5.4]** InteropEHRate Consortium, Design of an integrated EHR web app for HCP - V1 (deliverable D5.4). https://www.interopehrate.eu/resources/