



Call identifier: H2020-ICT-2016 - **Grant agreement no:** 732907

Topic: ICT-18-2016 - Big data PPP: privacy-preserving big data technologies

Deliverable 4.2 MHMD Ontological Resources

Due date of delivery: April 30th, 2018

Actual submission date: April 30th, 2018

Start of the project: 1st November 2016

Ending Date: 31st October 2019

Partner responsible for this deliverable: **HES-SO**

Version: **1.0**



D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
---------------------------------	------------------------------

Document Classification

Title	MHMD Ontological Resources
Deliverable	D4.2
Reporting Period	01.11.2016 - 30.04.2018
Authors	Douglas Teodoro, Emilie Pasche, Rudolf Mayer, Patrick Ruch
Work Package	WP4
Security	Public
Nature	Report
Keyword(s)	Ontology; Data model; Data stewardship

Document History

Name	Remark	Version	Date
Douglas Teodoro	First Version	0.1	15.03.2018
Douglas Teodoro	Description of catalogue, dictionary and normalization services	0.2	10.04.2018
Emilie Pasche	Prototype description	0.3	11.04.2018
Rudolf Mayer	Description of track and provenance strategies	0.5	
Douglas Teodoro	Consolidation	0.9	15.04.2018

List of Contributors

Name	Affiliation
Douglas Teodoro	HES-SO
Emilie Pasche	HES-SO
Rudolf Mayer	SBA

List of reviewers

Name	Affiliation
Minos Garofalakis	Athena

1	PREFACE	5
2	EXECUTIVE SUMMARY	6
2.1	SCOPE	6
2.2	PROBLEM BEING ADDRESSED	6
2.3	SCIENTIFIC APPROACH AND WORK UNDERTAKEN	6
2.4	ACHIEVEMENTS	6
2.5	RELATIONSHIP TO THE REST OF THE PROJECT	6
2.6	CONFORMANCE TO THE “DESCRIPTION OF WORK”	6
2.7	NEXT STEPS	6
3	INTRODUCTION	8
3.1	OBJECTIVES	8
3.2	SCOPE AND CONTEXT	8
3.3	STATE-OF-THE-ART	9
3.4	DEFINITIONS, ACRONYMS AND ABBREVIATIONS	10
3.5	OVERVIEW	10
4	CATALOGUE OF MHMD DATA SOURCES	11
4.1	QMUL DATA SAMPLE	11
4.2	DIGI.ME DATA SAMPLES	14
4.2.1	<i>Health data</i>	14
4.2.2	<i>Social media data</i>	19
4.2.3	<i>Financial data</i>	22
5	ANALYSIS OF TERMINOLOGIES AND ONTOLOGIES FOR REPRESENTING MHMD DATASETS AND CONSTRUCTION OF MHMD DATA DICTIONARY	24
5.1	RESOURCES	24
5.1.1	<i>Anatomical Therapeutic Chemical (ATC)</i>	24
5.1.2	<i>International Classification of Diseases (ICD)</i>	24
5.1.3	<i>Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)</i>	24
5.1.4	<i>Medical Subject Headings (MeSH)</i>	25
5.1.5	<i>digi.me ontology</i>	25
5.2	ONTOLOGY ASSESSMENT	25
5.2.1	<i>Top down approach</i>	25
5.2.2	<i>Bottom-up ontology approach</i>	26
5.3	DATA DICTIONARY FOR THE MHMD SAMPLE SETS	26
6	UNIFIED DATA CATALOGUE METADATA MODEL	31
6.1	EXISTING METADATA MODELS FOR BIOMEDICINE	31
6.1.1	<i>i2b2 data mart model</i>	31
6.1.2	<i>Bioschemas specifications</i>	31
6.1.3	<i>DATS metadata model</i>	32
6.2	DATS PHYSICAL MODEL AND INSTANTIATION OF MHMD DATASETS	32
6.3	LOGICAL MODEL ON TOP OF DATS	34
7	DATA NORMALIZATION SERVICE	35
8	DATA STEWARDSHIP	ERRORE. IL SEGNALE NON È DEFINITO.
8.1	PROVENANCE	ERRORE. IL SEGNALE NON È DEFINITO.
8.2	VERSIONING	ERRORE. IL SEGNALE NON È DEFINITO.
9	DATA CATALOGUE PROTOTYPE	41
9.1	INTERFACE	41
9.2	QUERY EXAMPLE	44
9.3	DATA CATALOGUE INDEX	45
9.4	CATALOGUE API	45
9.4.1	<i>View of the data</i>	45
9.4.2	<i>Search for several records</i>	47

9.4.3 Search for one record..... 48

10 CONCLUSION AND NEXT STEPS.....50

REFERENCES.....51

APPENDIX.....52

SAMPLE SETS52

QMUL.....52

digi.me.....52

DIGI.ME ONTOLOGY60

1 Preface

This document describes the work performed in the context of WP4 Data Harmonisation Module Extensions to create the MHMD harmonized data catalogue service. The report provides a description of the representative synthetic datasets shared within the consortium and describes the methodologies used to homogenize these datasets into a single semantic metadata model. The work performed is mostly related to tasks T4.1 Construction of Data Catalogue and T4.2 Data Dictionary, which ended in M12 and M18 respectively. Nevertheless, to show the big picture of WP4, we also include the work performed in the context of tasks T4.3 Development of Data Stewardship Modules (section 8) and T4.4 Semantic Querying and Cohort User Interface (section 9), both of which will start in M18.

2 Executive Summary

2.1 Scope

The MyHealthMyData (MHMD) platform recognizes four main stakeholders in the data value chain - hospitals, citizens, research centres and industry - with different interests. While citizens and hospitals will share very heterogeneous and privacy sensitive datasets in the network, research centres and industry need streamlined and homogeneous ways to search, discovery and access these datasets. In this context, WP4 provides the services for harmonizing, ingesting, cataloguing and discovering dataset metadata across the MHMD network.

2.2 Problem being addressed

In WP4, we face the challenge to make data sources and their respective datasets, shared in the MHMD platform, searchable, findable and accessible. As data shared in the network are highly heterogeneous and little formalized, methods for harmonising these data need to be developed to improve the interoperability and actual data reutilization within MHMD framework. These challenges are accentuated due to the sensitivity of the data. Thus, these methods should be designed taking into account privacy and confidentiality issues.

2.3 Scientific approach and work undertaken

In this deliverable, we describe the construction of the MHMD data catalogue and the elaboration of the data dictionary. Using a top-down approach, we describe and organize the datasets shared using the repository provided in deliverable D4.1 "Repository to Share Data Samples" into modalities. Then, using a top-down approach, we specify the formal model to represent MHMD metadata and the reference terminology to represent MHMD metadata content. Eventually, we develop an automatic methodology, based on the Metamap algorithm, to normalize the dataset contents against reference terminologies.

2.4 Achievements

We detail the data dictionary for the datasets shared in the data sample repository, using MeSH as global reference terminology. Additionally, we implemented a REST service to harmonise dataset attributes, in particular, demographics, diagnosis, medication, and procedures. Finally, we developed a prototype of metadata catalogue, for registering the synthetic dataset shared in the consortium.

2.5 Relationship to the rest of the project

This deliverable is based on the use-cases developed in WP1 - Requirement Analyses and is informed by WP2 on issues related to synthetic data and metadata privacy. Moreover, it is tightly aligned with WP3, concerning the registration of metadata in the onboarding process, and with WP5, for the definition of and access to harmonised cohorts.

2.6 Conformance to the “Description of work”

This report describes the work developed in the context of deliverable D4.2 MHMD Ontological Resources, which aims to describe the reference terminologies selected in task T4.2 "Data Dictionary", the strategies to normalize data, and track provenance and versioning (T4.3). It describes the work to construct the data catalogue, data dictionary, and data and model normalization into a unified metadata model.

2.7 Next steps

As next steps, we will work on the data endpoints and cohort construction prototypes, allowing access to data and description of endpoints. In particular, enabling different views via semantic querying. Additionally, we will develop the data stewardship modules to enhance the local MHMD repositories with provenance tracking and data subset citation capabilities.

3 Introduction

The MHMD platform recognizes four main stakeholders in the data value chain - hospitals, citizens, research centres and industry - with different interests. The MHMD architecture aims to align these stakeholders in a transparent and efficient way, allowing research centres and businesses to seek access to large volumes of longitudinal data to develop novel medical services and analyse trends. Thus, MHMD focuses on an information architecture bridging personal data sources with clinical histories, lab tests and diagnostic images, providing an end-to-end platform for knowledge discovery at both population and individual level. In this context, WP4 provides the services for harmonizing, cataloguing and discovering datasets across the MHMD network.

3.1 Objectives

The main goal of WP4 is to provide data harmonisation services, featuring functionalities for data normalisation, publication and search using a minimal set of semantic descriptors and dataset properties. The data harmonization layer will allow data to be searchable, findable, interoperable and re-usable within the MHMD network. To achieve these goals, a few issues are being investigated, such as the level of published metadata that enables traceability of a specific dataset while keeping individual privacy, and its integration with the privacy preserving and application layers. Indeed, the main challenge for the data harmonisation layer will be to demonstrate how it conforms to the privacy preserving features identified in the course of the project while allowing data to be catalogued and discovered.

3.2 Scope and context

The implementation of the data harmonisation layer, necessary for achieving a homogeneous network of datasets for data mining and analytics, is based on four main phases. The first phase comprises the construction of the *Data Catalogue*, where the data is organized into modalities, with the metadata granularity defined along with the blockchain infrastructure, so to accommodate possible performance limitations. Then, the *Data Dictionary* is generated through the normalisation of data contents against reference terminologies (e.g., UMLS/MeSH, WHO-ATC and ICD-10) with transcoding tables and specialized algorithms. For site- or country-specific data, existing mappings are being customized to normalize the new data types. As third step, the development of *Data Stewardship Modules* will provide solutions for provenance tracking and versioning of evolving data sources, and further implement the Research Data Alliance recommendations on data subset identification and citation. Conclusive step will be the development of a *Semantic Querying and Cohort User Interface*, containing endpoints for real-time dataset search. Here, different views or ontologies will be dynamically made possible to adapt to the platform use cases. Particular attention will be devoted to the cohort definition interface, which will allow users to select sets of pertinent data subjects, as data sinks for the later setup of smart contracts. In this deliverable, we describe the development and current status of these phases.

Processing highly heterogeneous data requires the development of a set of data normalization services (Figure 1). In MHMD, each data source owner is responsible for sharing a representative dataset, so that targeted algorithms can be designed to normalize data using standard vocabularies specific to particular domains (e.g. the European epSOS Master Value Catalogue or the Medical Subject Headings (MeSH) for clinical records). Structured metadata is being stored via dedicated endpoints, so to easily find information and send specific data sharing requests through the network. Normalisation is being achieved, as described in this report, through a series of steps, starting from the preparation and sourcing of data, necessary for the application of specialized normalization services. Then, to serve as a representation of datasets available in the network, the generation of a minimum set of metadata for each dataset is provided. The online cataloguing of data, consisting on the publication and indexing of metadata, and the development of a cohort search service will provide

the functionalities for finding existing datasets and facilitate data consultation. As last step, the definition of flexible data sharing pipelines over data exposed through the catalogue will help speeding up data transactions.

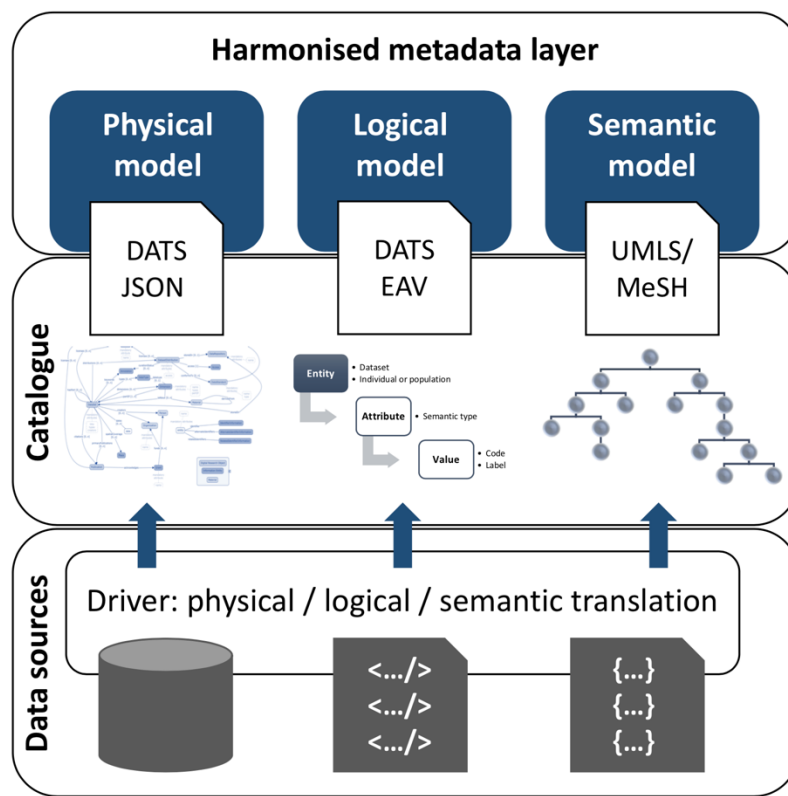


Figure 1 - MHMD metadata ingestion and cataloguing model

3.3 State-of-the-art

Several projects have been implementing solutions to integrate and share individual and patient healthcare data in networks for secondary usage purposes [1–3]. MD-Paedigree [1] integrates and shares highly heterogeneous biomedical information, data, and knowledge to support evidence-based translational medicine at the point of care. It focuses on modelling different paediatric disease to provide better disease understanding and predictive analytics to improve therapy. Similarly, Cardioproof [2] builds on large healthcare datasets to create predictive modelling and simulation tools for cardiology. It uses clinical data to train and validate predictive models to help with early diagnosis, predicting disease behaviour and evolution, and predicting treatment outcomes. Due to the need of big horizontal datasets, these projects cannot afford to re-contact individual patients to request consent and have to use fully anonymised data. On the other hand, EHR4CR [3], which aims to provide a platform for enabling the execution of clinical trials in distributed healthcare networks, follows a different approach where basic queries are run against pseudo-anonymised hospital databases. The main goal of EHR4CR is to provide ways to validate research protocol and then engage identified cohorts into clinical trials.

For research and business purposes, distributed data need to be *findable, accessible, interoperable and reusable* (FAIR principles) [4]. The MHMD architecture leverages on the FedEHR Infostructure [5], an information infrastructure developed for two previous EU-funded FP7 projects, namely MD-Paedigree and Cardioproof, with the FAIR principles in mind. This infrastructure is being extended in the context of MHMD to ingest and semantically integrate additional, non-medical data sources, such as personal well-being data from mobile applications, and provide dataset search homogeneity in a highly-secure and private network. To this aim, the MHMD platform architecture organises the

heterogeneity of data types and components into functional layers and contributors. In the first place, the "Private data sources" layer feeds the system with heterogeneous medical records (such as hospitals and FedEHR datasets) and other individual data (such as, digi.me personal clouds and mobile devices). The "Data harmonisation" layer, developed in WP4, then normalizes these data and provides an integrated search platform to locate datasets from the distributed sources. On top of these layers, the "Privacy middleware" layer includes components for core security and encryption functions, a blockchain ledger and personal data management back-end, and a data profiler for curation and identification of common patterns. Finally, the "Application" layer contains the main functionalities for users, from data analytics to dynamic consent and personal record management.

3.4 Definitions, acronyms and abbreviations

Acronym	Definition
API	Application Program Interface
ATC	Anatomical Therapeutic Chemical
CRS	Care records system
CVI	Cardiovascular imaging
CVS	Comma separated value
DATS	Data Tag Suite
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
i2b2	Informatics for Integrating Biology & the Bedside
ICD	International Disease Classification
JSON	Java Script Object Notation
MeSH	Medical Subject Headings
MHMD	MyHealthMyData
PID	Persistent Identifier
QMUL	Queen Mary University London
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
UMLS	Unified Medical Language System
WP	Work package

3.5 Overview

To define the reference terminologies and the level of published metadata for the harmonisation services, we adopt a mixed bottom-up/top-down approach. In the first, we analysed the source dataset while in the latter we analysed relevant biomedical terminologies. In the next section, we introduce, describe and analyse the sample datasets shared in deliverable WP4.1. In section 5, we introduce the terminological resources used to harmonize these datasets and the dataset data model for the data cataloguing service. Then, in section 6 we describe the data model used to represent datasets shared within the MHMD network and in section 7 we show how the attributes of the datasets described in section 4 can be translated to the terminological resources described in section 5. Then, in section 8 and 9 we describe the work performed to provide track and provenance for the registered datasets and the status of the MHMD data catalogue API, respectively. Finally, in section 10 we conclude this report.

4 Catalogue of MHMD data sources

To define the reference terminologies and the level of published metadata for the harmonization services, we adopt a mixed bottom-up/top-down approach. In the bottom-up phase, data from the partner data sources provided in the data sample sharing repository (deliverable D4.1 'Repository to Share Data Samples') were analysed and catalogued. In this process, we identified protected health identifiers (Yes) attributes, that is, information created or collected that could be linked to a specific individual [6], such as entity identifiers (both semantic, e.g., social security number, and logic, e.g., table identifiers), timestamps, etc., following HIPAA and UK's Data Protection Act guidelines [7,8]. Then, we identified attributes that could be semantically represented using standard biomedical terminologies (e.g., age = 10 --> MeSH:D002648 Child; is_hypertension = True --> MeSH:D006973 Hypertension, etc.).

Two data partners shared representative synthetic data in the MHMD data sample repository: QMUL and digi.me. QMUL datasets were generated based reference ranges from the UK Biobank population cohort [9], and represent inpatient visits. An example record for each of these datasets is provided in the Appendix section "Sample sets". It is important to note that both datasets are fully synthetic, that is, created using a mix of aggregated statistics for each attributed and mock values, as for QMUL, or using purely mock values, as for digi.me datasets. On the other hand, Digi.me sample contains data from individuals, which use the platform to manage and control their digital footprint. The samples include health data but also social media and financial information.

In the next sections, we detail these datasets and classify their attribute into PHI and non-PHI terms.

4.1 QMUL data sample

QMUL dataset contains two files: care records system (CRS), which simulates secondary care visits, and cardiovascular imaging (CVI). In the former, one can find diagnosis lists and patient demographics, linked via fake NHS identifiers to the second file, which contains imaging parameters. Risk factors in the first dataset (CRS) are used to introduce artificial biases in the second (CVI) dataset. These datasets are available in comma separated value (CSV) format, where each line corresponds to an event for a patient.

CRS_IDENTIFIED

This dataset contains care records system information, simulating secondary care visits and include patient demographics, diagnoses, procedure codes, imaging types.

Attribute	Range		Type	Label	PHI
Episode Start Date	27.May.10	18.Sep.09	Date	episode of care start date	Yes
Diagnosis Codes	ICD9:427.31 ICD9:785.2 ICD9:518.83 ICD10:J96.21	ICD10:J96.9 ICD10:J96.20 ICD9:360.42 ICD9:642.91 ICD9:V45.01 ICD10:Z45.01 ICD10:R01 ICD10:F10.188	ICD-9/ICD-10	diagnosis codes	No
Procedure Codes	0SG14A1 0BCG4ZZ 051N0ZY 0D1K4J4 024F0KJ	021V0JU	ICD-10 PCS	procedure codes	No
NHS Number	1759961146	1793786449	Integer	national health service number	Yes
Given Name	Philipa	Ruprecht	String	given name	Yes
Family Name	Cully	Capuano	String	family name	Yes

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

Address Line 1	672 Billingbauk Street	61 Paulyn Avenue	String	address (number, street)	Yes
Address Line 2	West Dunbartonshire	Aylesbury Vale	String	address (town)	Yes
Address Line 3	Scotland	England	String	address (country)	No
Address Line 4	G60 9DB	MK18 9WW	String	address (postcode)	Yes
Postcode	G60 9DB	MK18 9WW	String	postcode	Yes
Date of Birth	30.Aug.76	13.Oct.42	Date	date of birth	Yes
Date of Death	-	19.Dec.82	Date	date of death	Yes
Marital Status	widowed	yes	String	marital status	No
Ethnic Origin	other	bangladeshi	String	ethnic origin	No
Gender	female	male	String	gender	No
Health Authority of Residence	West Dunbartonshire	Aylesbury Vale	String	health authority residence	Yes
Registered GP	Cully,Philipa	Capuano,Ruprecht	String	registered general practitioner	Yes
Registered GP Practice	Park GP Practise	Cucumber GP Practise	String	registered general practitioner practice	Yes
index	0	382205	Integer	database index	Yes
Imaging types	Cardiac MRI Echocardiography	Echocardiography	Code (local)	imaging types	No

CVI_IDENTIFIED

This dataset contains cardiovascular imaging information, describing cardiac structure and function parameters with symbolic correlations with risk factors.

Attribute	Range		Type	Label	PHI
Address	45 George Julius Rd,Wokingham, England, RG11 6MT	771 Sather Rd,Torbay,England, TQ4 0ET	String	address	Yes
Phone	555778197	555202699	Integer	phone number	Yes
Last Name	Schaner	Graw	String	last name	Yes
First Name	So	Shari	String	first name	Yes
Middle Name			String	middle name	Yes
Medical Record Number	2159268637	2463714947	Integer	medical record number	Yes
Prior Names			String	prior names	Yes
Ethnicity	white	other asian	Code (local)	ethnicity	No
Gender	female	female	Code (local)	gender	No
Birthdate	19.Dec.35	01.Feb.60	Date	birthdate	Yes
Study Date	05.Jul.08	18.May.08	Date	study date	Yes
Patient Age	72	48	Integer	patient age	No
National Identifier	1462919379	1762234735	Integer	national identifier	Yes
Blood Pressure	123/101	121/101		blood pressure	No
Heart rate	90	85	Integer	heart rate	No

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

Height (cm)	146.841259	143.114784	Float	height	No
Weight (kg)	61.9437057	58.1564489	Float	weight	No
Scanner Manufacturer	Siemens	Siemens	String	scanner manufacturer	No
LVEDV (ml)	118.362334	136.079787	Decimal	left ventricle end-diastolic volume	No
LVESV (ml)	37.6978334	38.0472194	Decimal	left ventricle end-systolic volume	No
LVSV (ml)	80.6645007	98.0325677	Decimal	left ventricle stroke volume	No
LVEF (%)	68.1504816	72.0405064	Decimal	left ventricle ejection fraction	No
LV Mass (g)	76.0802965	127.628197	Decimal	left ventricle mass	No
RVEDV (ml)	159.864557	155.537916	Decimal	right ventricle end-diastolic volume	No
RVESV (ml)	60.3277696	70.0258922	Decimal	right ventricle end-systolic volume	No
RVSV (ml)	99.5367879	85.5120234	Decimal	right ventricle stroke volume	No
RVEF (%)	62.2631992	54.9782496	Decimal	right ventricle ejection fraction	No
RV Mass (g)	31.6330351	27.3067596	Decimal	right ventricle mass	No
BMI (kg/msq)	28.7277083	28.3941504	Decimal	body mass index	No
BSA	1.58953974	1.52051231	Decimal	body surface area	No
BSA (msq)	1.58953974	1.52051231	Decimal	body surface area (msq)	No
CO (L/min)	7.25980506	8.33276825	Decimal		No
Central PP(mmHg)	12.9655991	12.6649743	Decimal		No
DBP (mmHg)	101	101	Integer	diastolic blood pressure	No
Diabetic RF	non-diabetic	non-diabetic	Boolean	diabetic reference	No
Hypertension RF	non-hypertensive	non-hypertensive	Boolean	hypertension reference	No
LVEF (ratio)	0.68150482	0.72040506	Decimal	left ventricle ejection fraction (ratio)	No
MAP	110.709333	109.706667	Decimal		No
PAP (mmHg)	11.1965179	11.9196549	Decimal		No
PP (mmHg)	22	20	Integer		No
RVEF (ratio)	0.62263199	0.5497825	Decimal	right ventricle ejection fraction (ratio)	No
SBP (mmHg)	123	121	Integer		No
SVR (mmHg/L/min)	15.2496289	13.1656928	Decimal		No

Smoking RF	smoker	smoker	Boolean	smoking reference	No
Vascular RF	atherosclerotic	non-atherosclerotic	Boolean	vascular reference	No
risk_name	intermediate	intermediate	Code (local)	risk name	Yes

4.2 digi.me data samples

In this section, we describe digi.me data samples. Digi.me data samples are available in the JSON format, in a proprietary model designed by the company. Individual records are connected using a specific attribute, entityid, which is common for all datasets. Each record in a JSON object corresponds to an event for an individual. We start our description by the health samples, which have 230 records stored in the sharing repository. Digi.me health data covers several health care events in the patient's pathway through the health service, such as admission, diagnosis, prescription and medication. Then, we list the social media and financial sample sets, which have only one record each per dataset.

4.2.1 Health data

ADMISSIONS

Provide individual health data related to health care admissions.

Attribute	Range		Type	Label	PHI
createddate	1178729040000	1363170420000	Integer (TSE ¹)	admission date	Yes
dischargedate	1192669080000	1450458780000	Integer (TSE)	discharge date	Yes
organization	Heilbrigðisstofnun	Saga_skema - TMS	String	organization	No
responsiblephysician	3Sturluson	Karl Sigurðsson	String	responsible physician	Yes
servicegroup	Almennar lyflækningar(T)	REL Barnaskurðeild	String	service group	No
spectreatmentid	4290	165812	Integer	treatment identifier	Yes
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

ALLERGY

Provide individual health data related to allergies.

Attribute	Range		Type	Label	PHI
atc	A01AB04	QG51AG01	ATC	WHO-ATC code	No
classification	Bráðaofnæmi	Ofnæmi	String	drug classification	No
comment			String	comment	No
component	123456879012345687901234568790123456879	12345687901234568790123456879012345687901	Integer	component	Yes

¹ TSE --> time since the epoch (ms)

createdate	1319735847000	1449761521000	Integer (TSE)	creation date	Yes
type	Lyfjafnæmi		String	drug type	No
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

ARRIVAL_AMBULATORY

Provide individual health data related to arrivals at ambulatory care.

Attribute	Range		Type	Label	PHI
admissionstartdate	1128436560000	1412938380000	Integer (TSE)	admission start date	Yes
createddate	1135696466000	1414396800000	Integer (TSE)	arrival date	Yes
arrivalid	1001	84445	Integer	arrival identifier	Yes
responsiblephysician	Anna Mýrdal Helgadóttir	Valdi	String	responsible physician	Yes
responsiblephysicianid	1987	601603	Integer	responsible physician identifier	Yes
responsiblephysicianmdno	221	6663	Integer	responsible physician MdNo	Yes
servicegroup	Almennar lyflækningar(T)	Slysa- og bráðalækningar(H)	String	service group	No
servicegrouparrivaldischarge	Almennar lyflækningar(T) 10.10.2014	Slysa- og bráðalækningar(H) 27.01.2010	String	service group arrival discharge	No
treatmentid	719	281336	Integer	treatment identifier	Yes
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

ARRIVAL_EMERGENCY

Provide individual health data related to arrivals at emergency care.

Attribute	Range		Type	Label	PHI
createddate	1135696466000	1414396800000	Integer (TSE)	arrival date	Yes
arriveid	30729	98501	Integer	arrival identifier	Yes
organizationname	Saga_skema - TMS	84445	String	organization name	No
responsiblephysician	Anna Mýrdal Helgadóttir	Valdi	String	responsible physician	Yes
spectreatmentid	719	281336	Integer	treatment identifier	Yes
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

ARRIVAL_PRIMARY_HEALTH

Provide individual health data related to arrivals at primary care services.

Attribute	Range		Type	Label	PHI
arriveid	67877	98501	Integer	arrival identifier	Yes
arrivetime	1408113034000	1452850200000	Integer (TSE)	arrival time	Yes
contactname	Fjölskylduviðtal	Matur	String	contact name	Yes
departmentorgname	Saga_skema - TMS		String	department name	No
resourcename	Bárður Sigurgeirsson	Gunnar A. Baarregaard	String	resource name	Yes
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

DIAGNOSIS

Provide individual diagnosis health data.

Attribute	Range		Type	Label	PHI
code	A00	Z34	ICD-10	diagnosis code	No
codingsystem	ICD-10		String	terminology	No
firstregistration	1122336000000	1472515200000	Integer (TSE)	first registration date	Yes
islongterm	TRUE		Boolean	is long term	No
lastregistration	1122336000000	1483056000000	Integer (TSE)	last registration date	Yes
name	ÁKOMIN AFLÖGUN Á MJADMAGRIN D	VIROSIS	String	diagnosis label	No
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

MEDICATION

Provide individual medication consumption (action) data.

Attribute	Range		Type	Label	PHI
atccode	A02BC05	S01EE01	ATC	WHO-ATC code	No
autoseponate			?		
conceptcode	2348	1115876	?	internal concept code	No
createddate	1392116168000	1472552301000	Integer (TSE)	confirmation date	Yes
daysleft	Lokið		String	days left	No
directions	1 2 sinnum á dag	3 ml þrisvar á dag	String	medication direction	No
form	augndr	töflur	?	medication form	No

lastchanged	1392076800000	1472515200000	Integer (TSE)	last changed date	Yes
lastprescribed	1392076800000	1472515200000	Integer (TSE)	last prescribed date	Yes
name	ABILIFY	ZITROMAX	String	medication name	No
nrnorr	2033	497289	?		No
numberofpackings	1	3	Integer	number of packages	No
numeroftimes	1 sinni	4 sinnum á 80 daga fresti	String	number of times	No
onetimeonly	TRUE		Boolean	one time only (single dose)	No
prescriptionends	1395705600000	1477612800000	Integer (TSE)	date of prescription ending	Yes
quantity	1	500	Integer	medication quantity per dose	No
skammtaaskja	0		Integer	closed box	No
strength	18 mcg/hylk	750 mg	String	medication strength	No
totalquantity	1	500	Integer	medication quantity (total)	No
usedfor	augnsjúkdómi	verk	String	used for (condition)	No
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

PRESCRIBED_ITEMS

Provide individual prescribed items (instruction) data associated to a prescription.

Attribute	Range		Type	Label	PHI
atccode	A01AB04	QG51AG01	ATC	WHO-ATC code	No
dosageinstructions	1 2 sinnum á dag	3 ml þrisvar á dag	String	prescribed direction	No
form	augndr	töflur	?	prescribed form	No
ismonitored	false	true	Boolean	is monitored	No
itemno	0	100000	Integer	item number	Yes
name	ABILIFY	ZITROMAX.	String	drug name	No
createdate	1319735847000	1449761521000	Integer (TSE)	creation date	Yes
numberofpackages	1	3	Integer	number of packages	No
prescriptionid	16_3112754029_103_273167	20_4250247387_425_2345027	String	prescription identifier	Yes
createddate	1395705600000	1477612800000	Integer (TSE)	date of prescription creation	Yes
productid	116584	1477612	Integer (TSE)	product identifier	Yes

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

quantity	1	500	Integer	medication quantity per dose	No
strength	18 mcg/hylk	750 mg	String	prescription strength	No
unit	stk	mg	String	prescription unit	No
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

PRESCRIPTIONS

Provide individual prescribed (instruction) data.

Attribute	Range		Type	Label	PHI
daysbetweendispen sations	22	80	Integer	days between dispensation	No
earliestdispensatio ndate	1389830400000	1472515200000	Integer (TSE)	earliest dispensation date	Yes
id	266268	279038	Integer	identifier	Yes
iscanceled	TRUE		Boolean	is cancelled	No
createddate	1389880260000	1472552280000	Integer (TSE)	created date	Yes
latestdispensationd ate	1421280000000	1504051200000	Integer (TSE)	latest dispensation date	Yes
prescribeddispensa tions	1	4	Integer	prescribed dispensations	No
prescribeditems	16_3112754029_109_N05AH03	36_320239_302_MA2349A	String	prescribed items	Yes
prescribercontactin fo	543 1000	545-3300	String	prescriber contact information	Yes
prescriberid	195	9999	Integer	prescriber identifier	Yes
prescribername	Ásta Rún Ásgeirsdóttir	Sveinn Magnússon	String	prescriber name	Yes
prescriptiontype				prescription type	No
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

VACCINATIONS

Provide individual vaccination data.

Attribute	Range		Type	Label	PHI
code	J07AE02	J07BM02	ATC	code	No
codename	Bcg-imovax® berklar / tuberculosis	Vaxigrip® influenza	String	code name	No
codes	Barnaveiki	Stífkrampi	String	codes	No

codingsystem	ATC		String	coding system	No
createddate	1183420800000	1483056000000	Integer (TSE)	created date	Yes
senderdescription	Saga 3.1.39 release	Vantar upplýsingar um sendanda	String	sender description	Yes
sendergateway	SAGA_3139A	SAGA_RELEASE	String	sender gateway	Yes
senderid	RA11-3100-6101	RAxx0761	String	sender identifier	Yes
sendersystem	SAGA.NET		String	sender system	Yes
entityid	16_3112754029_100_1110553080000	20_898379452937_100_24520712341790	String	entity identifier	Yes

4.2.2 Social media data

This dataset contains 3 subsets - media, post and comment - which contain data about interactions of individuals in social networks, such as Facebook and Twitter.

MEDIA

This dataset describes media resources, such as images, posted by individual in their social media.

Attribute	Sample	Type	Label	PHI
baseid	4_1542271900260495153_182042	String	base identifier	Yes
cameramodelentityid		String	camera model entity identifier	Yes
commentcount	0	Integer	comment count	No
commententityid		String	comment entity identifier	Yes
createddate	1490832000000	Integer	created date	Yes
description		String	description	No
displayshorturl		Boolean	display short URL	No
displayurlindexend	0	Integer	display URL index end	No
displayurlindexstart	0	Integer	display URL index start	No
entityid	4_1542271900260495153_182042	String	entity identifier	Yes
filter	Gingham	String	filter	No
interestscore	0	Integer	interest score	No
itemlicenceentityid		String	item license entity identifier	Yes
latitude	0	Integer	latitude	Yes
likecount	14	Integer	like count	No
link	https://www.instagram.com/p/BVnQB86F-8xIEERyS_PisN_g2p5-V-o3aUBd7A0/	URL	link	Yes
locationentityid		String	location entity identifier	Yes
longitude	0	Integer	longitude	Yes
mediaAlbumEntityID	4_182042_182042_1	String	media album entity identifier	Yes
mediaalbumname		String	media album name	No
mediaid	1542271900260495153_182042	String	media identifier	Yes
mediaobjectid		String	media object identifier	Yes
mediaobjectlikeid		String	media object like identifier	Yes
name	Team's getting smaller by 1. Good luck @mheap	String	name	Yes

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

originatortype	2	Integer	originator type	No
personentityid	4_182042_182042	String	person entity identifier	Yes
personfilerelativepath		String	person file relative path	Yes
personfileurl	https://scontent.cdninstagram.com/t51.2885-19/11356960_1625044604420020_264796266_a.jpg	URL	person file URL	Yes
personfullname	Pascal Wheeler	String	person full name	Yes
personusername	pascalw	String	person user name	Yes
postentityid	4_182042_1542271900260495153_182042	String	post entity identifier	Yes
resources.aspectratio.accuracy	100	Integer	resources aspect ratio accuracy	No
resources.aspectratio.actual	1:1	Ratio	resources aspect ratio actual	No
resources.aspectratio.closest	1:1	Ratio	resources aspect ratio closest	No
resources.height	150	Integer	resources height	No
resources.mimetype	image/jpeg	String	resources mime type	No
resources.type	0	Integer	resources type	No
resources.url	https://scontent.cdninstagram.com/t51.2885-15/s150x150/e35/c135.0.809.809/19367043_122168111710592_7572307112722694144_n.jpg	URL	resources URL	Yes
resources.width	150	Integer	resources width	No
tagcount	0	Integer	tag count	No
taggedpeoplecount	0	Integer	tagged people count	No
type	0	Integer	type	No
updateddate	1490832000000	Integer	updated date	Yes

POST

This dataset contains posts, including the interaction, such as like counts, of individuals in their social networks.

Attribute	Sample	Type	Label	PHI
annotation		String	annotation	No
baseid	4_182042_1542271900260495153_182042	String	base identifier	Yes
cameramodelentityid			camera model entity identifier	Yes
commentcount	0	Integer	comment count	No
commententityid			comment entity identifier	Yes
createddate	1490832000000	Integer (TSE)	created date	Yes
description			description	No
entityid	4_182042_1542271900260495153_182042	String	entity identifier	Yes
favouritecount	0	Integer	favourite count	No
iscommentable	1	Integer	is commentable	No
isfavourited	0	Integer	is favourited	No
islikeable	1	Integer	is likeable	No
islikes	0	Integer	is likes	No

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
---------------------------------	------------------------------

isshared	0	Integer	is shared	No
istruncated	0	Integer	is truncated	No
latitude	0	Decimal	latitude	Yes
likecount	14	Integer	like count	No
longitude	0	Integer	longitude	Yes
originalcrosspostid	0	Integer	original cross post identifier	Yes
originalpostid	0	Integer	original post identifier	Yes
originalposturl		URL	original post URL	Yes
personentityid	4_182042_182042	String	person entity identifier	Yes
personfilerelativepath			person file relative path	Yes
personfileurl	https://scontent.cdninstagram.com/t51.2885-19/11356960_1625044604420020_264796266_a.jpg	URL	person file URL	Yes
personfullname	Pascal Wheeler	String	person full name	Yes
personusername	pascalw	String	person user name	Yes
postentityid			post entity identifier	Yes
postid	1542271900260495153_182042	String	post identifier	Yes
postreplycount	0	Integer	post reply count	No
posturl	https://www.instagram.com/p/BVnQB86F-8xIEERyS_PisN_g2p5-V-o3aUBd7A0/	URL	post URL	Yes
rawtext		String	raw text	No
referenceentityid	4_182042	String	reference entity identifier	Yes
referenceentitytype	15	Integer	reference entity type	No
sharecount	0	Integer	share count	No
socialnetworkuserentityid	4_182042	String	social network user entity identifier	Yes
source			source	No
text	Team's getting smaller by 1. Good luck @mheap	String	text	Yes
title			title	No
type	20	Integer	type	No
updateddate	1490832000000	Integer (TSE)	updated date	Yes
visibility			visibility	No

COMMENT

This dataset contains comments of individuals in their social networks.

Attribute	Sample	Type	Label	PHI
appid			app identifier	Yes
baseid	4_182042_17858982733161786	String	base identifier	Yes
commentcount	0		comment count	No
commentid	1490832000000	Integer	comment identifier	Yes
commentreplyid			comment reply identifier	Yes
createddate	1490832000000	Integer (TSE)	created date	Yes
entityid	4_182042_17858982733161786	String	entity identifier	Yes
likecount	0	Integer	like count	
link			link	Yes

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

metaid			meta identifier	Yes
personentityid	4_182042_1545946064	String	person entity identifier	Yes
personfilerelativepath			person file relative path	Yes
personfileurl	https://scontent.cdninstagram.com/t51.2885-19/s150x150/19051047_1129050077200786_761246497533591552_a.jpg	URL	person file URL	Yes
personfullname	Sarah Lamb	String	person full name	Yes
personusername	iamgirlygeekdom	String	person user name	Yes
privacy	0	Integer	privacy	No
referenceentityid	4_1536332454209789316_182042	String	reference entity identifier	Yes
referenceentitytype	1	Integer	reference entity type	No
socialnetworkuserentityid	4_182042	String	social network user entity identifier	Yes
text	Oooh I missed out! Mint magnums are my favourite!	String	text	No
updateddate	1490832000000	Integer (TSE)	updated date	Yes

4.2.3 Financial data

This dataset contains data of financial transactions of individuals.

TRANSACTION

This dataset describes financial transaction of digi.me users.

Attribute	Sample	Type	Label	PHI
accountentityid	17_10019612	String	account entity identifier	Yes
amount	2.43	Decimal	amount	No
basetype	DEBIT	String	base type	No
category	Restaurants	String	category	No
categoryid	22	Integer	category identifier	Yes
categorysource	SYSTEM	String	category source	No
categorytype	EXPENSE	String	category type	No
checknumber			check number	No
consumerref			consumer ref	Yes
createddate	1490832000000	Integer (TSE)	created date	Yes
currency	GBP	String	currency	No
entityid	17_10019612_10945571	String	entity identifier	Yes
highlevelcategoryid	10000011	Integer	high level category identifier	Yes
id	10945571	Integer	identifier	Yes
ismanual	false	Boolean	is manual	No
merchantaddress1		String	merchantaddress1	Yes
merchantaddress2		String	merchantaddress2	Yes
merchantcity		String	merchant city	No
merchantcountry		String	merchant country	No
merchantid	costacoffee	String	merchant identifier	Yes
merchantname	Costa Coffee	String	merchant name	No
merchantstate		String	merchant state	No
merchantzip		String	merchant zip	Yes

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

originalref	Debit COSTA COFFEE ON 26 MAR CPM	String	original ref	Yes
postdate	1490832000000	Integer (TSE)	postdate	Yes
runningbalance	113.55	Decimal	running balance	No
runningbalancecurrency	GBP	String	running balance currency	No
simpleref	Costa Coffee	String	simple ref	No
status	POSTED	String	status	No
subtype	PURCHASE	String	subtype	No
transactiondate	0	Integer	transaction date	Yes
type	PURCHASE	String	type	No

5 Analysis of terminologies and ontologies for representing MHMD datasets and construction of MHMD data dictionary

To define the reference terminologies for the harmonization services, in the top-down phase performed we analysed some standard biomedical terminologies and ontologies as candidate resources to model and harmonize the MHMD datasets presented previously and create a dictionary of the data source attributes. These resources were selected due to their relevance to the health and life sciences community, to MHMD according to the use-cases presented in deliverable D1.1, and to the existing expertise within the project. In the next sections, we present a brief description of the analysed terminologies and ontologies (a more detailed overview can be found elsewhere [10–13]). Then, we describe our process to select the best and more appropriate resource to harmonize MHMD datasets in the metadata catalogue.

5.1 Resources

5.1.1 Anatomical Therapeutic Chemical (ATC)

ATC [10] is a drug classification system developed and maintained by the World Health Organization (WHO). In the healthcare domain, it is the drug classification system most used internationally. ATC classifies drug's active substances into five groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. The classification system favours international non-proprietary drug names and is available only in the English language. Translation to other languages is available from the international non-proprietary names' catalogue. As MHMD shall cope with datasets from different geographical regions, ATC is a good candidate for representing drug concepts. However, as expected, this terminology cannot cover the whole spectrum of data available in MHMD.

5.1.2 International Classification of Diseases (ICD)

ICD [11] is an international terminology that organizes and codes diagnostic health information for statistics and epidemiology, healthcare management, monitoring and evaluation, research, primary care, prevention and treatment. Like ATC, ICD is sponsored and managed by the WHO and has been adopted by the WHO Member States since 1967. Currently, it is in the operational version 10 (ICD-10) and is accessible in 42 languages, including Arabic, Chinese, English, French, Russian and Spanish. ICD version 11 is under development in a collaborative and interactive process between health and taxonomy experts, and users. In addition to diagnosis, ICD has an extension, ICD Procedure Code System (ICP PCS), which provides concepts for representing procedural events. Similar to ATC, ICD is a ubiquitously employed in health care settings to code diseases, and covers only few of MHMD dataset domains, i.e., diseases and procedures.

5.1.3 Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)

SNOMED CT [12] is a standard ontology for healthcare. SNOMED CT terms are available originally in the English language and the ontology is processable by machines. It provides the most comprehensive clinical vocabulary available in any language, covering large part of the medical vocabulary such as diseases, symptoms, operations, treatments, devices and drugs. Currently, it contains more than 300k active concepts. The intellectual property of SNOMED CT belongs to the International Health Terminology Standards Development Organisation, which owns and administers the terminology. SNOMED CT is one of the most adopted ontologies internationally, having its initial vocabulary translated into Spanish, French and a few other languages. Based on the licensing agreements with countries, the terminology is being translated into different languages from English. However, the ontology is complex to express some concepts, which might need sometimes a specific

SNOMED CT syntax, and given SNOMED CT's level of concept specification, it is hard to navigate and locate concepts in the hierarchy.

5.1.4 Medical Subject Headings (MeSH)

MeSH [13] is the National Library of Medicine curated medical vocabulary resource. The main purpose of MeSH is to provide a hierarchically-organized terminology for indexing and cataloguing of biomedical information such as MEDLINE/PubMed and other NLM databases. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels of the thirteen-level hierarchy, such as "Ankle" and "Conduct Disorder". There are over 28,000 descriptors in MeSH with over 90,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid". In addition to these headings, there are more than 240,000 supplementary concept records within a separate file. Generally, supplementary concept records contain specific examples of chemicals, diseases, and drug protocols. They are updated more frequently than descriptors. Each supplementary concept record is assigned to a related descriptor via the heading map field. The heading map is used to rapidly identify the most specific descriptor class and include it in the citation. While SNOMED CT is designed to represent actual clinical concepts in a healthcare setting, MeSH focuses on cataloguing biomedical information. Thus, despite a good coverage of concepts in several themes, it has a simpler representation when compared to SNOMED CT.

The MeSH ontology is provided at <https://www.ncbi.nlm.nih.gov/mesh>.

5.1.5 digi.me ontology

The digi.me Medical Object Baseline manages the complexity of medical information and, in testing experiments with small scale sample sets, has proved suitable to support the initial mappings of a number of source specific medical standards. The ontology is still under development and is expected to expand to cover off additional elements and properties as required (see Appendix "digi.me ontology" for detailed view). A total of 6 different medical data standards have been mapped to the Medical Ontology Baseline in the digi.me ontology: Icelandic Health (existing); HL7 CDA; C-CDA R2.1; Blue Button Clear; Blue Button VA; and Epic. In any particular area of interest, such as health, finance, or social networks, the normalized ontology is typically based on the most mature, well-defined and expressive data format. For example, the digi.me Social Network normalized ontology is based on the Facebook structure and set of data attributes. Facebook was selected as a baseline because it provides a mature and comprehensive set of relevant attributes.

5.2 Ontology assessment

5.2.1 Tod down approach

To decide which resource to use to represent MHMD datasets, we analysed the terminologies and ontologies presented previously and compared them against four dimensions: comprehensiveness, generality, complexity and availability of annotated resources. The first dimension, comprehensiveness, measures how complete the resource is describing its domain, i.e., the expressivity of the ontology language to enable representing the complexities of the domain as comprehensive as possible [14]. Second, generality measures how the resource can generalize in terms of domain coverage, i.e., how broad is the coverage of the terminology. Despite MHMD datasets having a major focus on healthcare, they are also expected to originate from non-clinical/medical domain. Thus, it is important that the resource is able to cover non-clinical/healthcare concepts, such as devices, sensors, etc. The third dimension, complexity, measures the complexity of the resource, i.e., how easy is for data sources to find concepts and map them into the ontology. Finally, the availability of annotated resources dimension measures the amount of existing annotated

resources using the ontology. For example, PubMed abstracts are annotated using MeSH terms, and thus a very large annotated set is available for MeSH. Similarly, digi.me ontology is mapped to several other terminologies, such as HL7 CDA, Blue Button and Epic model. As manually mapping all existing data from the data sources to a common global ontology is an extensive work, the existence of annotated examples is crucial to guarantee the sustainability of the mapping process via (semi-)automated mapping algorithms.

In Table 1, we provide a score for each of these dimensions for the resources described above. Using these dimensionalities, we can calculate the suitability score of a resource to represent the MHMD catalogue metadata according to the formula:

$$\text{suitability} = \text{comprehensiveness} + \text{generality} - \text{complexity} + \text{annotation}$$

As we can see from Table 1, from a high-level perspective, MeSH provides a good suitability score for representing MHMD datasets. Of course, other dimensions could be used in the assessment methodology and the score of the attributes is only an estimate. Nevertheless, it already provides a reasonable direction for knowledge representation, in particular, as comparing these ontological/terminological resources is out of the scope of this work [14].

Table 1 - Evaluation of knowledge representation resources

Resource	Comprehensiveness	Generality	Complexity	Annotation	Suitability score
ATC	+++	+	+	+	4
ICD	+++	+	+	++	5
SNOMED CT	+++	+++	+++	++	5
MeSH	++	+++	++	+++	6
digi.me	+	+++	++	++	4

5.2.2 Bottom-up ontology approach

To complement the previous semantic resource analyses, we performed a bottom-up analyses based on the description of the datasets provided in section 4. Using the label and category attributes of the sample sets, we can group the attributes into 4 main semantic classes (as showed in the next section):

- Demographics (e.g., gender, age, weight)
- Administrative (e.g., organization, department, service)
- Clinical concepts (e.g., diagnosis, medication and procedure)
- Miscellaneous concepts (e.g., quantity, devices)

From the terminologies and ontologies assessed in the previous section, SNOMED CT, MeSH and digi.me are able to represent these categories. As MeSH has lower complexity compared to SNOMED CT, allowing data sources to more easily map their concepts, and higher level of annotated resources (e.g., MEDLINE/PubMed collection), compared to both digi.me and SNOMED CT, allowing (semi-)automatic tools to be implemented (see section 7), we take it as the reference ontology to represent MHMD concepts in the cataloguing service.

5.3 Data dictionary for the MHMD sample sets

In this section, we provide the data dictionary for the datasets presented in section 4 using MeSH concepts. Digi.me's social media and financial data were not included in the dictionary due to their low number of samples available.

After the analysis of semantic categories for the sample datasets, we have the following metadata modalities:

- **Cardiovascular Diseases; C14**
 - Hypertension; C14.907.489
- **Nutritional and Metabolic Diseases; C18**
 - Diabetes Mellitus; C18.452.394.750
- **Inorganic Chemicals; D01**
- **Organic Chemicals; D02**
- **Heterocyclic Compounds; D03**
- **Polycyclic Compounds; D04**
- **Pharmaceutical Preparations; D26**
 - Dosage Forms; D26.255
- **Diagnosis; E01**
 - Diagnostic Techniques and Procedures; E01.370
 - Body Height; E01.370.600.115.100.160.100
 - Body Surface Area; E01.370.600.115.100.231
 - Body Weight; E01.370.600.115.100.160.120
 - Blood Pressure; E01.370.600.875.249
 - Diagnostic Imaging; E01.370.350
- **Therapeutics; E02**
 - Long-Term Care; E02.760.476
- **Anesthesia and Analgesia; E03**
- **Surgical Procedures, Operative; E04**
- **Investigative Techniques; E05**
 - Risk Assessment; E05.318.740.600.800.715
- **Dentistry; E06**
- **Equipment and Supplies; E07**
- **Circulatory and Respiratory Physiological Phenomena; G09**
 - Heart Rate; G09.330.380.500
 - Cardiac Volume; G09.330.380.249
 - Stroke Volume; G09.330.380.124.882
- **Persons; M01**
 - Age Groups; M01.060
 - Population Groups; M01.686
 - Smokers; M01.808
- **Population Characteristics; N01**
 - Marital Status; N01.824.308.500
- **Health Care Facilities, Manpower, and Services; N02**
 - Hospital Departments; N02.278.216.500.968
 - Hospital Administration; N02.278.216.500
 - Hospital Units; N02.278.388
 - Prescriptions; N02.421.668.778
- **Health Care Economics and Organizations; N03**
- **Environment and Public Health; N06**
 - Body Mass Index; N06.850.505.200.100.175
- **Publication Formats; V02**
 - Terminology; V02.310.750
- **Geographic Locations; Z01**

Each modality is associated to its respective MeSH tree code, which uniquely identifies a concept in the MeSH hierarchy. In the following tables, we provide the mapping between the MeSH codes and the attributes in the source datasets grouped by high data categories (demographics, administrative, clinical and miscellaneous).

DEMOGRAPHICS CONCEPTS

Dataset	Attribute label	Semantic category	MeSH concept
CRS_IDENTIFIED	address (country)	demographics	Geographic Locations;Z01
CRS_IDENTIFIED	marital status	demographics	Marital Status;N01.824.308.500
CRS_IDENTIFIED	ethnic origin	demographics	Population Groups;M01.686
CRS_IDENTIFIED	gender	demographics	Persons;M01
CVI_IDENTIFIED	ethnicity	demographics	Population Groups;M01.686
CVI_IDENTIFIED	gender	demographics	Population Groups;M01.686
CVI_IDENTIFIED	patient age	demographics	Age Groups;M01.060
CVI_IDENTIFIED	height	demographics	Body Height;E01.370.600.115.100.160.100
CVI_IDENTIFIED	weight	demographics	Body Weight;E01.370.600.115.100.160.120

ADMINISTRATIVE CONCEPTS

Dataset	Attribute label	Semantic category	MeSH concept
ARRIVAL_PRIMARY_HEALTH	department name	location	Hospital Departments;N02.278.216.500.968
ARRIVAL_EMERGENCY	organization name	location	Health Care Economics and Organizations;N03
ARRIVAL_AMBULATORY	service group	location	Hospital Administration;N02.278.216.500 Hospital Units;N02.278.388
ARRIVAL_AMBULATORY	service group arrival discharge	location	Hospital Administration;N02.278.216.500 Hospital Units;N02.278.388
ADMISSIONS	organization	location	Health Care Economics and Organizations;N03
ADMISSIONS	service group	location	Hospital Administration;N02.278.216.500 Hospital Units;N02.278.388

CLINICAL CONCEPTS

Dataset	Attribute label	Semantic category	MeSH concept
DIAGNOSIS	diagnosis code	diagnosis	Diagnosis;E01
DIAGNOSIS	terminology	diagnosis	Terminology;V02.310.750
DIAGNOSIS	is long term	diagnosis	Long-Term Care;E02.760.476
DIAGNOSIS	diagnosis label	diagnosis	Diagnosis;E01
DIAGNOSIS	used for (condition)	diagnosis	Diagnosis;E01
CRS_IDENTIFIED	diagnosis codes	diagnosis	Diagnosis;E01
PRESCRIPTIONS	prescription type	drug	Prescriptions;N02.421.668.778
PRESCRIBED_ITEMS	WHO-ATC code	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
PRESCRIBED_ITEMS	prescribed form	drug	Dosage Forms;D26.255
PRESCRIBED_ITEMS	drug name	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

VACCINATIONS	code	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
VACCINATIONS	code name	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
VACCINATIONS	codes	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
VACCINATIONS	coding system	drug	Terminology;V02.310.750
MEDICATION	WHO-ATC code	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
MEDICATION	internal concept code	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
MEDICATION	medication form	drug	Dosage Forms;D26.255
MEDICATION	medication name	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
ALLERGY	WHO-ATC code	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
ALLERGY	drug classification	drug	Terminology;V02.310.750
ALLERGY	drug type	drug	Inorganic Chemicals;D01 Organic Chemicals;D02 Heterocyclic Compounds;D03 Polycyclic Compounds;D04
CRS_IDENTIFIED	procedure codes	procedure	Therapeutics;E02 Anesthesia and Analgesia;E03 Surgical Procedures, Operative;E04 Investigative Techniques;E05 Dentistry;E06 Equipment and Supplies;E07

MISCELLANEOUS CONCEPTS

Dataset	Attribute label	Semantic category	MeSH concept
CVI_IDENTIFIE	blood pressure	quantity	Blood Pressure;E01.370.600.875.249
CVI_IDENTIFIE	heart rate	quantity	Heart Rate;G09.330.380.500
CVI_IDENTIFIE	left ventricle end-diastolic volume	quantity	Cardiac Volume;G09.330.380.249
CVI_IDENTIFIE	left ventricle end-systolic volume	quantity	Cardiac Volume;G09.330.380.249
CVI_IDENTIFIED	left ventricle stroke volume	quantity	Stroke Volume;G09.330.380.124.882

D4.2 MHMD Ontological Resources	MHMD-H2020-ICT-2016 (732907)
--	-------------------------------------

CVI_IDENTIFIED	right ventricle end-diastolic volume	quantity	Cardiac Volume;G09.330.380.249
CVI_IDENTIFIED	right ventricle end-systolic volume	quantity	Cardiac Volume;G09.330.380.249
CVI_IDENTIFIED	right ventricle stroke volume	quantity	Stroke Volume;G09.330.380.124.882
CVI_IDENTIFIED	body mass index	quantity	Body Mass Index;N06.850.505.200.100.175 Body Weight;E01.370.600.115.100.160.120
CVI_IDENTIFIED	body surface area	quantity	Body Surface Area;E01.370.600.115.100.231
CVI_IDENTIFIED	body surface area (msq)	quantity	Body Surface Area;E01.370.600.115.100.231
CVI_IDENTIFIED	diastolic blood pressure	quantity	Blood Pressure;E01.370.600.875.249
CVI_IDENTIFIED	hypertension reference	boolean	Hypertension;C14.907.489
CVI_IDENTIFIED	diabetic reference	boolean	Diabetes Mellitus;C18.452.394.750
CVI_IDENTIFIED	smoking reference	boolean	Smokers;M01.808
CVI_IDENTIFIED	risk name	code (local)	Risk Assessment;E05.318.740.600.800.715
CRS_IDENTIFIED	imaging types	code (local)	Diagnostic Imaging;E01.370.350

6 Unified data catalogue metadata model

To allow the registration, cataloguing, search and discovery of MHMD datasets, we assessed existing metadata models that could be used to harmonize disparate formats and data models found in biomedical datasets. We identified 4 main requirement principles that this model should respect in order to achieve our objectives:

- Generality: the model shall be able to generalize to different types of datasets, allowing representation of disparate data sources, such as EHR, sensors and social media;
- Expressiveness: the model shall allow comprehensive expression of datasets so that they can be findable;
- Complexity: the model must not be complex so that they data sources can readily integrated into the network architecture; and
- Flexibility: the model must be flexible so that it can adapt to specific needs of datasets as new data sources join the MHMD network.

6.1 Existing metadata models for biomedicine

In this section, we analyse i2b2, Bioschemas and DATS (Data Tag Suite) models for representing biomedical data. While i2b2 is a robust and proved model, Bioschemas and DATS are emerging technologies, with novel data management concepts in mind, such as the FAIR principles.

6.1.1 i2b2 data mart model

The i2b2 data mart follows a star schema structure to model its data based on relational databases. The data mart, provided as open source, contains one central fact table surrounded by five dimension tables: patient, concept, visit, provider and modifier. In general, the most important concept regarding the construction of a star schema is identifying what constitutes a fact. In healthcare, a logical fact is usually an observation on a patient. Thus, the fact table can contain the basic attributes about the healthcare observation, such as the patient and provider numbers, a concept code for the concept observed and a start and end time for a procedure. The dimension tables contain further descriptive and analytical information about attributes in the fact table. Despite i2b2's model being generic enough to host other types of facts than healthcare oriented, there is a need for adapting the dimension tables to represent, for example, user interactions with social media or his/her own sensor measurements. This is true for any other type of non-specific healthcare related data in i2b2. Thus, such a model might pose some issues as a general model to represent MHMD datasets. Another issue is the complexity to load data in i2b2. As it is well known within i2b2's community, there is steep learning curve to get datasets ingested into the data mart.

6.1.2 Bioschemas specifications

Bioschemas is a collection of specifications that provide guidelines to facilitate a more consistent adoption of schema.org metadata model within the life sciences. Schema.org is an effort supported by the main search engines and is already widely implemented across the web. Bioschemas builds on it to improve data interoperability in life sciences, encouraging the community to use schema.org markup and have their websites and services with consistently structured information. This structured information then makes it easier to discover, collate and analyse distributed data. Similar to i2b2, Bioschemas operates as an open community initiative. The base of Bioschemas reference, i.e., schema.org, provides a way to add semantic markup to web pages. It describes 'types' of information, which then have 'properties'. For example, 'Event' is a type that has properties like 'startDate', 'endDate' and 'description'. If types or properties needed in the life sciences are missing, then Bioschemas defines common generic types, like events and datasets, and proposes their adoption by schema.org. While Bioschema provides a solid ontology for representation of metadata in life

sciences, particularly due to the utilization of schema.org, Bioschemas itself needs yet more tests to demonstrate how its metadata model can cope heterogeneous biomedical datasets.

6.1.3 DATS metadata model

DATS is the underlying model powering metadada ingestion, indexing and searches in DataMed, a NIH (National Institute of Health - US) funded project that aims to represent for biomedical datasets what PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>) is for the biomedical literature. Currently, DATS is used to index more than 70 repositories, including dbGaP (the database of Genotypes and Phenotypes), ClinicalTrials.gov and ClinVar, and catalogues more than 2.3 million biomedical datasets. DATS is designed in a modular approach with a core model, containing the most essential metadata elements, and an extended module, with specific elements for life, environmental and biomedical science domains and can be further extended as needed. The model is able to represent more than 15 data types, including phenotype, gene expressions, imaging data and clinical trials. It has been designed with the FAIR principles for data management in mind, allowing the assignment of persistent identifiers, enrichment of formal metadata, provenance tracking and licensing.

Given the flexibility, low complexity and proved success of DATS for representing and cataloguing biomedical datasets, we selected this model to ingest, index and catalogue MHMD metadata.

6.2 DATS physical model and instantiation of MHMD datasets

As showed in Figure 2, the DATS model is designed around the *Dataset*, an element that intends to cater for any unit of information stored by repositories. Dataset covers both (i) experimental datasets, which do not change after deposition to the repository, and (ii) datasets in evolving databases, describing dynamic facts and concepts, such as EHR and genes, whose definition morphs over time. The model includes three main types of elements: *Digital Research Object* element, such as the *Dataset* itself, but also *DataRepository* and others; *Information Entity* element, such as *License* and *Access*; and *Material* element, such as *Organization* and *Person*. The *Dataset* element is also linked to other digital research objects, such as *License* and *DatasetDistribution*. Due to the aim for the maximum coverage of ingesting and search use cases, the model may appear quite detailed in some places. Nevertheless, DATS anticipates that not all use cases can be fulfilled, and that it is difficult to foresee all type of data sources the model should represent in a distributed and heterogeneous data sharing network.

DATS formal specification is provided and maintained at https://github.com/biocaddie/DATS/blob/master/dataset_schema.json.

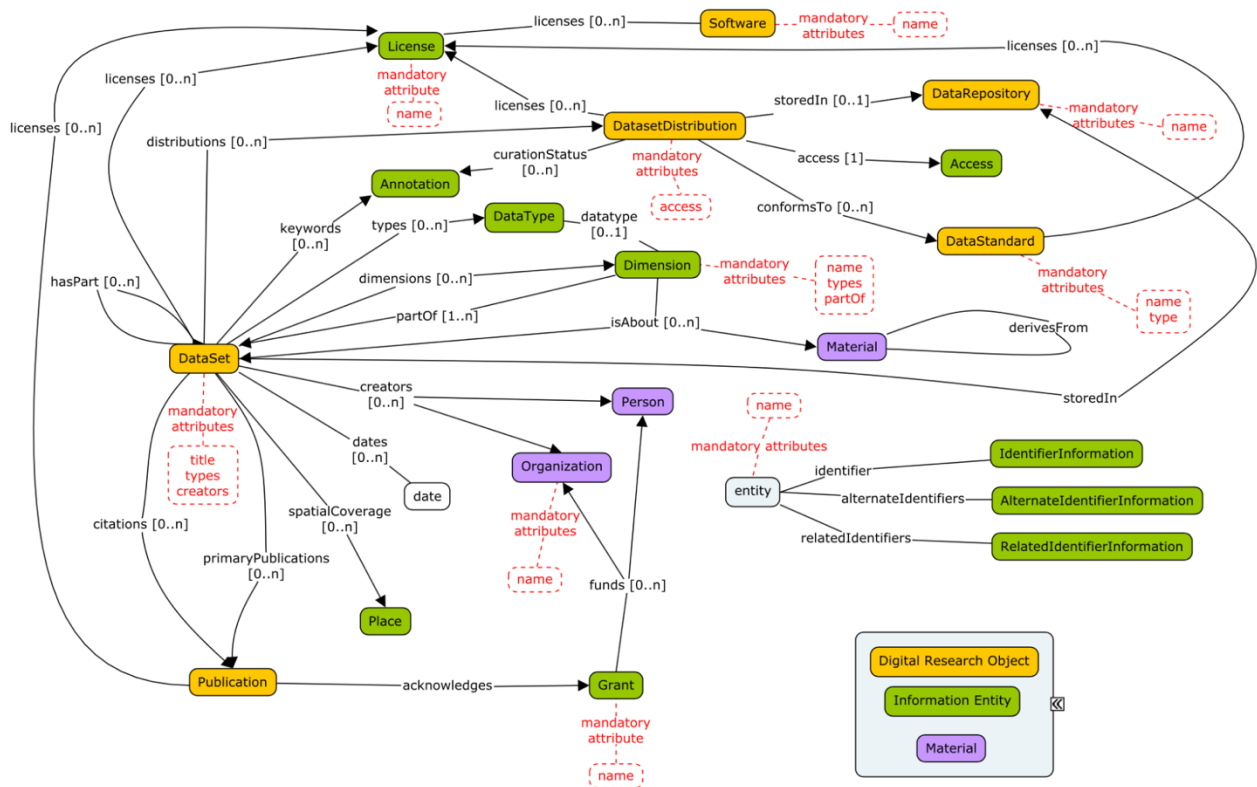


Figure 2 - DATS core data model [15]

Figure 3 shows an example of how the DATS model is being used to represent metadata for a MHMD dataset. We use 8 high level elements from DATS to represent the full MHMD metadata: identifier, title, description, types, creators, producedBy, keywords, and distributions. The *identifier* element contains a code uniquely identifying an entity locally to a system or globally; the *title* element provides the name of the dataset, usually one sentence or short description of the dataset; the *description* element contains a textual narrative comprised of one or more statements describing the dataset; the *types* element is a term, ideally from a controlled terminology, identifying the dataset type or the nature of the data; the *creators* element describe the person(s) or organization(s) which contributed to the creation of the dataset; the keywords element contains tags associated with the dataset, which will help in its discovery; the *producedBy* element describes the study process which generated a given dataset, if any; and the *distributions* element describes how datasets are made available, including relevant dates associated to the distribution and what is the provenance of the dataset.

```

{
  "identifier": {2 items},
  "title": "CRS identified",
  "description": "QMUL health data - CRS identified",
  "types": [{
    "value": "Magnetic Resonance Imaging (MRI) of Heart"
  }],
  "creators": [1 item],
  "keywords": [{
    "valueIRI": "https://uts.nlm.nih.gov/metathesaurus.html?cui=C1706428",
    "value": "Male Phenotype"
  }, {
    "valueIRI": "https://uts.nlm.nih.gov/metathesaurus.html?cui=C0018802",
    "value": "Congestive heart failure"
  }],
  "distributions": [
    {
      "identifier": {2 items},
      "access": {2 items},
      "dates": {2 items},
      "storedIn": {5 items}
    }
  ],
  "producedBy": {
    "name": "QMUL MHMD connector",
    "studyGroups": [1 item],
    "selectionCriteria": [1 item],
    "input": [1 item]
  }
}

```

Figure 3 - Example of a dataset metadata represented using DATS

6.3 Logical model on top of DATS

Combining the DATS model with the MeSH ontology, MHMD metadata can be modelled and queried as an entity-attribute-value (EAV) model (Figure 4). In this scenario, each *entity* of the EAV model represents a dataset, which can refer to a single individual or to a population (or study cohort), according to the data source specifications, and is described by the attributed *identifier* in the DATS model. Then, the *keywords* tag encodes the concept *values* for the entity, associated to a respective *attribute* provided by the MeSH semantic types.

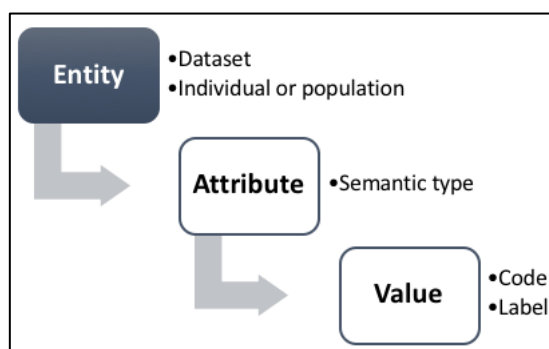


Figure 4 - EAV model for MHMD data catalogue

7 Data normalization service

As we could see from the datasets described in section 4, datasets that will be shared in MHMD are very heterogeneous in both format and content. To allow data sources to publish their data in the network, using the MeSH ontology described in section 5 and the metadata model described in section 6, we developed a normalization service called TransMeSH.

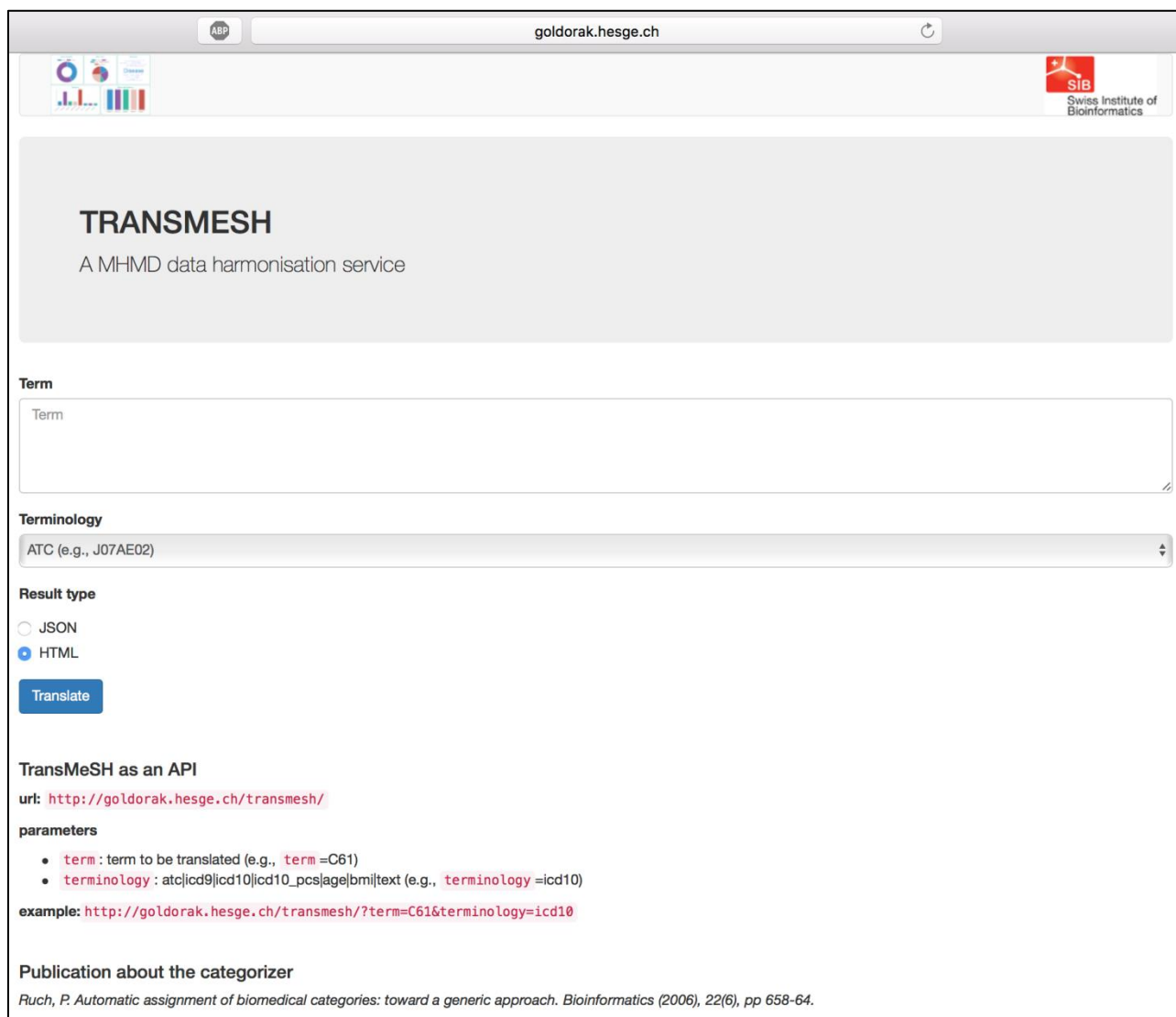
TransMeSH translates data values available in the datasets to MeSH concepts using MetaMap as proximity score algorithm for text attributes [16], rules for continuous scale attributes, and a mix of translation table and MetaMap to standard terminologies, such as ATC and ICD PCS. MetaMap identifies biomedical concepts from free-form textual input and maps them into concepts from the Unified Medical Language System (UMLS) Metathesaurus. On off-the-shelf experiments, the algorithm achieves precision of 95% to 98% [17]. The algorithm works by breaking the text into phrases and then, for each phrase, it returns the mapping options ranked according to the strength of mapping. On the other hand, for continuous scale attributes, such as age and BMI, our normalization algorithm implements rules to convert the input directly into MeSH codes. For example, for age = 85, the algorithm returns the MeSH code D000369: 'Aged, 80 and over'. In this sense, the harmonization tool also helps with the anonymization of the dataset. Finally, for standard terminologies, a code --> label expansion table is used to extract the label for the specific code, which is then fed to MetaMap for annotation.

Currently, the normalization service harmonizes the following data types:

- ICD-9 (clinical --> diagnosis)
- ICD-10 (clinical --> diagnosis)
- ICD-10 PCS (procedure)
- ATC (clinical --> medication/prescription/vaccination)
- gender (demographics)
- age (demographics)
- BMI (demographics)
- time (miscellaneous)
- text (miscellaneous)
- diabetic (clinical --> diagnosis)
- hypertension (clinical --> diagnosis)

The attributes of digi.me and QMUL datasets were manually assigned to each of these normalization types. For the moment, attributes that do not have a normalization function, such as administrative organization concepts (in Icelandic, arrival_emergency dataset), are not being loaded in the metadata catalogue model.

TransMeSH is available as a REST API at <http://goldorak.hesge.ch/transmesh>. As showed in Figure 5, the service takes as input an attribute value, the type of attribute (ICD9, ATC, etc.) and the result format (JSON or HTML), and it returns the UMLS/MeSH codes associate to it. For example, for the input term J07AE02, the algorithm returns the MeSH codes showed in Figure 6: D022121, D014612 and D001428. ATC code J07AE02 represents the concept 'cholera, live attenuated' (https://www.whocc.no/atc_ddd_index/?code=J07AE02), which is under the 'Cholera vaccines' (J07AE) and 'BACTERIAL VACCINES' (J07A) super classes. With 'Vaccines', 'Bacterial Vaccines' and 'Cholera Vaccines', the result of TransMeSH seems like a good match for the code J07AE02.



ABP goldorak.hesge.ch

TRANSMESH
A MHMD data harmonisation service

Term

Term

Terminology

ATC (e.g., J07AE02)

Result type

☐ JSON

☒ HTML

Translate

TransMeSH as an API

url: <http://goldorak.hesge.ch/transmesh/>

parameters

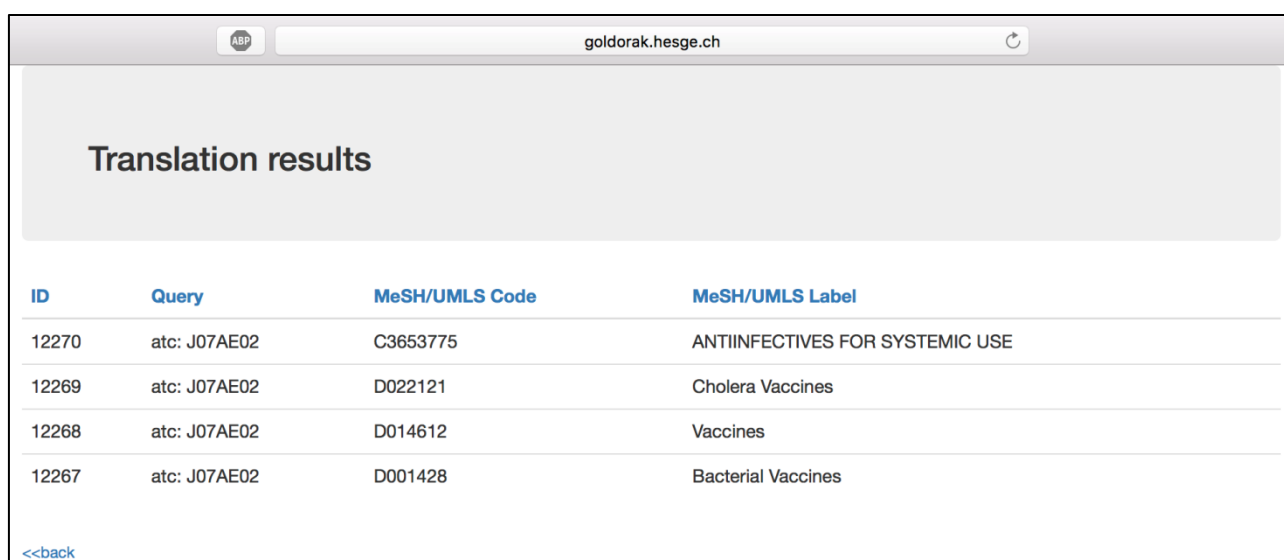
- term**: term to be translated (e.g., **term** = C61)
- terminology**: atc|icd9|icd10|icd10_pcs|age|bmi|text (e.g., **terminology** = icd10)

example: <http://goldorak.hesge.ch/transmesh/?term=C61&terminology=icd10>

Publication about the categorizer

Ruch, P. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics* (2006), 22(6), pp 658-64.

Figure 5 - TransMeSH - a MHMD data harmonisation service



ABP goldorak.hesge.ch

Translation results

ID	Query	MeSH/UMLS Code	MeSH/UMLS Label
12270	atc: J07AE02	C3653775	ANTIINFECTIVES FOR SYSTEMIC USE
12269	atc: J07AE02	D022121	Cholera Vaccines
12268	atc: J07AE02	D014612	Vaccines
12267	atc: J07AE02	D001428	Bacterial Vaccines

[<<back](#)

Figure 6 - Normalisation results for query ATC:J07AE02

8 Data stewardship

The goal of the data stewardship activities in MyHealthMyData is to ensure that the evolution of data is captured at the right level of granularity and detail, to be able to identify when data was created, modified or deleted. This provides a basis for reproducing data sets as they were used at a specific moment in time. Such reproduction is important to verify and analyse what exact data records were retrieved and accessed by consumers, to allow for a transparent account of the data usage.

The data records actually utilised for an investigation in an eco system such as MyHealthMyData are difficult to precisely identify. On the one hand, they are volatile in nature (they might change by addition, deletion or updating of records), and on the other hand, they are composed of records from several distributed source systems. At the same time, consumers of the data generally access only a specific subset of the data, namely those that enable them to answer their research questions. These subsets might be both a subset of available records (rows) and available attributes (columns), as specific research questions might only need a subset of the available data, and data minimisation concerns would require only obtaining necessary data.

To address these issues, the MyHealthMyData system will rely on implementing the Research Data Alliance (RDA)² recommendations on data subset identification and citation published by the Data Citation Working Group of the RDA³. When implemented, these recommendations rely on versioning of the data, which also enables provenance tracking, and also to precisely identify arbitrary subsets of data. The MyHealthMyData architecture requires a specific solution as the nature of the data is distributed, but also these can be addressed with the recommendations.

As the work in this task only starts in the month of the delivery of this deliverable, the description of the strategy for data stewardship outlined below is on an abstract, and specific implementation details are still subject to change. Deliverable D4.4 Data Stewardship modules, due in M36, will report on the system eventually provided.

8.1 Data Versioning, timestamping and identification

The recommendation from the Data Citation working group comprises the following 14 recommendations:

- Preparing the Data and the Query Store
 - R1 - Data Versioning
 - R2 – Timestamping
 - R3 - Query Store Facilities
- Persistently Identifying Specific Data Sets
 - R4 - Query Uniqueness
 - R5 - Stable Sorting
 - R6 - Result Set Verification
 - R7 - Query Timestamping
 - R8 - Query PID
 - R9 - Store the Query
 - R10 - Automated Citation Texts
- Resolving PIDs and Retrieving the Data

² <https://www.rd-alliance.org/>

³ <https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html>

- R11 - Landing Page
- R12 - Machine Actionability
- Upon modifications to the Data Infrastructure
 - R13 - Technology Migration
 - R14 - Migration Verification

Figure 7 RDA Data Citation Recommendations as components.

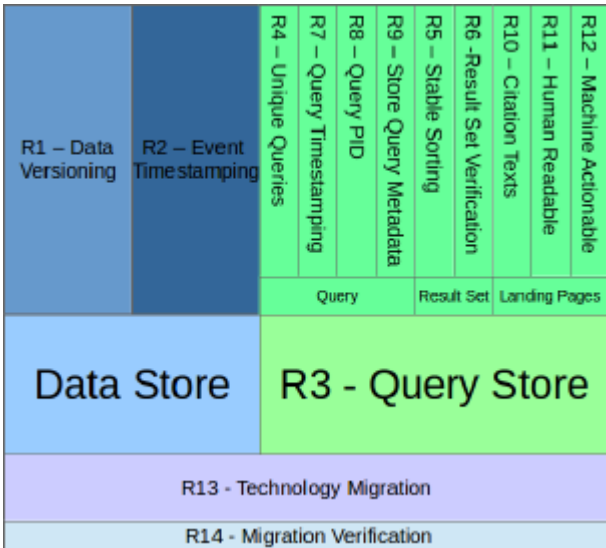


Figure 7 RDA Data Citation Recommendations as components⁴

Note that not all activities are required for each application scenario, and one activity in the task 4.4. Data Stewardship will be in identifying the components that are import for MyHealthMyData.

The core activities are therefore the following:

1. Data versioning
2. Data time stamping
3. Data identification

As data changes over time, a system that wants to be able to provide data as it was accessed at a certain point in time needs to keep versioning information on the data, sometimes also referred to as temporal tables in database systems. Such versioned data storage does not directly insert, delete or update records. The system keeps additional information (markers) on each record that indicates the operation. In fact, updates and deletions are rather emulated by marking certain records as not current, instead of actually deleting over overwriting the information.

Next to versioning, timestamping of the operations, i.e. recording the exact moment when the operation happened (e.g. an creation timestamp), specifies the validity of a specific record (a time interval). In this way, data can be produced that reflected the state of the database at a certain moment in time.

⁴ Andreas Rauber and Ari Asmi and Dieter van Uytvanck and Stefan Proell. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. Bulletin of IEEE Technical Committee on Digital Libraries (TCDL). 2016

Besides versioning, such information also provides a basic provenance of the data, as the creation, evolution and eventual deletion is recorded. While other information often regarded as data provenance, such as the authors/creators of data and changes, is not recorded, this information is not of primary interest to the MyHealthMyData system, as the concern is on the access and usage of data.

A further important aspect of the data citation solution is the query store, i.e. the component that stores information on which queries were executed, at which point in time, as well as fixity and checksum information on the provided result set.

8.2 Data Citation in MyHealthMyData

This section gives a high-level overview on aspects to be considered when implementing the data stewardship modules in MyHealthMyData. The exact definition on the system finally adopted is subject to further investigation in Task 4.4, which ends in M36.

One challenge in the MyHealthMyData architecture is the distributed nature of the data sources. Thus, the recommendations from the RDA working group that address distributed data settings are of high importance in this task, which concerns e.g. the fixity information and the federation of results obtained, as well as the broadcasting of the query to the various data sources.

As versioning, for efficiency reasons, needs to happen close to the data source, and it is further not desired to disclose any data, also not the versioned data, more than necessary, the technical implementation of the versioning will happen in each of the data source providers in the MyHealthMyData ecosystem. During this project, the primary focus will be on the FEDEHR system, which is the major system deployed at the data sources, and thus covers initial providers. For future other data source providers, implementation of the versioning and further functionality is a requirement to fully participate in the exchange.

Besides the data sources, also the data stored in the consent given by data subjects for their data being used in investigations needs to be addressed, thus data provenance and citation capabilities will also be integrated in these data bases.

From a technical point of view, in database systems, versioning can happen in the same data base as the currently valid data is stored, or in a separate database table. The latter is preferable in case frequent access to the table requires a high performance⁵, and will be considered for MyHealthMyData.

The technical solution for providing data versioning will depend on the capabilities of the underlying data base management system at each data source. As much as possible, built-in functionality for temporal databases as defined in the ANSI SQL 2011 standard will be utilised.

The query store, i.e. the component recording metadata on the queries and results obtained, is to be centrally available for access. It will thus be a component close to the Data Catalogue.

⁵ Snehil Gupta, Connie Zabarovskaya, Brian Romine, Daniel A. Vianello, Cynthia Hudson Vitale, and Leslie D. McIntosh. Incorporating Data Citation in a Biomedical Repository: An Implementation Use Case. AMIA Jt Summits Transl Sci Proc. 2017; 2017: 131–138.

A further research aspect in this task is the combination of the blockchain as a central audit trail and log of data provisioning and usage, and the combination of this with the data stewardship modules. It will be investigated to what extent the blockchain shall contain information that is traditionally provided in the query store, to provide also a decentralised authority on this information, which in turn could increase the trust in the overall system.

9 Data catalogue prototype

A prototype of the data catalogue has been developed by HES-SO. The objective of the data catalogue is first to give the user a view of the data available in the catalogue, as well as enabling to search for records given a set of keywords. Finally, the user must be able to request access to data.

9.1 Interface

A graphical user interface has been developed for the data catalogue. The prototype is accessible at the following URL: http://candy.hesge.ch/MHMD_Catalogue. It is to be mentioned that the data catalogue is based on a non-static set of records. Therefore, the screenshot presented in this deliverable represents the view of the data catalogue at a given point-in-time.

The user is first presented with a page where the data available in the data catalogue can be visualized (Figure 8). Several diagrams are presented, displaying specific information. First, a pie chart shows some demographic information (i.e. gender) about the people populating the data catalogue. For the moment, only the gender is represented in the demographic pie chart, but additional demographic information could be added in a near future (e.g. age-related groups, location, etc.) Second, a cloud is displaying the most frequent keywords stored in the data catalogue. This list is based on the MeSH terminology. This diagram aims at giving a quick view on the data available. Third, the user is presented with four bar diagrams, each representing a different MeSH category (i.e. diseases, drugs and chemicals, procedures and anatomy). Additional MeSH categories can optionally be added in the future, depending on the data available in future datasets. For each of these diagrams, the view is based on the tree structure. Indeed, by clicking on a bar, the user will see a more detailed view of the data. For instance, by clicking on the “Cardiovascular diseases” bar, the user will see that 974 cases are representing “Heart diseases”, and 565 cases are representing “Vascular diseases”. The data can be browsed until the leaves are reached.



MHMD DATA CATALOGUE

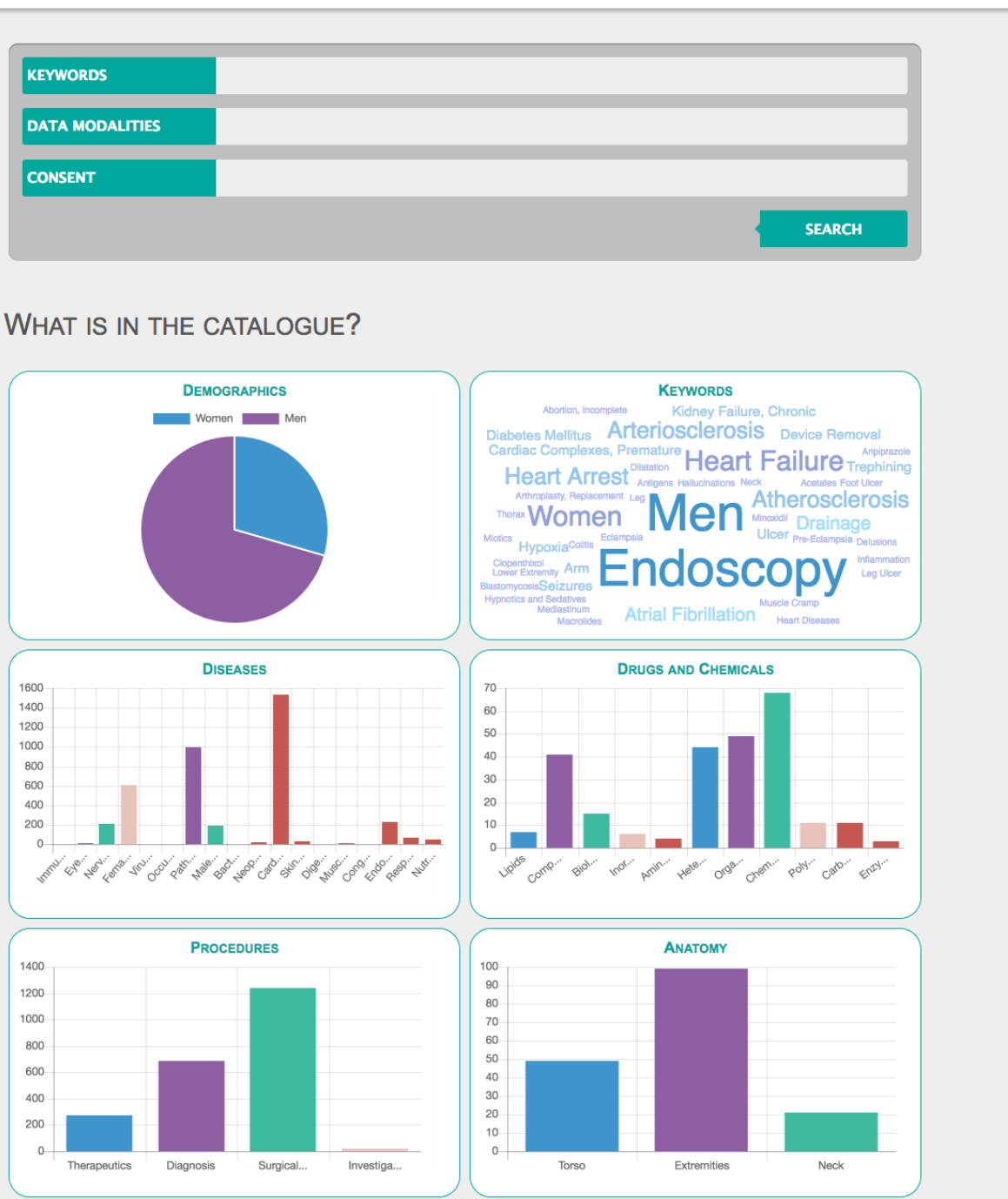


Figure 8 View page of the data catalogue

On top of the page, a search box is available to search for specific records. Three fields are proposed: keywords, data modalities and consents. The search by keywords enables to search for all records mentioning one or several MeSH terms (e.g. “heart diseases”). The search by data modalities enables to filter returned records to one or several data modalities (e.g. “prescription”). Finally, the consent field can be used to filter returned records to one or several consent types (e.g. “synthetic data”). Results are displayed in a table (Figure 9). If a given study is containing several records (e.g. corresponding to different modalities or consent types), the records are grouped together. For each record, a short description is displayed, as well as the consent type.



MHMD DATA CATALOGUE

KEYWORDS	Heart Diseases x
DATA MODALITIES	
CONSENT	
SEARCH	

RESULTS

Page 1/66 (653 individuals / 653 records)			
	278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442		<input type="checkbox"/>
PID	Description	Consent	
278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442	QMUL health data - crs identified	consent not required (synthetic data)	<input type="checkbox"/>
	f377ba0c34fd82ad354deeb4310550d042823f8bf72422cd918aacb0		<input type="checkbox"/>
PID	Description	Consent	
f377ba0c34fd82ad354deeb4310550d042823f8bf72422cd918aacb0	QMUL health data - crs identified	consent not required (synthetic data)	<input type="checkbox"/>
	f71766949585031e89685cc021e63ae2b881d582a75e0555624becca		<input type="checkbox"/>
PID	Description	Consent	
f71766949585031e89685cc021e63ae2b881d582a75e0555624becca	QMUL health data - crs identified	consent not required (synthetic data)	<input type="checkbox"/>
	a9c21f14a3fc70e0b25ad6e8504ddccea84656e5a1f3f181c4792238		<input type="checkbox"/>

Figure 9 Search page of the data catalogue

Finally, the user can select a few records of interest and request access to these records. The query will be sent to the blockchain, which will dispatch it to the different data providers. A temporary request page (Figure 10) has been set up but will be modified once the blockchain services will be ready.



MHMD DATA CATALOGUE

DATA ACCESS REQUEST

QUERY

Query: keywords:Heart Failure;

PID: -593102693

SELECTED DATA

You have requested access to the following records:

- 278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442
- 0f40f6c311bc7dbedfd919edb749f4e2efa7d9c849b7df48695861f8
- de9178147801c8f10afe173c65303754284c44a6d0dd02a8fa7cda9b
- a9c21f14a3fc70e0b25ad6e8504ddccea84656e5a1f3f181c4792238
- a0b79490cd8703a6f3ac482748ab391383c4660f2459ff62b9d48679
- f9cf9804bc619972a5d35059459d2323e4d76f05acf30ffb3b297555

ADHERENCE COMMITMENT

Please fill the [adherence commitment form](#) and send a signed copy to sibtextmining@gmail.com

REQUEST FINALIZATION

Enter your email address

Click on the "Request data" button to finalize your request. Once your adherence commitment form will have been accepted, you will receive an email providing you access to the selected data.

[Back](#)

REQUEST DATA

Figure 10 Request data page

9.2 Query example

We present in this section an example of a query. Two keywords have been entered: “heart failure” and “pregnancy complications”. The user also required retrieving only synthetic data. It results in a list of 101 records corresponding to 101 individuals (i.e. each individual owns only one record in this dataset).

The screenshot displays a search interface with three filter sections: KEYWORDS (Heart Failure, Pregnancy Complications), DATA MODALITIES, and CONSENT (synthetic data). A SEARCH button is located on the right. Below the filters, the RESULTS section shows a table of 101 records. The table has columns for PID, Description, and Consent. The first three records are visible, each with a document icon, a person icon, and a checkbox.

PID	Description	Consent
278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442	QMUL health data - crs identified	consent not required (synthetic data)
e28bee41368fa4005ec45aed884920015fba162dd17d3a1705e42080	QMUL health data - crs identified	consent not required (synthetic data)
5b637c3e9ccc7a4277e52aea0c70450dcfcaa01c396726bea3d7e652	QMUL health data - crs identified	consent not required (synthetic data)

Figure 11 Example of a query

9.3 Data catalogue index

The data catalogue is based on an Elasticsearch index (version 6.3.2). Data are formatted in the DATS format (see section 6 for more details) and loaded in the index. Currently, the data index contains 2240 records. The data catalogue index can be directly queried at the following URL: <http://goldorak.hesge.ch:8991/catalogue>. Elasticsearch syntax must be used to perform queries on this endpoint.

9.4 Catalogue API

In addition to the Elasticsearch endpoint, several APIs are proposed to query the data catalogue: an API to visualize the data content, an API to search for records and an API to retrieve one record.

9.4.1 View of the data

We propose an API to get a view on the data. Given a data type, this service returns each concept and the corresponding number of occurrences. For instance, you can use this service to get the list of all contract types and their proportion in the data catalogue. This API is used to display the diagrams on the main page of the data catalogue.

This API is available at the following URL:

http://candy.hesge.ch/MHMD_Catalogue/api/getViewJson.jsp

The list of parameters is presented in Table 2.

Parameter name	Value	Description
type	datatype	List of all data modalities present in the data catalogue (e.g. prescription)
	consents	List of all contract types present in the data catalogue (e.g. synthetic data)
	cloud	List of all keywords present in the data catalogue (i.e. MeSH terms)
	codeA	List of all anatomy terms present in the data catalogue
	codeC	List of all drugs and chemicals present in the data catalogue
	codeD	List of all diseases present in the data catalogue
	codeE	List of all procedures present in the data catalogue
	codeM	List of all information related to “Persons” present in the data catalogue
level	A/C/D/E/M	Enable to retrieve the first level of the tree categories. Can only be used with a type value of codeX.
	A MeSH terms	Enable to retrieve the children of a given MeSH terms present in the data catalogue. Can only be used with a type value of codeX.

Table 2 Parameters for the *getViewJson.jsp* API

The output data is formatted using JSON format (Figure 12). The data JSON object contains an array of label-value objects. The label field contains the MeSH terms identified in the data catalogue, while the value field contains the number of occurrences of the MeSH terms in the data catalogue. In this example, they are four concepts corresponding to heart diseases in the data catalogue. Each of these concepts may also contain sub concepts (children).

```
{
  "data": [{
    "label": "Arrhythmias, Cardiac",
    "value": 394
  }, {
    "label": "Heart Arrest",
    "value": 232
  }, {
    "label": "Myocardial Ischemia",
    "value": 6
  }, {
    "label": "Heart Failure",
    "value": 333
  }
]}
```

Figure 12 Example of the output of this API for the following query:

http://candy.hesge.ch/MHMD_Catalogue/api/getViewJson.jsp?type=codeD&level=Heart%20Diseases

9.4.2 Search for several records

We propose an API to search for records. Given a set of keywords and filters, the service returns all records corresponding to the query. Records are grouped by study/patient. This API is used to search for records in the data catalogue.

This API is available at the following URL: http://candy.hesge.ch/MHMD_Catalogue/api/search.jsp

The list of parameters is presented in Table 3.

Parameter name	Value	Description
keywords	Any MeSH term (e.g. Heart failure)	To retrieve records based on a set of MeSH terms. MeSH terms are separated by semi-colons.
datatypes	Any data modality (e.g. prescription)	To filter records to only one or several data modalities (e.g. prescription). Data modalities are separated by commas.
consents	Any consent description type (e.g. synthetic data)	To filter records to only one or several contract types (e.g. synthetic data). Consents are separated by commas.

Table 3 Parameters for the search.jsp API

The output data is formatted using JSON format (Figure 13). The data JSON object contains an array of study/patient objects. For each study/patient, the following fields are displayed:

- A study/patient identifier
- An array of records

For each record, the following fields are proposed:

- An Elasticsearch identifier (esid)
- A persistent identifier (pid)
- A short description of the data
- The data provider name
- The type of consent for this record.

```

{
  "hitsrecords": 101,
  "hitspatients": 101,
  "data": [{
    "study": "278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442",
    "records": [{
      "esid": "278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442",
      "pid": "278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442",
      "desc": "QMUL health data - crs identified",
      "dataprovder": "QMUL",
      "consent": "consent not required (synthetic data) "
    }
  ],
  // ...
  {
    "study": "0fcffa8873e355ac57788de0d47926ff44c3aa08cb72742a4971955a",
    "records": [{
      "esid": "0fcffa8873e355ac57788de0d47926ff44c3aa08cb72742a4971955a",
      "pid": "0fcffa8873e355ac57788de0d47926ff44c3aa08cb72742a4971955a",
      "desc": "QMUL health data - crs identified",
      "dataprovder": "QMUL",
      "consent": "consent not required (synthetic data) "
    }
  ]
}

```

Figure 13 Example of the output of this API for the following query:

http://candy.hesge.ch/MHMD_Catalogue/api/search.jsp?keywords=Heart%20Failure;Pregnancy%20Complications&consents=synthetic

9.4.3 Search for one record

We propose an API to search for one specific record. Given the Elasticsearch id, the service returns the full DATS-formatted record. This API is used to display each record in the data catalogue.

This API is available at the following URL:

http://candy.hesge.ch/MHMD_Catalogue/api/getRecord.jsp

This service is only enabling one parameter: the Elasticsearch ID (Table 4).

Parameter name	Value	Description
esid	Any Elasticsearch ID	To retrieve details about one specific record

Table 4 Parameters for the *getRecord.jsp* API

The output data is formatted using JSON format (Figure 14). It returns the DATS-formatted record, as described in section 7.


```

{
  "distributions": [{
    "identifier": {
      "identifier": "278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442",
      "identifierSource": "catalogue"
    },
    "access": {
      "landingPage": "https://myhealthmydata.eu/data/QMUL/",
      "accessURL": "https://myhealthmydata.eu/data/QMUL/local_mapping:
278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442"
    },
    "dates": [{
      "type": {
        "valueIRI": "",
        "value": "creation"
      },
      "date": "2018-04-03"
    }, {
      "type": {
        "valueIRI": "",
        "value": "publication date"
      },
      "date": "2018-04-03"
    }],
    "storedIn": {
      "identifier": {
        "identifier": "https://myhealthmydata.eu/data/QMUL/",
        "identifierSource": ""
      },
      "licenses": [{
        "identifier": {
          "identifier": "https://consent.uri/id=consent not required",
          "identifierSource": "https"
        },
        "version": "",
        "name": "consent not required",
        "extraProperties": [{
          "category": "synthetic data",
          "values": [{
            "consent": "fully anonymised"
          }]
        }]
      }],
      "version": "",
      "name": "QMUL",
      "types": [{
        "valueIRI": "",
        "value": "csv"
      }]
    },
    "keywords": [{
      "valueIRI": "https://uts.nlm.nih.gov/metathesaurus.html?cui=C1706428",
      "value": "Male Phenotype"
    }, {
      "valueIRI": "https://uts.nlm.nih.gov/metathesaurus.html?cui=C1706429",
      "value": "Male, Self-Reported"
    }
  ]
}

```

Figure 14 Example of the output of this API for the following query:

http://candy.hesge.ch/MHMD_Catalogue/api/getRecord.jsp?esid=278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442

10 Conclusion and next steps

In this report, we describe the main steps to harmonize datasets in the MHMD data catalogue. The source datasets used are synthetic representation of QMUL and digi.me datasets. As expected, these datasets are very heterogeneous concerning their format and content. We applied a mixed bottom-up and top-down approach to specify the ontological resources and metadata model. After the analyses of the sources, and of biomedical ontologies and metadata models, we decided for MeSH and DATS to represent the content and hierarchy of concepts, and the unified metadata model, respectively. Additionally, we implemented a tool, TransMeSH, which can help data sources to (semi-)automatically convert their dataset content into MeSH concepts. Finally, we describe a prototype data catalogue explorer, which can be used to search and discover datasets in the MHMD network.

For the next steps, we will work on the design of the ingestion and automatic indexing of metadata in the MHMD catalogue. We will also investigate how we can connect the metadata publish in the catalogue to the actual source data using only metadata tags, so to avoid disclosure of any local identifier or provenance information.

References

1. MD-Paedigree.
2. CardioProof.
3. EHR4CR.
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* [Internet]. 2016;3:160018. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4792175&tool=pmcentrez&rendertype=abstract>
5. Cipi re S, Ereteo G, Gagnard A, Boujelben N, Gaspard S, Breton V, et al. Global Initiative for Sentinel e-Health Network on Grid (GINSENG): Medical data integration and semantic developments for epidemiology. In: *Proceedings - 14th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2014*. 2014. p. 755–63.
6. HIPAA.com. HIPAA “Protected Health Information”: What Does PHI Include? [Internet]. HIPAA.com. 2017. p. 1–25. Available from: <https://www.hipaa.com/hipaa-protected-health-information-what-does-phi-include/>
7. Centers for Disease C, Prevention. HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. *MMWR - Morb Mortal Wkly Rep*. 2003;52 Suppl:1–17.
8. Information Commissioner’s Office Anonymisation Code [Internet]. Available from: <https://ico.org.uk/media/1061/anonymisation-code.pdf>
9. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J Cardiovasc Magn Reson*. 2017;19(1).
10. WHO Collaborating Centre for Drug W, Norwegian Institute of Public Health N. ATC/DDD Index 2014. 2014.
11. WHO. ICD-10 Version:2016. Who. 2016. <http://apps.who.int/classifications/icd10/browse/2>.
12. NIH-NLM. SNOMED Clinical Terms  (SNOMED CT ) [Internet]. NIH-US National Library of Medicine. 2015. Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
13. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. Vol. 8, *Journal of the American Medical Informatics Association*. 2001. p. 317–23.
14. Keet CM, Rodriguez M. Comprehensiveness versus scalability: guidelines for choosing an appropriate knowledge representation language for bio-ontologies. 2007.
15. Sansone SA, Gonzalez-Beltran A, Rocca-Serra P, Alter G, Grethe JS, Xu H, et al. DATS, the data tag suite to enable discoverability of datasets. *Sci Data*. 2017;4.
16. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings AMIA Symp* [Internet]. 2001;17–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11825149>
17. Chiamello E, Pincioli F, Bonalumi A, Caroli A, Tognola G. Use of “off-the-shelf” information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *J Biomed Inform* [Internet]. 2016;63:22–32. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27444186>

Appendix

Sample sets

QMUL

CRS_IDENTIFIED

Episode Start Date,Diagnosis Codes,Procedure Codes,NHS Number,Given Name,Family Name,Address Line 1,Address Line 2,Address Line 3,Address Line 4,Postcode,Date of Birth,Date of Death,Marital Status,Ethnic Origin,Gender,Health Authority of Residence,Registered GP,Registered GP Practice,index,Imaging types

27 May

2010,ICD9:427.31|ICD9:785.2|ICD9:518.83|ICD10:J96.21|,0SG14A1|0BCG4ZZ|051N0ZY|0D1K4J4|024F0KJ,1759961146,Philipa,Cully,672 Billingbawk Street,West Dunbartonshire,Scotland,G60 9DB,G60 9DB,30 Aug 1976,,widowed,other,female,West Dunbartonshire,"Cully,Philipa",Park GP Practise,0,Cardiac MRI|Echocardiography

CVI_IDENTIFIED

Address,Phone,Last Name,First Name,Middle Name,Medical Record Number,Prior Names,Ethnicity,Gender,Birthdate,Study Date,Patient Age,National Identifier,Blood Pressure,Heart rate,Height (cm),Weight (kg),Scanner Manufacturer,LVEDV (ml),LVESV (ml),LVSV (ml),LVEF (%),LV Mass (g),RVEDV (ml),RVESV (ml),RVSV (ml),RVEF (%),RV Mass (g),BMI (kg/msq),BSA,BSA (msq),CO (L/min),Central PP(mmHg),DBP (mmHg),Diabetic RF,Hypertension RF,LVEF (ratio),MAP,PAP (mmHg),PP (mmHg),RVEF (ratio),SBP (mmHg),SVR (mmHg/L/min),Smoking RF,Vascular RF,risk_name

"45 George Julius Rd,Wokingham,England,RG11 6MT",555778197

,Schaner,So,,2159268637,,white,female,19 Dec 1935,05 Jul

2008,72,1462919379,123/101,90,146.84125917106218,61.94370568696988,Siemens,118.36233409488048,37.697833418100814,80.66450067677967,68.1504815646151,76.0802964766289,159.86455749285932,60.32776962611274,99.53678786674658,62.2631991904726,31.63303506646367,28.72770833816111,1.589539743028041,1.589539743028041,7.25980506091017,12.965599081585946,101,non-diabetic,non-

hypertensive,0.6815048156461511,110.70933333333333,11.196517943141346,22,0.622631991904726,123,15.249628936930929,smoker,atherosclerotic,intermediate

digime

ADMISSION

```
{
  "createddate": 1110553080000,
  "dischargedate": 1128533400000,
  "organization": "R171 Test organization",
  "responsiblephysician": "Nemandi_2",
  "servicegroup": "R171 BarnaskurÃ°deild(K)",
  "spectreatmentid": 491,
  "entityid": "16_3112754029_100_1110553080000"
```

}

ALLERGY

{

"atc": "",
 "classification": "BrÃ¸Ã¸aofnÃ¸mi",
 "comment": "",
 "component":

"1234568790123456879012345687901234568790123456879012345687901234568790123456879012345687901234568790",

"createddate": 1449247364000,
 "type": "EfnaofnÃ¸mi",
 "entityid": "16_3112754029_107_"

}

ARRIVAL_AMBULATORY

{

"admissionstartdate": 1258453980000,
 "createddate": 1258737992000,
 "arrivalid": 20367,
 "responsiblephysician": "",
 "responsiblephysicianid": "",
 "responsiblephysicianmdno": "",
 "servicegroup": "BlÃ¸Ã¸lÃ¸kningar(T)",
 "servicegrouparrivaldischarge": "BlÃ¸Ã¸lÃ¸kningar(T) 17.11.2009",
 "treatmentid": 27732,
 "entityid": "16_3112754029_101_20367"

}

ARRIVAL_EMERGENCY

{

"createddate": 1282733466000,
 "arrivalid": 30729,
 "organizationname": "Saga_skema - TMS",
 "responsiblephysician": "Agnar Tegnander",
 "spectreatmentid": 63533,
 "entityid": "16_3112754029_108_30729"

}

ARRIVAL_PRIMARY_HEALTH

{

"arriveid": 100814,
 "createddate": 1454344200000,
 "contactname": "20 mÃ¸n viÃ¸tal",
 "departmentorgname": "Saga_skema - TMS",
 "resourcenname": "GuÃ¸leif HarÃ¸ardÃ¸ttir (lÃ¸knir)",
 "entityid": "16_3112754029_102_100814"

```
}
```

DIAGNOSIS

```
{
  "code": "D55.0",
  "codingsystem": "ICD-10",
  "createddate": 1273795200000,
  "islongterm": true,
  "lastregistration": 1320019200000,
  "name": "ANAEMIA DUE TO GLUCOSE-6-PHOSPHATE DEHYDROGENASE [G6PD]
DEFICIENCY",
  "entityid": "16_3112754029_105_D55.0"
}
```

MEDICATION

```
{
  "atccode": "J07BF03",
  "autoseponate": "",
  "conceptcode": 2348,
  "createddate": 1395408305000,
  "daysleft": "Lokið",
  "directions": "1 ml að morgni, 1 um miðjan dag, 0 að kvöldi",
  "form": "stl",
  "lastchanged": 1395360000000,
  "lastprescribed": 1395360000000,
  "name": "IMOVAX POLIO",
  "nrnorr": "042556",
  "numberofpackings": 1,
  "numeroftimes": "1 sinni",
  "onetimeonly": true,
  "prescriptionends": 1395705600000,
  "quantity": "10",
  "skammtaaskja": "0",
  "strength": "",
  "totalquantity": "10",
  "usedfor": "",
  "entityid": "16_3112754029_104_J07BF03"
}
```

PRESCRIBED_ITEMS

```
{
  "atccode": "M01AE01",
  "dosageinstructions": "1 tafla Á dag",
  "form": "tÁflur",
  "ismonitored": false,
  "itemno": 0,
  "name": "IBUFEN",

```

```

    "numberofpackages": 1,
    "prescriptionid": "16_3112754029_103_273167",
    "createddate": 1424096880000,
    "productid": "116584",
    "quantity": 100,
    "strength": "400 mg",
    "unit": "stk",
    "entityid": "16_3112754029_109_M01AE01"
  }

```

PRESCRIPTION

```

{
  "daysbetweendispendations": 0,
  "earliestdispensationdate": 0,
  "id": 273132,
  "idbygateway": "",
  "iscanceled": false,
  "createddate": 1423785600000,
  "latestdispensationdate": 0,
  "prescribeddispensations": 1,
  "prescribeditems": "16_3112754029_109_N05AH03",
  "prescribercontactinfo": "",
  "prescriberid": "1520",
  "prescribername": "Birgir Andri Briem",
  "prescriptiontype": 1,
  "entityid": "16_3112754029_103_273132"
}

```

VACCINATION

```

{
  "code": "J07AE01",
  "codename": "Dukoral® kÃ³lera / cholera, inactiv, whole cell",
  "codes": "KÃ³lera:A",
  "codingsystem": "ATC",
  "createddate": 1088640000000,
  "senderdescription": "SÃ¶gu dev skema",
  "sendergateway": "SAGA_DEV",
  "senderid": "RAxx0761",
  "sendersystem": "SAGA.NET",
  "entityid": "16_3112754029_106_J07AE01"
}

```

MEDIA

```

{
  "baseid": "4_1542271900260495153_182042",
  "cameramodelentityid": "",
  "commentcount": 0,
  "commententityid": ""
}

```

```

    "createddate": 1498073159000,
    "description": "",
    "displayshorturl": "",
    "displayurlindexend": 0,
    "displayurlindexstart": 0,
    "entityid": "4_1542271900260495153_182042",
    "filter": "Gingham",
    "interestscore": 0,
    "itemlicenceentityid": "",
    "latitude": 0,
    "likecount": 14,
    "link": "https://www.instagram.com/p/BVnQB86F-8xIEERyS_PisN_g2p5-V-
o3aUBd7A0/",
    "locationentityid": "",
    "longitude": 0,
    "mediaAlbumEntityID": "4_182042_182042_1",
    "mediaalbumname": "",
    "mediaid": "1542271900260495153_182042",
    "mediaobjectid": "",
    "mediaobjectlikeid": "",
    "name": "Team's getting smaller by 1. Good luck @mheap",
    "originatortype": 2,
    "personentityid": "4_182042_182042",
    "personfilerelativepath": "",
    "personfileurl": "https://scontent.cdninstagram.com/t51.2885-
19/11356960_1625044604420020_264796266_a.jpg",
    "personfullname": "Pascal Wheeler",
    "personusername": "pascalw",
    "postentityid": "4_182042_1542271900260495153_182042",
    "resources": [{
      "aspectratio": {
        "accuracy": 100,
        "actual": "1:1",
        "closest": "1:1"
      },
      "height": 150,
      "mimetype": "image/jpeg",
      "type": 0,
      "url": "https://scontent.cdninstagram.com/t51.2885-
15/s150x150/e35/c135.0.809.809/19367043_122168111710592_7572307112722694144_n.jpg",
      "width": 150
    }],
    {
      "aspectratio": {
        "accuracy": 99.581589958159,
        "actual": "320:239",
        "closest": "4:3"
      },
      "height": 239,
      "mimetype": "image/jpeg",
      "type": 0,

```



```

        "url": "https://scontent.cdninstagram.com/t51.2885-
15/s320x320/e35/19367043_122168111710592_7572307112722694144_n.jpg",
        "width": 320
    },
    {
        "aspectratio": {
            "accuracy": 99.79123173277662,
            "actual": "640:479",
            "closest": "4:3"
        },
        "height": 479,
        "mimetype": "image/jpeg",
        "type": 0,
        "url": "https://scontent.cdninstagram.com/t51.2885-
15/s640x640/sh0.08/e35/19367043_122168111710592_7572307112722694144_n.jpg",
        "width": 640
    }
}],
"tagcount": 0,
"taggedpeoplecount": 0,
"type": 0,
"updateddate": 1498073159000
}

```

POST

```

{
    "annotation": "",
    "baseid": "4_182042_1542271900260495153_182042",
    "cameramodelentityid": "",
    "commentcount": 0,
    "commententityid": "",
    "createddate": 1498073159000,
    "description": "",
    "entityid": "4_182042_1542271900260495153_182042",
    "favouritecount": 0,
    "iscommentable": 1,
    "isfavourited": 0,
    "islikeable": 1,
    "islikes": 0,
    "isshared": 0,
    "istruncated": 0,
    "latitude": 0,
    "likecount": 14,
    "longitude": 0,
    "originalcrosspostid": 0,
    "originalpostid": 0,
    "originalposturl": "",
    "personentityid": "4_182042_182042",
    "personfilerelativepath": "",
    "personfileurl": "https://scontent.cdninstagram.com/t51.2885-
19/11356960_1625044604420020_264796266_a.jpg",

```

```

    "personfullname": "Pascal Wheeler",
    "personusername": "pascalw",
    "postentityid": "",
    "postid": "1542271900260495153_182042",
    "postreplycount": 0,
    "posturl": "https://www.instagram.com/p/BVnQB86F-8xIEERyS_PisN_g2p5-V-
o3aUBd7A0/",
    "rawtext": "",
    "referenceentityid": "4_182042",
    "referenceentitytype": 15,
    "sharecount": 0,
    "socialnetworkuserentityid": "4_182042",
    "source": "",
    "text": "Team's getting smaller by 1. Good luck @mheap",
    "title": "",
    "type": 20,
    "updateddate": 1498073159000,
    "visibility": ""
}

```

COMMENT

```

{
    "appid": "",
    "baseid": "4_182042_17858982733161786",
    "commentcount": 0,
    "commentid": "17858982733161786",
    "commentreplyid": "",
    "createddate": 1497373638000,
    "entityid": "4_182042_17858982733161786",
    "likecount": 0,
    "link": "",
    "metaid": "",
    "personentityid": "4_182042_1545946064",
    "personfilerelativepath": "",
    "personfileurl": "https://scontent.cdninstagram.com/t51.2885-
19/s150x150/19051047_1129050077200786_761246497533591552_a.jpg",
    "personfullname": "Sarah Lamb",
    "personusername": "iamgirlygeekdom",
    "privacy": 0,
    "referenceentityid": "4_1536332454209789316_182042",
    "referenceentitytype": 1,
    "socialnetworkuserentityid": "4_182042",
    "text": "Oooh I missed out! Mint magnums are my favourite!",
    "updateddate": 1497373638000
}

```

TRANSACTION

```

{

```

```
"accountentityid": "17_10019612",
"amount": 2.43,
"basetype": "DEBIT",
"category": "Restaurants",
"categoryid": 22,
"categorysource": "SYSTEM",
"categorytype": "EXPENSE",
"checknumber": "",
"consumerref": "",
"createddate": 1490832000000,
"currency": "GBP",
"entityid": "17_10019612_10945571",
"highlevelcategoryid": 10000011,
"id": "10945571",
"ismanual": false,
"merchantaddress1": "",
"merchantaddress2": "",
"merchantcity": "",
"merchantcountry": "",
"merchantid": "costacoffee",
"merchantname": "Costa Coffee",
"merchantstate": "",
"merchantzip": "",
"originalref": "Debit COSTA COFFEE ON 26 MAR CPM",
"postdate": 1490832000000,
"runningbalance": 113.55,
"runningbalancecurrency": "GBP",
"simpleref": "Costa Coffee",
"status": "POSTED",
"subtype": "PURCHASE",
"transactiondate": 0,
"type": "PURCHASE"
```

```
}
```

Figure 15 - digi.me Medical Object Baseline