Knowledge Graph Construction for Health, Lifestyle and Fitness Applications

Carlo Allocca [a,*], Alessio Antonini [b], Riccardo Pala [b], Angelo Salatino [b], Iman Naja [b], Rohit Ail [a], Muhammad Salman Haleem [c,d], Laura Lopez-Perez [e], Eugenio Gaeta [e], Leandro Pecchia [c,f] and Giuseppe Fico [e]

[a] *Emerging Solutions Lab, Samsung, UK*
E-mails: carlo.allocca@samsung.com, salman.haleem@warwick.ac.uk
[b] *Knowledge Media Institute, The Open University,UK*
E-mails: alessio.antonini@open.ac.uk, riccardo.pala@open.ac.uk, angelo.salatino@open.ac.uk, iman.naja@open.ac.uk
[c] *School of Engineering, University of Warwick,UK*
E-mails: salman.haleem@warwick.ac.uk, l.pecchia@warwick.ac.uk
[d] *School of Electronic Engineering and Computer Science, Queen Mary University of London,UK*
E-mail: m.haleem@qmul.ac.uk
[e] *Life Supporting Technologies Research Group, Universidad Politécnica de Madrid,ES*
E-mails: llopez@lst.tfo.upm.es, eugenio.gaeta@lst.tfo.upm.es, gfico@lst.tfo.upm.es
[f] *Università Campus Bio-Medico,IT*
E-mail: leandro.pecchia@unicampus.it

Abstract.
Digital coaching for healthcare is challenging due to the heterogeneous nature of data sources. This often leads to the development of ad-hoc pipelines customised to different combinations of formats, which are hard to maintain and easily fall out of date. In this paper, we present MatKG and HeLiFit (**He**alth, **Li**festyle and **Fit**ness), which consist of a pipeline and an extended ontological model for scalable construction of a Knowledge Graph integrating electronic medical records, medical devices with consumer behavioural, and bio data. This fully-developed solution effectively addresses the challenge of *semantic interoperability* between healthcare institutions and consumer technology providers, using standards such as FHIR and RML supporting the construction of the cross-organisation health data space needed for powering a new generation of AI solutions. Its design and development were driven by a wide range of use cases and an equivalent number of digital coaching solutions for promoting health and lifestyle recommendations on different patient cohorts and healthcare institutions in 11 countries across Europe and Asia. We extended HeLiFit to accommodate a broader range of applications, including sleep and nutrition recommendations. The infrastructure is being piloted, involving thousands of users and different pools of experts engaged in the validation of the generated recommendations. The presented system is available as an off-the-shelf scalable solution that can fast-track innovation in the field of semantic AI for healthcare.

Keywords: Digital Coaching, Semantic Interoperability, Health Ontologies, Knowledge Graphs, FHIR, RML

*Corresponding author. E-mail: carlo.allocca@samsung.com.

1. Introduction

The efficacy of healthcare prevention actions and therapies relies on the sustained effort of patients and the general public in following recommendations. Recommendations can range from simple lifestyle guidelines against sedentary habits and food education to specific physical activity and dietary prescriptions. Regardless of their type, long-term (life-long) adherence is a common challenge with a clear impact on personal well-being and health, as well as on the cost and sustainability of health and social care services. The increase in life expectancy and the number of chronic patients stress the need for better prevention and, in general, adherence that has a direct positive effect on lowering the risks of hospitalisation and developing or worsening chronic conditions.

In this scenario, adherence is a widely addressed use case for digital-driven innovation and an object of great expectations about the potentially positive impact, specifically for consumer applications targeting the general public. While apparently simple, this class of applications presents unique challenges based on a combination of competing requirements, like the need to be easy and practical yet precise, informative yet resilient to heterogeneous sources and spot use, and general yet tailored to individual needs. Furthermore, an essential aspect of their design concerns a strong involvement and collaboration of specialists in developing the recommendation mechanisms, analysing and addressing potential risks related to wrong recommendations and validating these technologies for well-being applications or to be certified as medical devices.

These challenges translate into well-known issues for the semantic web community, specifically in relation to data integration and ontology alignment, data ingestion for knowledge graphs and reasoning.

In this paper, we present MatKG and HeLiFit (**He**alth, **Li**festyle and **Fit**ness), which consist of a pipeline and an extended ontological model for knowledge graph construction in the healthcare domain. Specifically, MatKG uses RDF Mapping Language (RML) to ingest data from heterogeneous sources, like medical health records (MHR), data streams from IoT and wearable devices, self-assessment data and surveys from apps in native and standard FHIR format[1]. The resulting knowledge graph is used to generate real-time personalised recommendations (digital coaching) for lifestyle interventions, using the RDFox reasoner[2] with validated medical guidelines also implemented using the HeLiFit ontology. This solution is been developed as a part of the H2020 GATEKEEPER infrastructure in collaboration with five healthcare organisations across Europe.

We released the MatKG pipeline as open source and can be downloaded from Github [3] under the **MIT** License. In addition, the latest version of HeLiFit can be browsed and downloaded from the NCBO BioPortal [4] with **CC BY 4.0 DEED** License.

In summary, the main contributions of this paper are:

- MatKG, the pipeline for knowledge graph generation from health medical records, medical devices, wearables, surveys, IoT and other sources related to primary and secondary care and well-being;
- HeLiFit ontology (v2.0.0), a major expansion that takes into account stream and aggregated data as well as an extended set of applications;
- mapping rules from HL7 FHIR to the HeLiFit ontology;
- a reasoner for HeLiFit knowledge graph for generating health and wellbeing recommendations for digital coaching applications;
- applicative scenarios for demonstrating the benefits of MatKG as well as preliminary results about its use and impact.

The next section provides a description of the current situation with a focus on mappings, ontologies, and healthcare reasoning. Section 3 describes the methodology followed, emphasizing the background of MatGK applications on digital coaching and lifestyle interventions. Section 4 introduces the HeLiFit ontology, highlighting the major changes in its latest version and describes the pipeline and RML mappings from FHIR to HeLiFit, with a focus

[1] The HL7 FHIR https://www.hl7.org/fhir/
[2] The Oxford Semantic Technologies RDFox https://www.oxfordsemantic.tech/
[3] https://github.com/GATEKEEPER-OU/samsung-matkg
[4] https://bioportal.bioontology.org/ontologies/HELIFIT

on the practical challenges we encountered in parsing the FHIR JSON serialization using JSONPath, maintainability of mappings and testing of the pipeline. Lastly, we describe the reasoning for performing recommendations on nutrition and physical activities. Following, Section 5 presents the approach to ongoing and future evaluation and its results. Furthermore, we briefly discuss the early impact of MatKG. Lastly, Section 6 discusses open issues and future directions.

2. State of the Art

We review state of the art according to three key aspects: i) the integration of heterogeneous sources, ii) the design of an ontology, connecting sources, standards and application scenarios, and iii) the generation of recommendations based on validated clinical guidelines based on the different data source configurations.

2.1. Integration of heterogeneous sources

Integrating data from heterogeneous sources has been facilitated by the systematic use of schema mappings [1, 2]. Usually, these consist of sufficient specification to transform each instance of the source schema, typically raw data, into an instance of the target schema with the same meaning [3]. The mappings are typically expressed in declarative languages using logical formalism that are chosen considering two criteria: i) the expressive power and ii) desirable structural properties, including generating URIs, generating blank nodes, expressing paths, If-Conditional with regular expression and equality and others. In the literature, we can find a number of semantic web approaches for mapping raw data and then building knowledge graphs [4–6]. For instance, we can find IoT-Stream [7], which is a lightweight ontology integrating IoT data streams. This approach consists of lightly annotating all the data streams so that they can be easily and quickly (in near real-time) integrated.

Similarly, there is the Semantic Web of Things (SWoT), supporting a wide-scale integration and composition by annotating the representation of both digital and physical things from the Web [8]. Another research effort is SOSA, a lightweight ontology for sensors, observations, samples and actuators [9]. SOSA offers a flexible and coherent framework for representing the entities and relations involved in sensing, sampling, and actuation.

The application scenarios previously described (see Section 3.1) involve a combination of health medical records and IoT devices both in standard and proprietary formats. Overall, the pipeline aims to extract health-related information regardless of the type of source, granularity and time-frequency. Furthermore, future applications will likely involve a wider range of data sources about the life and behaviour of users.

In this view, we decided to use the RDF Mapping Language [5] (RML) that is domain agnostic. As one of the mainstream approaches for RDF Mapping, RML is a modular, interoperable, and extensible mapping language which supports the definition of mapping rules from heterogeneous data structures for serializations into RDF.

2.2. Ontologies for digital coaching

With regard to health coaching, in literature, we can find a number of ontologies for motivational messages [10] and recommendations for minimising sedentary time [11].

Pratiwi et al. [12] propose an ontology that supports the analysis of user profiles (demographic, health, impairment and preference) and provides coaching recommendations. Another work comes from Chatterjee et al. [11], who designed an ontology that integrates daily activity level classification results, personal preferences, and recommendation messages with its content and intent. This ontology supports recommendations for maximising individuals' physical activity. However, this ontology does not take into account health medical records.

Other efforts are the Physical Activity Ontology [13] (PACO), which is designed to study the interoperability of physical activity data, and HeLis ontology [14], involves both physical activity and nutrition to monitor unhealthy behaviours. However, they have knowledge and concept limitations due to the implementation of the old version of WHO guidelines, and the task-oriented application of the knowledge graph. This limitation can also be observed in other top-level ontologies, such as DOLCE [15], CIDOCCRM [16], and SUMO [17], which aim to maximize interoperability and facilitate the design of an approach based on ontology patterns.

Considering the aforementioned solutions and their limitations, we developed HeLiFit Ontology, a domain ontology that models the key concepts and properties of healthcare knowledge systems, including healthcare monitoring, nutrition, and physical activities. The application scenarios of HeLiFit Ontology are based on the latest version of WHO guidelines, which has been endorsed by WHO and the American College of Sports Medicine (ACSM) since the 1980s [18].

2.3. Reasoning for healthcare

Reasoning in healthcare mainly focuses on risk detection [19], early intervention [20], preliminary diagnosis [21]. A recent effort brought to the development of Do-Care, a rule-based monitoring system supporting the supervision and follow-up of patients suffering from chronic diseases [22]. After collecting data from heterogeneous sources (e.g., wearables, nearable or usable), the system determines the patient's health condition as *normal*, *abnormal* or *wrong*.

Another approach is Hapicare, a monitoring system with self-adaptive coaching based on inference rules in Bayesian belief network [23]. The system performs inferences based on the Semantic Sensor Network (SSN) and SNOMED-CT ontologies. Moreover, it leverages patient historical records and contextual information from sensors.

Other reasoning approaches present challenges at this stage of the project and requirements around safeguarding: i) lack of available data required for training and testing a machine learning model, ii) available data not complete in terms of features and granularity required for this type of intervention, e.g., short periods of time or aggregated, iii) needs to inspect the rationale behind digital coaching recommendations (i.e., explanation), iv) needs to refine recommendations based on specialist input and v) needs to maintain recommendations aligned with ongoing medical research.

Considering that the reasoning conducted on the healthcare knowledge deals with a large amount of real data from IOT sensors and patient health records, we implement our reasoning engine based on RDFox [24], a well-established and high-performance knowledge graph and reasoner[5].

3. Methodology

This work has been developed as part of a larger initiative funded by the European Commission (EC): the H2020 GATEKEEPER project[6]. GATEKEEPER is a project with the main goal of building an open innovation ecosystem for AI services in primary and secondary healthcare. The project involves technology providers, healthcare institutions, healthcare specialists, patients, and researchers collaborating in designing and implementing AI services piloted in eight European and three Asian countries.

GATEKEEPER proposes a new approach to AI for healthcare and well-being based on the integration between healthcare medical records, medical devices, and consumer data from apps and wearables. Indeed, the underlying hypothesis is that while medical data have great precision and reliability, a breakthrough in AI capabilities also requires large datasets of behavioural data that fall outside the usual scope of the medical domain. In other words, medical records and precision measures should be complemented with data about walking and sleeping and self-assessments about other well-being aspects, like socialization and level of support. In this view, the design of the information architecture focused on the semantic interoperability of heterogeneous sources, privacy, decentralisation, and data pipelines to enable the tailoring and processing of anonymised datasets from multiple organizations and sources in compliance with European and National regulations.

The project addresses two types of AI applications: 1) behavioural interventions for chronic conditions and well-being and 2) risk assessment of worsening hospitalisation events and the emergence of new conditions. Both services are enabled by the integration of batch and real-time data from heterogeneous sources like electronic medical records, medical devices, consumer apps and devices. In this view, MatKG is the component developed specifically

[5]https://www.oxfordsemantic.tech/product

[6]The H2020 GATEKEEPER Project https://www.gatekeeper-project.eu/

Table 1
Summary of application scenarios reporting the type of data sources - Electronic Medical Records (EMR) or Personal Health Records (PHR) - ingested through MatKG and the aim of the digital coaching.

Application	Data sources	Digital coaching
Type two diabetes	EMR, PHR	Nutrition, physical activity
Mental health	EMR, PHR	Social recommendations, physical activity
Cancer survivors	EMR, PHR	Nutrition, physical activity
Active life	PHR	Sleep, physical activity
Loneliness	PHR	Social recommendations

for the integration of heterogeneous data, while the revised HeLiFit supports digital coaching, a form of lifestyle intervention for a range of different patients and users.

The intervention scenarios of MatKG designed in collaboration with the healthcare institutions involved in the project are described next. Broadly, the interventions address either the prevention or management of chronic conditions. The interventions focus on one or a combination of physical and social activities, sleep and nutrition. The differences relate to the clinical guidelines implemented as reasoning rules for the knowledge graph, tailored to the underlying conditions, frailties and risks of patient cohorts, like diabetes, cancer or mental health.

3.1. Scenarios

The ontology, pipeline and knowledge graph have been developed considering five medical use cases. We developed these use cases by leveraging the expertise of several healthcare organisations, considering the most pressing issues and aligning them with their vision of digital healthcare. The scenarios range from prevention in the general population such as mental health, active life and loneliness, to specialist services for supporting the self-management of long-term chronic conditions such as type two diabetes and cancer survivors. Table 1 reports the list of applications, types of data sources, and digital coaching recommendations that the system is used for.

Each scenario concerns specific domains and requires the development of a specific set of recommendation rules based on the type of available data, healthcare objectives and appropriate form of treatment. For instance, the “type two diabetes” and “cancer survivors” applications both involve recommendations about nutrition, but the types of recommendations follow a different logic. In the first case, diabetes patients need to: a) control/lose weight and b) keep the glycemia level within a range. The recommendations are based on combining food intake and physical activity in personal health records (PHR) like wearable sensors, while the electronic medical record (EMR) is used to scope the intervention. In the second case, the clinical objective for cancer survivors is to mitigate cancer-related symptoms like nausea, fatigue or anxiety. These symptoms are the result of both cancer and treatments. However, the strategy to be employed varies greatly based on the type of treatment received. Indeed, some patients need to lose weight gained from hormonal therapy, while others need to gain weight as a consequence of bowel resection. In this case, EMRs are used to define the goal of the coaching, while the PHR identifies the symptoms trajectories and the best corrective action.

4. Results

4.1. HeLiFit Ontology

The health knowledge graph builds on top of the HeLiFit ontology, specifically developed for digital coaching. The first initial version of the HeLiFit ontology (v1.0) has been presented in a previous work [18]. In this paper, we present the last version of the HeLiFit Ontology (v2.0.0), which we are now releasing to the wider scientific community. In this recent release, HeLiFit supports records of live data and binning[7] data, as illustrated in Figure 1 (within the dashed boxes). Specifically, HeLiFit can now model two levels of aggregations.

[7]Binning data refers to stream data, which has been aggregated in a given period of time, e.g., 10 seconds.

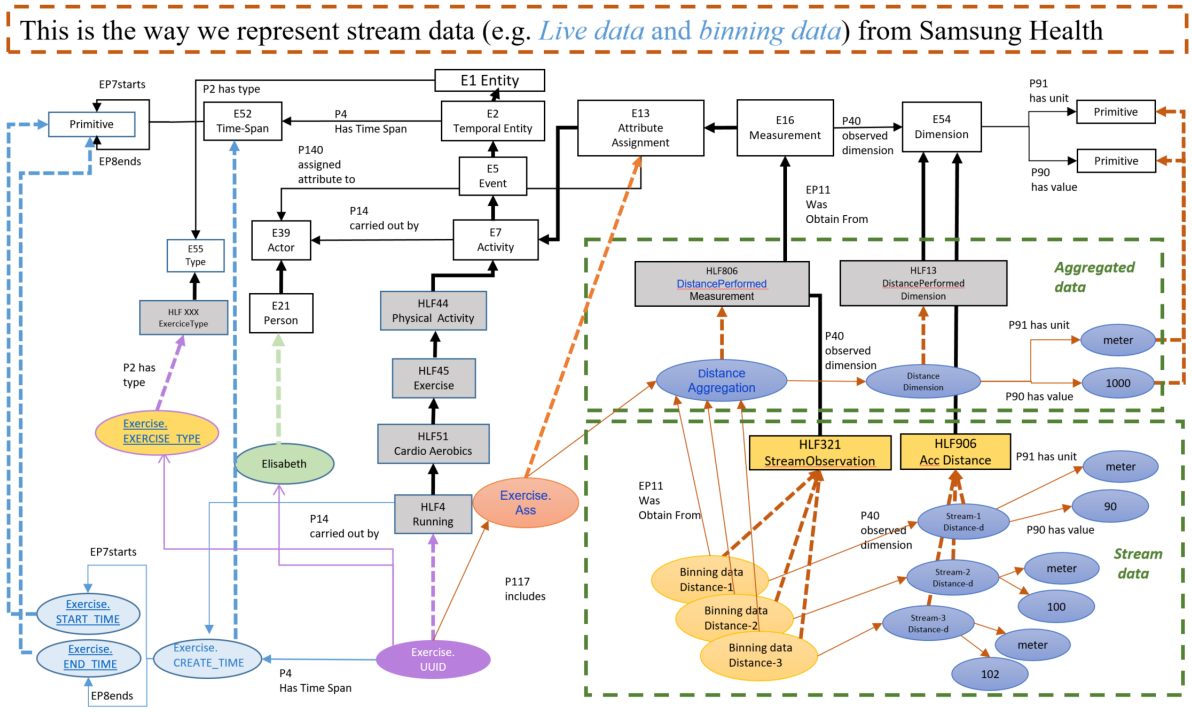

Fig. 1. Key changes to the new HeLiFit Ontology. We added stream data concepts and instances to monitor live data from sensors (dashed boxes). The aggregated data are inherited from the previous Ontology version [18].

The first level of aggregation is coarse-grained, providing statistics either through the mean values (e.g., average heart rate) or cumulative values (e.g., number of steps) during a long time span (e.g., a week).

The second level of aggregation is fine-grained, which we name stream data, for representing the data collection at a smaller time span (e.g., 10 seconds). Such stream data helps determine how the various metrics evolve over time and, hence, identify their trends. The stream data are directly collected from IoT sensors, so they contain more information than a statistical record of some measurement. This richness of information supports the healthcare recommending system in better understanding the intensity of physical activities of the users and evaluating if the physical activity is not suitable for them (leading to a high instantaneous heart rate).

The latest version of HeLiFit can be browsed on the Bioontology Portal[8].

4.2. *Processing and Pipeline*

The MatKG pipeline is a key component of a large infrastructure for the European Union healthcare data space. We chose open formats like FHIR and built an open ecosystem for other technology providers, in line with the constraints of the European funds and the vision for the European health data space that GATEKEEPER is helping to shape.

Given this premise, the rest of the section provides a schematic description of the software architecture and implementation of the pipeline that, in our opinion, best addresses the complex system of requirements emerging from stakeholders. Specifically, its design focuses on two competing factors: i) adoption of the FHIR standard in the medium-long term by both consumer and specialised technology providers (i.e., PHR and EMR), and ii) supporting

[8]https://bioportal.bioontology.org/ontologies/HELIFIT

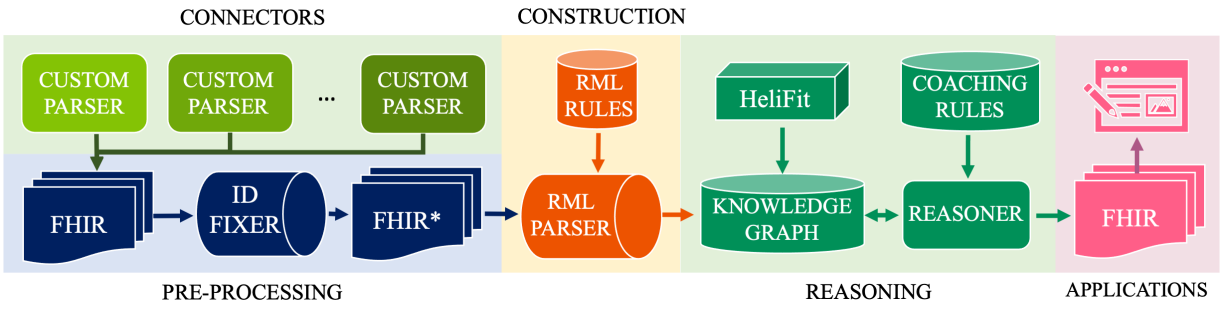

Fig. 2. The MatKG pipeline main components: 1) custom parsers for the different healthcare organisations, 2) FHIR pre-parser propagating IDs in nested documents, 3) graph constructor, and 4) reasoner that generates FHIR recommendations for, 5) digital coach applications.

the integration of new datasets from unexpected sources, e.g., new types of sensors or applications adopted by either primary, secondary carers or for self-management.

These competing needs have been addressed through a two-step processing: i) the use of a custom pre-processing parser using the IBM FHIR model[9] to convert both PHR and EMR into FHIR, and ii) an RML parser to convert FHIR into RDF triples following the HeLiFit model (see Fig. 2); described in the following subsections.

The MatKG pipeline is being developed in Java and can be downloaded from the GitHub repository[10].

4.2.1. Generating FHIR from EMR and PHR

The first step converts data from different formats in FHIR. This step has both practical and informative implications. On the one hand, it is used to identify potential cases that would require the extension of the FHIR format, contributing to the FHIR GATEKEEPER profile[11] developed during the project. On the other hand, it generates mappings that could be adopted and integrated within the originating infrastructure, expanding the adoption of the FHIR format and, therefore, interoperability of data. In this view, the first step is an interim measure used to take on board new sources, removing the bottleneck of requesting an upfront onsite FHIR conversion by new partner healthcare institutions (that may still use proprietary formats).

The processed sources include electronic medical records, i.e., records about patients' conditions and examinations, as well as personal health records coming from personal devices and apps reporting about habits, activities and patients' self-recorded assessments, e.g., symptoms management or food intake. Listing 1 is an example of an electronic medical record reporting the hypertension evaluation of a patient, which is given as input to the pipeline. Listing 2 is another example of input showing the personal health record about flights of stairs climbed in a day. Whereas an example of JSON FHIR generated output is available in Listing 3.

Listing 1 This is a sample of non-FHIR EMR reporting about the hypertension evaluation of a patient.

```
{
  "clinical_examinations": [
    {
      "patient": {
        "patient_id": "001"
      },
      "patient_age": 29,
      "date": "2021-03-30",
      "hypertension": false
    }
  ]
}
```

[9]IBM FHIR model for Java Maven https://mvnrepository.com/artifact/com.ibm.fhir/fhir-model
[10]MatKG - https://github.com/GATEKEEPER-OU/samsung-matkg.
[11]FHIR GATEKEEPER profile developed by HL7, URL: https://build.fhir.org/ig/gatekeeper-project/gk-fhir-ig/

Listing 2 A sample of non-FHIR PHR from a wearable device, recording flights of stairs climbed in a day.

```
{
  "data_source": "SH",
  "frequency": "day",
  "timestamp": "1623974400000",
  "data": [
    {
      "update_time": "1623989623494",
      "day_time": "20210618",
      "device_id": "XaIq54it7a",
      "type_id": "floorsClimbed",
      "data_uuid":"7685965849034eeeeee89",
      "values": [
        {
          "floor": "1.0",
          "start_time": "1623989575980",
          "end_time": "1623989594443",
          "time_offset": "UTC+0200"
        }
      ]
    }
  ]
}
```

In the design, we also considered using RML in this step, but this led to a series of difficulties concerning the structure and variety of documents to be considered. Indeed, RML uses JSONPath[12] to identify sub-sets of documents to be considered for conversion. JSONPath cannot express backreference and, therefore, it is not possible to convert individual measurements (like systolic and diastolic blood pressures), keeping the reference to the origin FHIR observation (blood pressure observation). This limitation required introducing a preprocessing step to down-propagate references within arrays.

4.2.2. From FHIR to RDF triples

The second step converts the data from FHIR to RDF triples, filling the knowledge graph: a one-to-one mapping between the two. This solution is used to simplify the maintainability of the pipeline. Indeed, this design choice addresses two potential challenges: a) the number of sources and b) the evolution of HeLiFit required by new application scenarios.

As discussed in the previous section, the limitations of the JSONPath regarding backreference can lead to the need for a pre-processing step. This case is limited to FHIR documents, including an array of "components" when observations like blood pressure have more than one measure (e.g. systolic e diastolic rate). To address this issue, each FHIR is run through a pre-processing step that injects the "resourceId"–a non-standard parameter–within each FHIR "component". The pre-processing propagates the reference to the observation (ID) within the RML scope for each measurement. The pre-processing also generates a second non-FHIR temporary field "systemDomainName". This field is used to generate a correctly formatted URI that combines system and code (avoiding nested URLs that would have been otherwise generated through RML transformations, like https://identifier.../https://loinc.org/...). Listing 4 presents a subset of RML rules to generate triples. Instead, Listing 5 shows a sample of the generated knowledge graph.

4.2.3. Health Knowledge Graph

The knowledge graph is the output of MatKG. MatKG uses RDFizer[13], a standalone module to convert FHIR formats to HeLiFit triples. In addition, MatKG includes unit testing using competency queries to check the consistency and validity of the pipeline and the generated knowledge graph.

The generated knowledge graph contains sensitive data about the users, which, due to privacy reasons and GDPR regulations, we cannot distribute. However, we have produced a synthetic sample of an alike knowledge graph, which can be downloaded and tested from https://doi.org/10.21954/ou.rd.21711050.

[12] JSONPath https://www.ietf.org/archive/id/draft-goessner-dispatch-jsonpath-00.html
[13] RDFizer - https://github.com/GATEKEEPER-OU/rdfizer-java

Listing 3 JSON FHIR generated from the input data in Listing 2.

```
{
    "resourceType": "Bundle",
    "type": "transaction",
    "entry": [
        {...}, // patient
        {
            "resource": {
                "resourceType": "Observation",
                "id": "123",
                "identifier": [
                    {
                        "system": "https://.../identifier",
                        "value": "Observation/123"
                    }
                ],
                "status": "final",
                "code": {
                    "coding": [
                        {
                            "system": "http://loinc.org",
                            "code": "floors_climbed",
                            "display": "FloorsClimbed"
                        }
                    ]
                },
                "subject": { ... }, // patient's ref
                "effectivePeriod": {
                    "start": "2022-12-02T12:10:43.40498Z",
                    "end": "2022-12-02T12:10:44.795442Z"
                },
                "valueQuantity": {
                    "value": 1.0,
                    "unit": "floor(s)",
                    "system": "http://unitsofmeasure.org",
                    "code": "..."
                },
                "device": { ... } //device's ref
            }
        }
    ]
}
```

4.3. Reasoning

To generate digital coaching recommendations, we used RDFox, an off-the-shelf and high-performance knowledge graph reasoner. In the following, we describe one of our case studies on physical activities and how the KG and the RDFox reasoner are currently employed.

4.3.1. Case study - WHO Rules for Physical Activity

The RDFox reasoner is driven by three rule groups written in OWL Functional Syntax: i) the HeLiFit Ontology; ii) the Physical Activity Conversion Rules; and iii) the Domain Rule Set.

The HeLiFit Ontology provides the concept of entities and relations to describe the elements defined in individual health records. The Physical Activity Conversion Rules support the analysis of physical activities taking into account the intensity, such as the energy consumption per unit time (e.g., running higher than walking), as well as the exercise categories (i.e., aerobic and anaerobic). The Domain Rule Set is interpreted from WHO guidelines, which provide a variety of codes (recommendations) based on individual circumstances, such as age, physical activities, health metrics, and so on.

When using the reasoner of health recommendation, individual data will be formalised into instances of the HeLiFit ontology. The recommender will account for users' physical activities in a certain time span (e.g., a week) and provide WHO codes as the healthcare recommendations. The data on physical activities mainly depict the status of four key metrics: frequency, intensity, modality, and duration.

The reasoning is conducted considering the observation of these metrics, which rely on the rules described in WHO guidelines, and the rules are transformed into the Datalog logistics as the input of RDFox. For instance, as

Listing 4 Example of RML rules used to generate the HeLiFit triples.

```
<#FloorsClimbedDimension> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "__RML_SRC__";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$.entry[?
        (@.resource.resourceType == 'Observation' &&
        @.resource.code.coding[0].code == 'floors_climbed')
    ]"
  ];

  rr:subjectMap [
    rr:template "https://.../{resource.code.coding[0].code}/{resource.id}";
    rr:class ho:HLF426FloorsClimbedDimension;
  ];

  rr:predicateObjectMap [
    rr:predicate ho:P91hasUnit;
    rr:objectMap [
      rml:reference "resource.valueQuantity.unit";
    ]
  ];

  rr:predicateObjectMap [
    rr:predicate ho:P90hasValue;
    rr:objectMap [
      rml:reference "resource.valueQuantity.value";
    ]
  ]; .
```

Listing 5 A sample of the HeLiFit turtle generated by MatKG.

```
<https://__base/sdn/floors_climbed> a
    <https://__base/HLF209IdentifierType>;
    <https://__base/P3hasNote> "http://loinc.org" .

<https://__base/type/E42/code/floors_climbed> a
    <https://__base/E42Identifier>;
    <https://__base/P2hasType> <https://__base/sdn/floors_climbed>;
    <https://__base/P3hasNote> "floors_climbed" .

...

<https://__base/.../id/7685965849034eeeeee89> a
    <https://__base/HLF331FloorsClimbedMeasurement>;
    <https://__base/L12happenedOnDevice>
        <https://__base/type/D8/id/XaIq54it7a>;
    <https://__base/P14carriedOutBy>
        <https://__base/id/user12%40puglia.gatekeeper.com>;
    <https://__base/P1isIdentifiedBy>
        <https://__base/type/E42/code/floors_climbed>;
    <https://__base/P40observedDimension>
        <https://__base/type/HLF426/code/floors_climbed/id/7685965849034eeeeee89>;
    <https://__base/P4hasTimeSpan>
        <https://__base/type/E52/code/floors_climbed/id/7685965849034eeeeee89>.

<https://__base/type/HLF426/code/floors_climbed/id/7685965849034eeeeee89> a
    <https://__base/HLF426FloorsClimbedDimension>;
    <https://__base/P90hasValue> "1";
    <https://__base/P91hasUnit> "floor(s)".
```

shown in Table 2, the intensity (e.g., sedentary or vigorous) can be evaluated from a range of objective measures, including the number of steps per minute.

Under the umbrella of the healthcare knowledge graph, all rules contained in the WHO guidelines are represented by Physical Activity Conversion Rules and Domain Rule Set so that the recommender is enabled to reason on input individual health records and select WHO codes for healthcare recommendations.

Table 2
Level of Intensities of Physical Activity.

Intensity Category	MET (M)	HR_{max} (H)	HRR (R)	VO_{2max} (V)	Steps per minute (S)
Sedentary	$M < 1.6$	$H < 40\%$	$R < 20\%$	$V < 20\%$	$S < 119$
Moderate	$3 < M < 6$	$55\% < H < 70\%$	$40\% < R < 60\%$	$40\% < V < 60\%$	$119 < S < 123$
Vigorous	$6 < M < 9$	$70\% < H < 90\%$	$60\% < R < 80\%$	$60\% < V < 80\%$	$137.8 < S < 140.7$
High	$9 < M$	$90\% < H$	$80\% < R$	$80\% < V$	$140.7 < S$

Listing 6 The code for sedentary behaviour recognition.

```
PREFIX sh: <https://opensource.samsung.com/projects/helifit/>

sh:IntensityType[?person, sh:sedentarybehavior]:-
    sh:E21Person[?person], sh:HLF152AgeAssignment[?ageAss],
    sh:P140assignedAttributeTo[?ageAss, ?person],
    sh:P04hasAge[?ageAss, ?age], FILTER(?age<=64 && ?age>=18),
    sh:HLF44PhysicalActivity[?PA], sh:P14carriedOutBy[?PA, ?person],
    sh:P117includes[?PA, ?PAAss], sh:E13AttributeAssignment[?PAAss],
    sh:P07personHasPAStepsTotal[?person, ?PAStepsTotal],
    sh:E52TimeSpan[?TS], sh:P4hasTimeSpan[?person, ?TS],
    sh:P03hasSpanValue[?TS, ?recommendSpan],
    FILTER(?PAStepsTotal <= 26775*?recommendSpan) .
```

Table 3
Summary of ongoing pilots including target users and advancements.

Application	Target users	Advancements
Type two diabetes	100	Validation of rules and ongoing evaluation of the recommendations
Mental health	10,000	Development of recommendation rules
Cancer survivors	100	Development of recommendation rules
Active life	9,500	Ongoing evaluation of recommendations
Loneliness	80	Validation of rules

5. Evaluation

5.1. Early Impact

The knowledge graph is currently being piloted by five healthcare organisations in the UK, Italy, Spain, and Germany. The piloting involves testing the digital coaching recommendations and their evaluation by trained doctors. In parallel, this provides the opportunity to stress the pipeline with real-time data and ontology through the assessments and demands from the clinical staff concerning data and recommendation analytics.

Each of the scenarios presented in Table 1 involves both an expert team and a cohort of patients generating data that is ingested in the knowledge graph and used to generate recommendations. Table 3 reports the advancement of each study in terms of the evaluation of the system.

At the moment of writing, the pipeline is yet to be included in a service adopted by a healthcare provider. Indeed, this step will be the final result of the piloting and the necessary process of certification of the digital coaching solution as a medical device. However, as a well-being application, the pipeline and ontology have been the object of several requests from Samsung teams working on the Samsung Health suite. This knowledge transfer from the Research & Development branch and collaboration with research institutions to the production teams is the first evidence of the potential impact of the presented solution.

We adopted a unit-test approach that evaluates the pipeline components. Specifically, the evaluation uses test data to check whether the behaviour of components is as expected in terms of data conversion, comparing the correctness of output against a pre-computed and manually validated result. Technically, the unit-test output and the validated data are both converted into hash strings, and the test fails if the hash does not perfectly match. It is worth noticing

that this approach does not just test the generation of the expected entries (e.g., FHIR observations) but excludes the generation of unexpected facts.

We tested the full pipeline using 35 competency questions, which we converted into SPARQL and ran on MatKG. Listing 7 shows the SPARQL query for testing the question "List all the physical activities performed by individuals in a particular time span." The full list of competency questions is available within the MatKG repository[14] on GitHub.

Listing 7 SPARQL query for the competency question "List all the physical activities performed by individuals in a particular time span."

```
prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix helifit: <https://opensource.samsung.com/projects/helifit/>

SELECT DISTINCT ?physicalAC WHERE {

  ?physicalAC rdf:type helifit:HLF49MuscleStrengtheningAnaerobics .
  ?phyActivity helifit:P14carriedOutBy ?ind .
  ?phyActivity helifit:P4hasTimeSpan  ?timeSpan .
  ?ind rdf:type helifit:E21Person .
  ?ind helifit:P1isIdentifiedBy ?userID .
  ?timeSpan helifit:EP9effectiveDatatime ?time .
  FILTER (?time > "2022-01-28T17:17:48Z"^^xsd:dateTime &&
  ?time <= "2022-02-28T17:17:48Z"^^xsd:dateTime)
}
```

The next release will include a test dataset and a similar unit-test approach to check the result of the queries against pre-computed, manually validated, and hashed results.

5.2. Sustainability Plan

Since 2022, both MatKG and HeLiFit have been core products of the Health Innovation team at Samsung Research and part of the Samsung Health app that reaches millions of customers. It is Samsung's priority to maintain and update both the ontology and the pipeline. Part of our team is allocated to expanding the MatKG pipeline with new additional parsers as well as an efficient way of ingesting knowledge graphs and reasoning. Another section of the team is working on the ontology, including new use cases.

Future use cases should focus on diseases that have a tradition of well-defined clinical guidelines that suggest user-specific behaviours belonging to many domains, such as:

- timely measurement of physiological parameters, e.g. daily measurement of blood pressure for patients with hypertension
- medication use, e.g. intake of specific drugs for patients with COPD (Chronic Obstructive Pulmonary Disease) or calcium or Vitamin D for patients with osteoporosis
- specific dietary prescriptions, e.g. number of calories per day for patients with obesity
- dietary restrictions, e.g. for patients with heart failure or with chronic kidney disease
- avoidance of triggers as in the case of irritants in the air for patients with asthma

We plan to have yearly evaluations to assess the quality of the recommendations and the uptake from the users and expand the HeliFit and MatKG application range.

6. Conclusions

The design, development and use of MatKG and, in general, the GATEKEEPER project showcased the crucial roles and strengths of the semantic web in AI development and complex inter-organizational services. We had the

[14] GitHub folder containing the full list of competency questions https://github.com/GATEKEEPER-OU/samsung-matkg/tree/develop/src/main/resources/queries

opportunity to identify issues and opportunities concerning the engineering of MatKG, specifically concerning the use of RML with real medical and IoT data and the use of FHIR records for data ingestion in MatKG.

The presented resource is effective in addressing several key challenges of digital healthcare. Indeed, the digital coaching system can implement clinically validated rules and it is able to "explain" the rationale behind each recommendation. Furthermore, the pipeline and knowledge graph provide effective solutions easy to expand and maintain for the integration of heterogeneous data sources ranging from electronic medical records and healthcare devices to consumer devices and apps.

Breaking the data silos of health and consumer data was identified in previous projects like the H2020 ActiveAge[15] as a key enabler for AI breakthrough in digital healthcare. The logic behind this is that consumer data would provide missing context about patients' lifestyles, complementing EMR and data from medical devices. In this view, the presented pipeline was designed in line with this vision: medical goals to be instantiated dynamically on a heterogeneous data space unique per cohort and individuals.

Arguably, the presented solution is a step forward in the right direction. However, the adopted approach to automatic reasoning presents well-known rigidness results of how rules are implemented: crisp thresholds and evaluation of conditions, like the number of minutes of activity and classification of users as active or not (see the first publication on WHO rules for an extensive description [18]).

This rigidity can be addressed through different strategies, like combining semantic reasoning with machine learning or by using fuzzy sets and fuzzy logic. Addressing this rigidity is necessary for real-world applications, in particular to prevention in the general population that is not following a strict medical regime. Indeed, from our experience, the flexibility of reasoning is critical in motivating users to achieve goals in a time frame and pace that works for each person. In other words, success or failure should be calculated considering progress toward a goal rather than the achievement of the goal itself.

It is worth noting that novel approaches for the automatic extraction of rules from narrative medical guidelines are emerging. Some applications of Large Language Models are going in this direction [25]. The comparison of these innovative approaches with the one described in our paper can be fruitfully explored in the future.

Lastly, in future applications, we will investigate how to integrate an AI-supported generation and validation of RML mappings to FHIR format. Indeed, the maintenance of mappings is one of the open issues as a significantly labour-intensive task that may be a barrier to scaling up this approach to a large number of use cases.

References

[1] B. ten Cate and P.G. Kolaitis, Structural characterizations of schema-mapping languages, *Communications of the ACM* **53**(1) (2010), 101–110.

[2] A. Chessa, G. Fenu, E. Motta, D.R. Recupero, F. Osborne, A. Salatino and L. Secchi, Enriching Data Lakes with Knowledge Graphs, in: *1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge, TEXT2KG 2022 and MK 2022*, 2022. http://oro.open.ac.uk/83013/.

[3] H. Kondylakis, M. Doerr and D. Plexousakis, Mapping language for information integration, *ICS-FORTH Technical Report* **385** (2006).

[4] L.E.T. Neto, V.M.P. Vidal, M.A. Casanova and J.M. Monteiro, R2RML by assertion: A semi-automatic tool for generating customised R2RML mappings, in: *Extended Semantic Web Conference*, Springer, 2013, pp. 248–252.

[5] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens and R. Van de Walle, RML: a generic language for integrated RDF mappings of heterogeneous data, in: *Ldow*, 2014.

[6] N. Minadakis, Y. Marketakis, H. Kondylakis, G. Flouris, M. Theodoridou, G. de Jong and M. Doerr, X3ML Framework: An Effective Suite for Supporting Data Mappings., in: *EMF-CRM@ TPDL*, 2015, pp. 1–12.

[7] T. Elsaleh, S. Enshaeifar, R. Rezvani, S.T. Acton, V. Janeiko and M. Bermudez-Edo, IoT-Stream: A lightweight ontology for internet of things data streams and its use with data analytics and event detection services, *Sensors* **20**(4) (2020), 953.

[8] A.J. Jara, A.C. Olivieri, Y. Bocchi, M. Jung, W. Kastner and A.F. Skarmeta, Semantic web of things: an analysis of the application semantics for the iot moving towards the iot convergence, *International Journal of Web and Grid Services* **10**(2–3) (2014), 244–272.

[9] K. Janowicz, A. Haller, S.J.D. Cox, D. Le Phuoc and M. Lefrançois, SOSA: A lightweight ontology for sensors, observations, samples, and actuators, *Journal of Web Semantics* **56** (2019), 1–10. doi:https://doi.org/10.1016/j.websem.2018.06.003. https://www.sciencedirect.com/science/article/pii/S1570826818300295.

[15] EU Horizon 2020 project "ACTivating InnoVative IoT smart living environments for AGEing well" https://cordis.europa.eu/project/id/732679/

[10] C. Villalonga, H.o. den Akker, H. Hermens, L.J. Herrera, H. Pomares, I. Rojas, O. Valenzuela and O. Banos, Ontological modeling of motivational messages for physical activity coaching, in: *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2017, pp. 355–364.
[11] A. Chatterjee, N. Pahari, A. Prinz and M. Riegler, Machine learning and ontology in eCoaching for personalized activity level monitoring and recommendation generation, *Scientific Reports* **12**(1) (2022), 1–26.
[12] P.S. Pratiwi, Y. Xu, Y. Li, S.G. Trost, K.M. Clanchy and D.W. Tjondronegoro, User profile ontology to support personalization for e-Coaching systems, in: *Proceedings of the CIKM 2018 Workshops: co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, Sun SITE Central Europe (CEUR), 2018.
[13] H. Kim, J. Mentzer, R. Taira et al., Developing a physical activity ontology to support the interoperability of physical activity data, *Journal of medical Internet research* **21**(4) (2019), e12776.
[14] M. Dragoni, T. Bailoni, R. Maimone and C. Eccher, HeLiS: an ontology for supporting healthy lifestyles, in: *International semantic web conference*, Springer, 2018, pp. 53–69.
[15] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider, Sweetening ontologies with DOLCE, in: *International conference on knowledge engineering and knowledge management*, Springer, 2002, pp. 166–181.
[16] M. Dörr, The cidoc crm-an ontological approach to semantic interoperability of metadata, 2001, *AI Magazine, Special Issue on Ontologies* (2002).
[17] I. Niles and A. Pease, Towards a standard upper ontology, in: *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, 2001, pp. 2–9.
[18] C. Allocca, S. Jilali, R. Ail, J. Lee, B. Kim, A. Antonini, E. Motta, J. Schellong, L. Stieler, M.S. Haleem et al., Toward a Symbolic AI Approach to the WHO/ACSM Physical Activity & Sedentary Behavior Guidelines, *Applied Sciences* **12**(4) (2022), 1776.
[19] S. Mishra, H.K. Thakkar, P.K. Mallick, P. Tiwari and A. Alamri, A sustainable IoHT based computationally intelligent healthcare monitoring system for lung cancer risk detection, *Sustainable Cities and Society* **72** (2021), 103079.
[20] A.C. Morales Tirado, E. Daga and E. Motta, Reasoning on Health Condition Evolution for Enhanced Detection of Vulnerable People in Emergency Settings, in: *Proceedings of the 11th on Knowledge Capture Conference*, K-CAP '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 9–16–. ISBN 9781450384575. doi:10.1145/3460210.3493551.
[21] A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, M.S. Riaz, K. Ali, C.N. John, M.I. Hussain and M. Nabeel, AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, *Informatics in Medicine Unlocked* **20** (2020), 100378.
[22] H.B. Elhadj, F. Sallabi, A. Henaien, L. Chaari, K. Shuaib and M. Al Thawadi, Do-Care: A dynamic ontology reasoning based healthcare monitoring system, *Future Generation Computer Systems* **118** (2021), 417–431.
[23] H. Kordestani, R. Mojarad, A. Chibani, A. Osmani, Y. Amirat, K. Barkaoui and W. Zahran, Hapicare: A healthcare monitoring system with self-adaptive coaching using probabilistic reasoning, in: *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2019, pp. 1–8.
[24] Y. Nenov, R. Piro, B. Motik, I. Horrocks, Z. Wu and J. Banerjee, RDFox: A highly-scalable RDF store, in: *International Semantic Web Conference*, Springer, 2015, pp. 3–20.
[25] M.C. Schubert, W. Wick and V. Venkataramani, Large Language Model-Driven Evaluation of Medical Records Using MedCheckLLM, *medRxiv* (2023). doi:10.1101/2023.11.01.23297684.