BROKERAGE AND MARKET PLATFORM

FOR PERSONAL DATA

D3.4 Final Data model and ledger for biomedical marketplace

www.krakenh2020.eu

This project has received funding from the European Union's Horizon 2020 (H2020) research and innovation programme under the Grant Agreement no 871473

D3.4 Final Data model and ledger for biomedical marketplace

Grant agreement	871473
Work Package Leader	ATOS
Author(s)	Davide Zaccagnini (LYN)
Contributors	George Prikamenos (LYN), Stefan More (TUG)
Reviewer(s)	Rob Holmes (TEX), Silvia Gabrielli, Michele Marchesoni (FBK)
Version	Final
Due Date	31/05/2022
Submission Date	23/05/2022
Dissemination Level	Public

Release History

Version	Date	Description	Released by
v0.1	18/03/2022	Initial version (ToC and initial text)	Davide Zaccagnini
v0.2	11/05/2022	Draft	Davide Zaccagnini
v0.3	16/05/2022	Draft revised	Silvia Gabrielli
v0.4	17/05/2022	Draft revised	Rob Holmes
v0.5	18/05/2022	Final Draft for submission	Davide Zaccagnini
v1.0	23/05/2022	Submitted version	Atos

Table of Contents

List of Figures

List of Acronyms

Acronym	Description
API	Application Programming Interface
CSV	Comma Separated Values
GDPR	General Data Protection Regulation
KCIT	KRAKEN Company Identification Tool
MeSH	Medical Subjects Heading
MPC	Multi Party Computation
NLM	National Library of Medicine
RDF	Resource Description Framework
URI	Uniform Resource Identifier
SSI	Self-Sovereign Identity
WP	Work Package

Executive Summary

The KRAKEN data model is a key component of the marketplace and the platform at large, linking and standardizing key information about users and data products across workflows and components of the platform. Key applications of this module include the tagging of Data Products, driving in turn their search and browsing on the platform front-end, the computation of data access permissions and the execution of computations from the MPC component.

1 Introduction

1.1 Purpose of the document

This document accompanies and provides an overview of the final version of the marketplace data model, and serves as a reference for consortium members to complete the integration of the platform in the last part of the project and for document readers to gain familiarity with this crucial component of the KRAKEN platform.

This document also extends the previously submitted D3.3[1], Data model and ledger for biomedical marketplace, First release.

1.2 Structure of the document

Building on D3.3, the first version of this document, this update presents, from that previous deliverable and for the sake of readability, an overview of the data model and ledger which hasn't changed since the submission of the previous deliverable. The following sections explain recent integrations and extensions focusing specifically on the standardization and processing of extended data access parameters defined after the second legal review of the platform, the handling of verified credentials' parameters for the KCIT developed in WP3, and the use of standardized medical and educational terms for the execution of distributed queries by the MPC analytical system.

[1] D3.3 Data model and ledger for biomedical marketplace, First release, July 2021

2 System overview (Excerpts from D3.3)

The two domain specific models cover respectively biomedical and educational Data Products. The main goal here has been to provide the highest level of data harmonisation and interoperability, not only within the platform but with external data ecosystems so to allow ease of use and accessibility by disparate types of users and data sellers fostering integration with the broader data landscape.

The biomedical space has over the years consolidated and established few terminologies that have become the de facto standards in healthcare. Among these, the Medical Subjects Heading serve as a bridge that maps these main references into an extensive semantic structure (ontology) used also to integrate data in various domains. It also provides multilingual capabilities so that data from different geographies and jurisdictions can be mapped into the KRAKEN marketplace model. No such references exist in the case of educational information. This forced the educational pilot teams to create an ad-hoc terminology including all the major concepts and terms utilised in the exchange of these type of data. Due to the much smaller scope of educational information compared to biomedical information, the ad hoc solution proved so far sufficient.

2.1 The MeSH Ontology

Medical Subject Headings (MeSH https://www.nlm.nih.gov/mesh/intro_trees.html) is a hierarchically organized terminology for indexing and cataloging biomedical information such as MEDLINE/PubMed and other NLM (National Library of Medicine) databases. It was introduced in 1960 by the then NLM director, Frank B. Rogers.

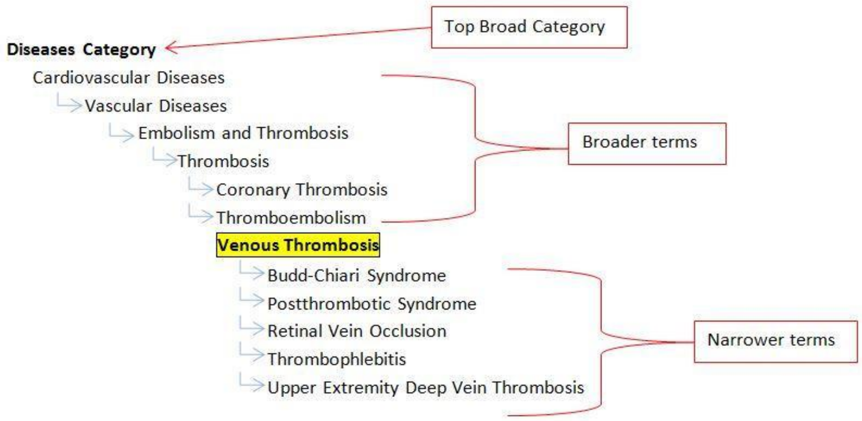
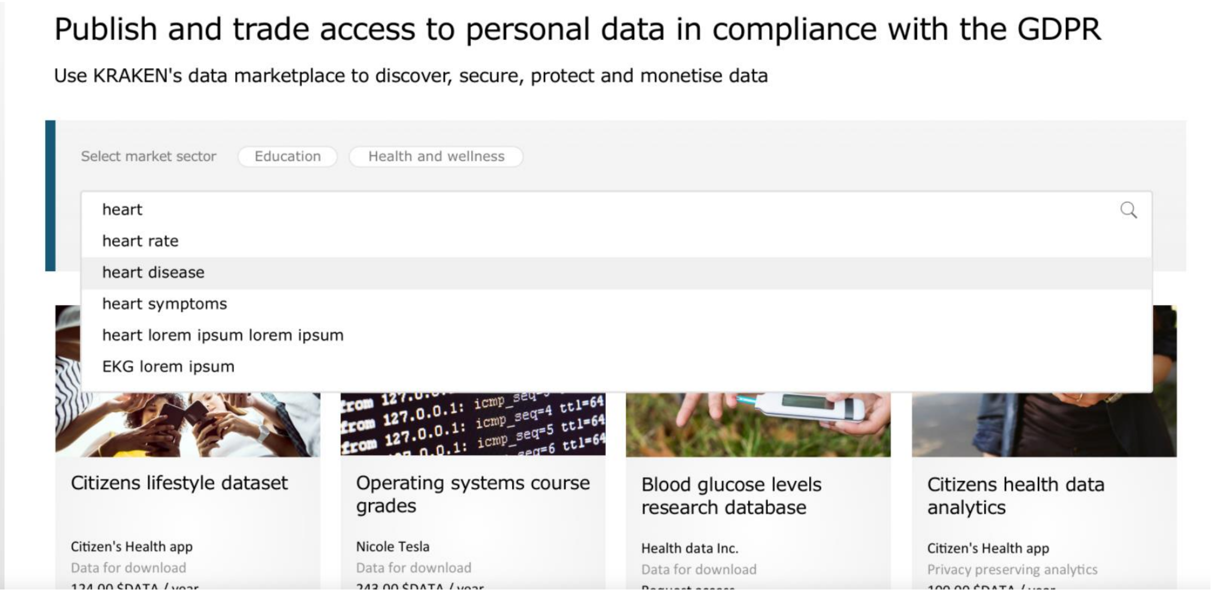


Figure 1 Example of MeSH term hierarchy



Publish and trade access to personal data in compliance with the GDPR

Use KRAKEN's data marketplace to discover, secure, protect and monetise data

heart
heart rate
heart disease
heart symptoms
heart lorem ipsum lorem ipsum
EKG lorem ipsum

Citizens lifestyle dataset
Citizen's Health app
Data for download
124.00 $DATA / year

Operating systems course grades
Nicole Tesla
Data for download
243.00 $DATA / year

Blood glucose levels research database
Health data Inc.
Data for download
Request access

Citizens health data analytics
Citizen's Health app
Privacy preserving analytics
100.00 $DATA / year

Figure 2 The MeSH ontology supporting semantic search in the marketplace

2.2 The Educational data model

Educational data are encoded and exchanged as Verified Credentials. The KRAKEN marketplace has been, in this view, connected to the Graz university information system. This is done using a connector component integrated in the university system. Using this connector component, students authenticate and export credentials into their KRAKEN mobile wallet. Those credentials are cryptographically signed by the university and prove education achievements such as source credentials and diplomas.

Using these exported credentials, students then use the KRAKEN marketplace components to create a Data Product with their data. This Data Product can be either a direct offer of their credentials, or an offer to use the education data in a Data Product containing aggregated student data accessed by privacy preserving analytic computations.

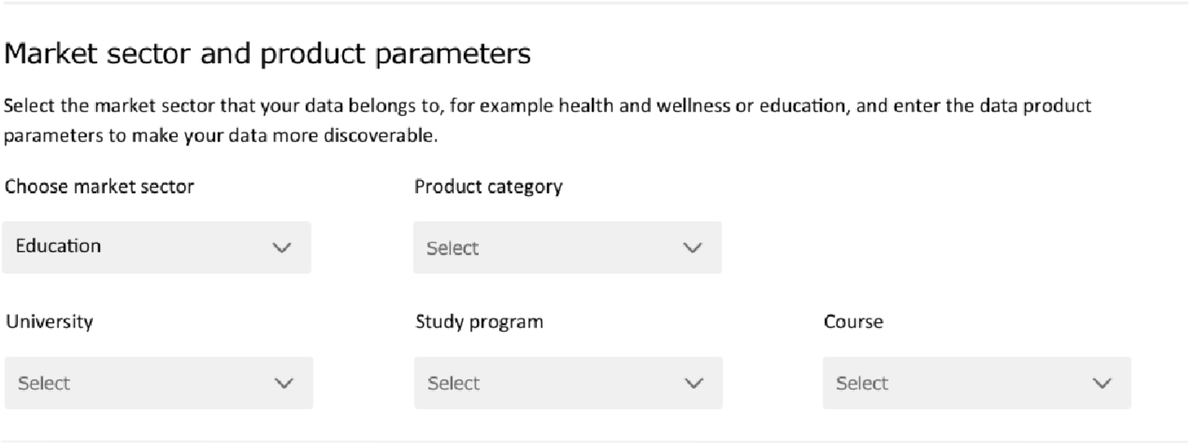
To enable querying for specific data by data buyers as well as to allow computations on multiple data sets, those education Data Products follow a simple data model:

- University Name
    - Study Program Identifier (name, number)
    - Course Identifier (name, number)

The attributes are organized in a hierarchical manner. While study program and course are in a level below the university (thus, they depend on the selected university), the two attributes are on the same level in the hierarchy. This is because most courses are part of multiple study programs and can often be freely assigned to a study program.

It is important to note that the values of the attributes depend on the university and the KRAKEN system does not use a global unified data model for courses and study programs as this information is currently not available for the used exported data.

This data model can in the future be extended by further attributes needed in upcoming use cases.



Market sector and product parameters

Select the market sector that your data belongs to, for example health and wellness or education, and enter the data product parameters to make your data more discoverable.

Figure 3: The Educational Data Model selectors in the "Create Data Product"

2.3 Interfaces

In KRAKEN the MeSH ontology defines the set of terms that can be used to categorize Data Products belonging to the health pilot in the marketplace. The categorization of Data Products happens through one or multiple tags.

The MeSH API (Application Programming Interface) is called on two occasions during the data publication process:

1. On the frontend as a suggestion tool to offer options to the user to choose a term belonging to the ontology. For this call, the entry point being used is:

   "https://id.nlm.nih.gov/mesh/lookup/term?match=startswith&limit=5&label=${search}"
   Where ${search} coincides with the word provided by the user. The returned value of this call consists of a set of maximum five terms that start with ${search}, if any, belonging to the ontology.

2. On the API to check that the provided terms assigned to the Data Product belong to the ontology. The entry point exploited in this occasion is:
   "https://id.nlm.nih.gov/mesh/lookup/term?label=${tag}&match=exact&limit=1"
   Where ${tag} coincides with one of the tags chosen by the user. The returned value of this call consists of one word corresponding to ${tag} if present in the ontology.

2.4 Deployment

The software described above belongs to the KRAKEN Marketplace Frontend and marketplace API code. A description of these two components, their interfaces, deployment, source code and background technologies and tools provided in D5.5.

2.5 Source Code

The MeSH RDF has no available source code but a file in RDF N-Triples format (found at ftp://ftp.nlm.nih.gov/online/mesh/rdf/mesh.nt.gz) is provided along with a SPARQL Query editor (https://id.nlm.nih.gov/mesh/query) and a Restful API (https://id.nlm.nih.gov/mesh/swagger/ui#/sparql/sparqlQuery) hosted by the U.S. National library of Medicine.

2.6 Baseline technologies and tools

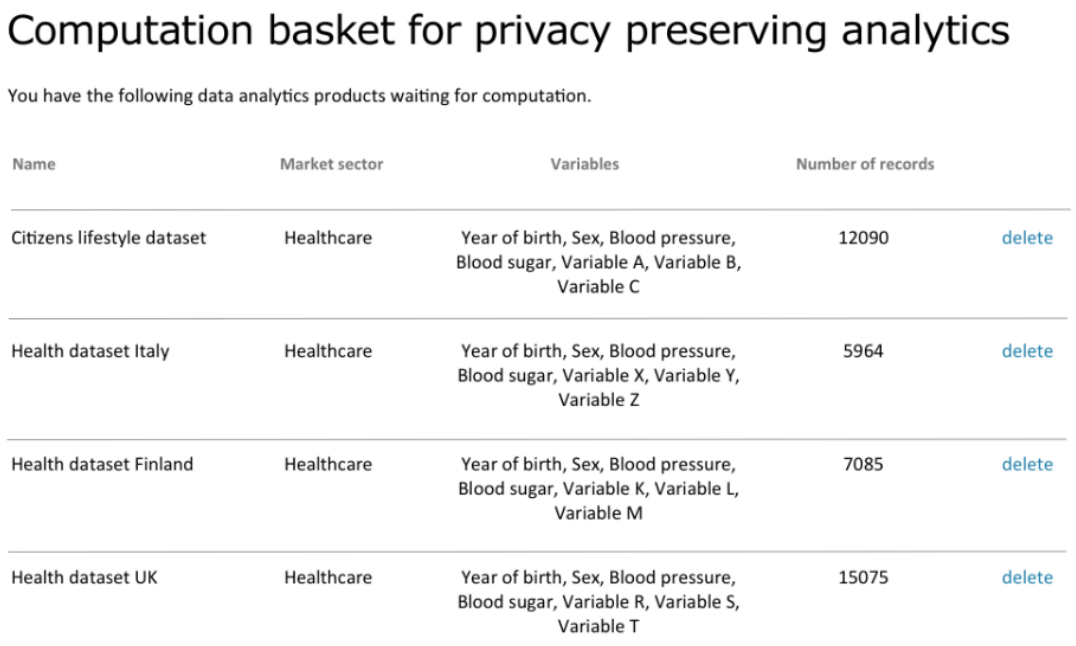
The MeSH biomedical vocabulary is made available by the U.S. NLM in Resource Description Framework (RDF) format, through the SPARQL query language and a Restful web api (see 3.4). These tools can be used to build web applications around MeSH terms. As a simple example, the lookup API supports a "match" parameter which specifies how the label parameter (e.g. typed descriptor) is matched against MeSH descriptors. This may take values "exact", "contains" and "startswith" and can be readily used to collect matching descriptors from a partial label being typed by a user and enable an autocomplete feature.

3 Extensions and new integration

During the second period of the project especially the biomedical data model was utilized to extend their representation of multiple access parameters including institutions, countries and intended users of Data Products as indicated by the legal team after the second legal review. In this regard the representation of jurisdictions as classified under the GDPR for compliant international data access was carried out through the selection of terms in the MeSh ontology along with the definitions of institutions for the handling of Verified Credentials issued by the Depute tool Developed in WP3.

MeSh terms were also used to map clinical concepts in reference databases utilized for initial testing of the MPC Analytical function and the integration of this module into the marketplace backend.

This is shown in the picture below where mostly cold data assets are harmonized through MeSh standard terms so that distributed queries can be consistently executed across multiple data sources even if their content overlaps only partially.



Computation basket for privacy preserving analytics

You have the following data analytics products waiting for computation.

Name	Market sector	Variables	Number of records	
Citizens lifestyle dataset	Healthcare	Year of birth, Sex, Blood pressure, Blood sugar, Variable A, Variable B, Variable C	12090	delete
Health dataset Italy	Healthcare	Year of birth, Sex, Blood pressure, Blood sugar, Variable X, Variable Y, Variable Z	5964	delete
Health dataset Finland	Healthcare	Year of birth, Sex, Blood pressure, Blood sugar, Variable K, Variable L, Variable M	7085	delete
Health dataset UK	Healthcare	Year of birth, Sex, Blood pressure, Blood sugar, Variable R, Variable S, Variable T	15075	delete

Figure 4 Front-end presentation of standard MeSh clinical terms

4 Conclusions and future considerations

With the submission of this document the corresponding task 3.5 is officially completed although additional testing, parameters mapping and possibly integration are foreseen, and planned for, from here to the deployment of the platform.

While the use of international standards for mapping of biomedical data (i.e., the MeSh Ontology) will greatly facilitate KRAKEN's Data Product interoperability within the broader health data ecosystem, the challenge of standardizing disparate data assets across market segments and specialties remains substantial. As of today, technologies that allow the automatic reconciliation and semantic integration of local data sources are still not viable. While these solutions are out of the scope of the project, a credible exploitation and sustainability plan for the platform will have to include the implementation of data standardization and curation services under appropriate incentives so that standardized, high quality data assets can be produced and monetized efficiently and at scale.

KRAKEN

Atos

FONDAZIONE BRUNO KESSLER

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

sic

LYNKEUS
STRATEGY CONSULTING | BLOCKCHAIN & SMART CONTRACTS | DATA ANALYTICS

XLAB

TX

KU LEUVEN CiTiP
CENTRE FOR IT & IP LAW

IAIK TU Graz

InfoCert
TINEXTA GROUP

@KrakenH2020

Kraken H2020

www.krakenh2020.eu

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871473