BROKERAGE AND MARKET PLATFORM
FOR PERSONAL DATA

D3.3 Data model and ledger for biomedical marketplace. First release

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

D3.3 Data model and ledger for biomedical marketplace, First release

Grant agreement	871473
Work Package Leader	ATOS
Author(s)	Davide Zaccagnini (LYN)
Contributors	Donato Pelegrino (TX), George Prikamenos (LYN), Stefan More (TUG)
Reviewer(s)	Rob Holmes (TEX), Silvia Gabrielli, Michele Marchesoni (FBK)
Version	Final
Due Date	31/07/2021
Submission Date	29/07/2021
Dissemination Level	Public

Release History

Version	Date	Description	Released by
v0.1	02/07/2021	Initial version (ToC and initial text)	Davide Zaccagnini
v0.2	24/06/2021	Draft with partners contributions	George Prikamenos
v0.3	26/07/2021	Draft for review	Davide Zaccagnini
v0.4	27/07/2021	Reviewed version by TEX	Rob Holmes
v0.5	28/07/2021	Reviewed version by FBK	Silvia Gabrielli
v0.6	28/07/2021	Integrated version	Davide Zaccagnini
V1.0	29/07/2021	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
MeSH	Medical Subjects Heading
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 work and data flows and front-end functionalities. Key applications include the definition of Data Products and users to compute data access permissions, but also browsing and searching on the data catalogue and mobile app. Future versions will add detail to parameters used in data access permissions and more fields to enable additional marketplace functionalities.

1 Introduction

1.1 Purpose of the document

This document accompanies and provides an overview of the first version of the marketplace data model, defining its current state as a reference for the document readers to gain familiarity with this crucial component of the KRAKEN platform. It also sets the current state of its development indicating the direction and rational for future extensions.

1.2 Structure of the document

The document is structured in 2 main sections describing the biomedical and educational data models respectively. The user-specific data model is discussed throughout the text. Aspects of deployment, source code and underlying technologies are described in corresponding sections.

2 Prototype overview

The KRAKEN data model is used to specify and manage information pertaining to both Data Products and users. It is composed of three sub-models, two of which cover Data Products, in both the biomedical and educational domains, while the third applies to users of the platform and their characteristics as data buyers and/or sellers. This information allows in turn the permissioning layer (Lynkeus blockchain) to compute data access permissions based on both users (ex. age and country) and Data Products (ex. intended uses) criteria. –All these parameters have been created in strict collaboration with WP7 over the course of multiple legal and ethical analyses.

The data models are exposed on the marketplace front-end supporting not only users and Data Products registration and tagging functionalities, but also search and browsing on the KRAKEN data catalogue.

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. Potential extensions will be performed as needed during the second part of the project.

3 The Medical Subject Headings

3.1 Description

3.1.1 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.

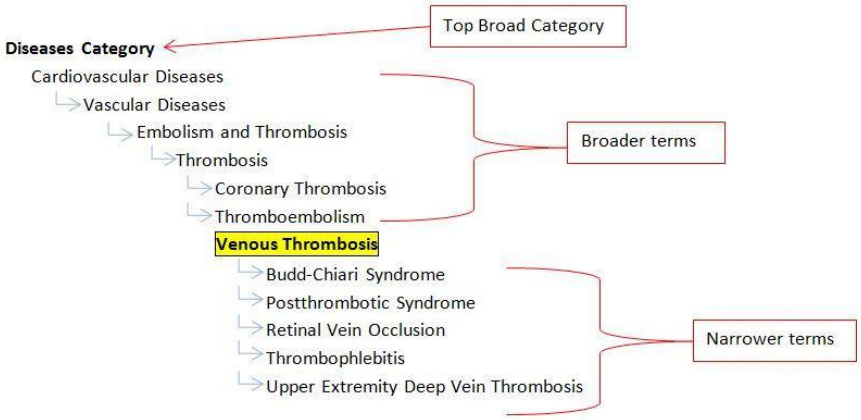

Figure 1 Example of MeSH term hierarchy

MeSH consists of descriptors/headings, qualifiers, concepts, terms, tree numbers and other identifiers.

MeSH tree numbers 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". Typically, MeSH terms also have subheadings. For example, common disease subheadings include ‘therapy’, ‘diagnoses’ and ‘etiology’ as subheadings. In more detail, MeSH terms may be thought of as nodes in the MeSH tree. Branches sprouting from that node correspond to related biomedical concepts. Because of the structure of MeSH, identifying which terms are below a certain node can be done very efficiently.

MeSH descriptors are divided into categories, namely:

Anatomy [A]

Organisms [B]

Diseases [C]

Chemicals and Drugs [D]

Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]

Psychiatry and Psychology [F]

Phenomena and Processes [G]

Disciplines and Occupations [H]

Anthropology, Education, Sociology, and Social Phenomena [I]

Technology, Industry, and Agriculture [J]

Humanities [K]

Information Science [L]

Named Groups [M]

Health Care [N]

Publication Characteristics [V]

Geographicals [Z]

In turn, these categories are subdivided into subcategories and within each category descriptors are listed in a hierarchical manner from general to specific forming the MeSH tree. When using MeSH, one should find the most specific (i.e. further away from root) MeSH descriptor to represent a concept of interest.

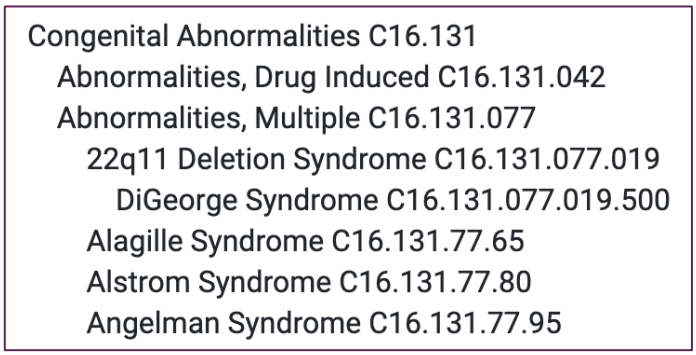

Figure 2 Part of MeSH tree with codes

3.1.2 MeSH URI Patterns

All MeSH RDF (Resource Description Framework) data is either represented as a URI (Unique Resource Identifier) reference or encoded as a literal string. NLM minted two base URIs for MeSH RDF:

- id.nlm.nih.gov/mesh/vocab# (meshv: prefix) for classes and predicates.
- id.nlm.nih.gov/mesh/ (mesh: prefix) for instances of classes.

Class instance URIs are constructed by appending id.nlm.nih.gov/mesh/ to identifiers. For example:

- Descriptor D015242: http://id.nlm.nih.gov/mesh/D015242
- Qualifier Q000008: http://id.nlm.nih.gov/mesh/Q000008
- Supplementary Concept Record C025735: http://id.nlm.nih.gov/mesh/C025735
- Concept M0000001: http://id.nlm.nih.gov/mesh/M0000001
- Term T000002: http://id.nlm.nih.gov/mesh/T000002
- Tree number A09.371.613: http://id.nlm.nih.gov/mesh/A09.371.613

Classes have a base URI of http://id.nlm.nih.gov/mesh/vocab#. For example:

- http://id.nlm.nih.gov/mesh/vocab#Descriptor
- http://id.nlm.nih.gov/mesh/vocab#Qualifier
- http://id.nlm.nih.gov/mesh/vocab#SupplementaryConceptRecord

Predicates also have a base URI of http://id.nlm.nih.gov/mesh/vocab#. For example:

- http://id.nlm.nih.gov/mesh/vocab#allowableQualifier
- http://id.nlm.nih.gov/mesh/vocab#annotation
- http://id.nlm.nih.gov/mesh/vocab#registryNumber

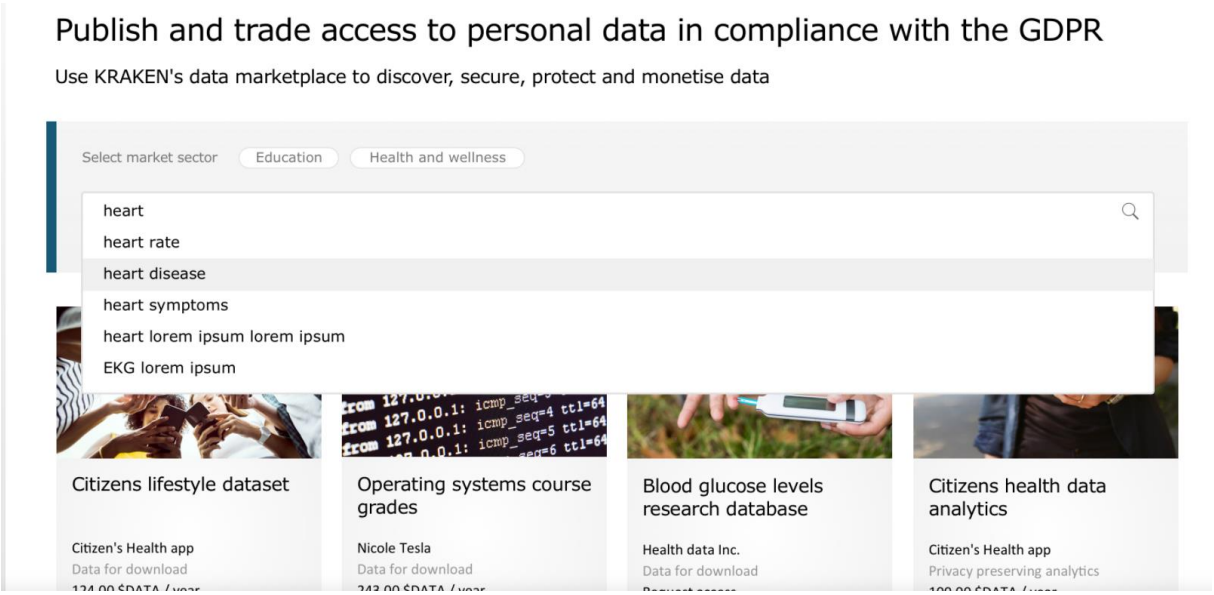

Figure 3 The MeSH ontology supporting semantic search in the marketplace

3.2 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.

3.3 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.

3.4 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.

3.5 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.

4 The Educational data model

In the context of KRAKEN's educational pilot students can export their academic information from a university system and provide access to various stakeholders such as other academic institutions, recruiters or employers. Additionally, students can use this data to create one of various Data Products on the KRAKEN marketplace.

4.1 Description

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 (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.

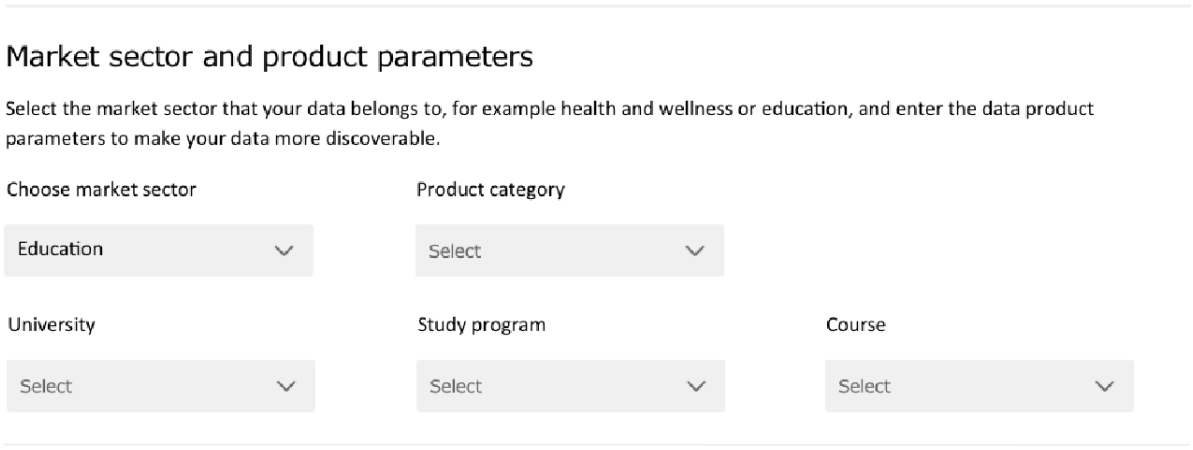

Figure 4: The Educational Data Model selectors in the "Create Data Product"

4.2 Interfaces

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

The Educational Data Model is used two times in 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. The set of terms that the user can choose from is fetched from the database.
2. On the API to check that the provided terms assigned to the Data Product belong to the ontology. The check is done by the API by looking for the provided tag in the set of available terms in the database.

4.3 Deployment

A description of these two components, their interfaces, deployment, source code and background technologies and tools will be provided in D5.5 KRAKEN marketplaces first release, due in August 2021.

4.4 Source code

The values for the data model described above are provided directly by the university from the university system. It is possible to add a dynamic data source for the values as soon as universities start to provide the required data in machine readable format.

The university connector components are available as open source from the KRAKEN GitHub organization at github.com/krakenh2020. The KRAKEN mobile app as well as its modules are currently only available at Atos' GitLab at https://scm.atosresearch.eu/ari/kraken.

4.5 Baseline technologies and tools

The data stored in the wallets of the student is encoded as an SSI (verifiable credential, a special JSON-based format. The KRAKEN marketplace uses the data encoded in the credential to create the various Data Products. This Data Products are then annotated using attributes from the data model mentioned above.

The possible values for this data model are provided by the university in a simple CSV value format and imported into the marketplace frontend. The frontend then queries the respective subset of the data and provides it to the user for selecting the matching values.

5 Conclusion

The first release of the KRAKEN data model has been successfully designed and implemented covering a large portion of the intended scope and functionalities. While domain specific sub-models may require extensions in the area of educational data, they are fully expected to cover all necessary terms and concepts in the biomedical area for the foreseeable future. The MeSH data model allows semantic interoperability both within the platform and outside of it, in the broader biomedical data ecosystem, supporting also multilingual definitions of Data Products.

User specific parameters will surely be extended to allow more nuanced expression of personal preferences on the data access permissioning layer. Additional legal parameters may also be needed as national regulations, as opposed to EU-wide GDPR (General Data Protection Regulation), will be taken into consideration in the second period of the project.

@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