European Olfactory Knowledge Graph, version 1

Deliverable D4.2

Version FINAL

NEGOTIATING OLFACTORY AND SENSORY EXPERIENCES IN CULTURAL HERITAGE PRACTICE AND RESEARCH

The Odeuropa project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004469. This document has been produced by the Odeuropa project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Grant Agreement No.	101004469
Project Acronym	ODEUROPA
Project full title	Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research
Funding Scheme	H2020-SC6-TRANSFORMATIONS-2020
Project website	http://odeuropa.eu/
Project Coordinator	Prof. Dr. Inger Leemans KNAW Humanities Cluster Email: inger.leemans@huc.knaw.nl
Document Number	Deliverable D4.2
Status & version	FINAL
Contractual date of delivery	December 2021
Date of delivery	December 2021
Type	Other
Security (distribution level)	Public
Number of pages	12
WP contributing to the deliverable	WP4
WP responsible	WP4
EC Project Officer	Hinano Spreafico
Authors: Pasquale Lisena,[1] Raphael Troncy,[1] Thibault Ehrhart[1]	
Internal reviewers: Inna Novalija,[2] Stefano Menini,[3] Daniel Schwabe[2]	
Affiliations: (1) EUR, (2) JSI, (3) FBK	
Keywords: Knowledge Graph, Linked Data, SPARQL, GraphDB	
Abstract: This document introduces the first version of the European Olfactory Knowledge Graph (EOKG). While this is intrinsically a resource, we are presenting in this document the first data that has been loaded as well as the required infrastructure	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	10 December 2021	First draft	Pasquale Lisena	All
0.2	16 December 2021	Improvement after internal review	Pasquale Lisena	all
0.3	23 December 2021	Proof Reading	Raphael Troncy	all
1.0	27 December 2021	Final check and approval by project manager	Marieke van Erp	-

Executive Summary

This document presents the first version of the European olfactory Knowledge Graph (EOKG). The work carried out for this deliverable consists of an engineering effort for setting up the technological support for the EOKG. In addition, we converted the data coming from the text annotation campaign carried out as par of WP3 into the RDF format following the Odeuropa ontology, including also some preliminary interlinking efforts. The results of this deliverable are tied with those of D4.1 (*Ontology and Vocabularies for Olfactory Information*), which offers the data structure to be used for the EOKG.

This document offers a list of pointers and descriptions related to:

- The Knowledge Graph (KG) infrastructure, that includes a GraphDB database, a SKOSMOS instance, a Web API and a repository containing dumps and scripts;
- The first data uploaded in the triple store, coming from the annotation of textual resources.

The EOKG will continue to be populated until the end of the Odeuropa project. In particular, an update of this document and of the related resources is foreseen at M30 and will be presented in Deliverable D4.3.

List of Acronyms

AI. Artificial Intelligence
API. Application Programming Interface
CQ. Competency Questions
EOKG. European Olfactory Knowledge Graph
JSON. JavaScript Object Notation
KG. Knowledge Graph
RDF. Resource Description Framework
SKOS. Simple Knowledge Organization System
SPARQL. SPARQL Protocol and RDF Query Language
UI. User Interface
URI. Uniform Resource Identifier

Contents

1 Introduction

Previous work has shown that Knowledge Graphs (KG) are suitable to represent and exploit the domain information in fields such as cultural heritage [Isaac and Haslhofer, 2013, Carriero et al., 2019], history [Koho et al., 2021], and art [Achichi et al., 2018, Dijkshoorn et al., 2018]. The graph paradigm ensures at the same time explicit semantics[1] and a high expressiveness granted by a dynamic data structure.

Knowledge Graphs, in combination with other Semantic Web technologies, provide methods and standards for integrating different data sources, promoting the re-use of resources through shared identifiers (URIs) and the interlinking with other datasets in the Web of Data. In this way, information about the chemical composition of a smell, the structured description of its perception by a person, data about the odorant object or place, and the contextual spatio-temporal information can cohabit in the same graph. In addition, it enables advanced use of domain knowledge, e.g. in smart search, reasoning and AI applications.

For this reason, the Odeuropa project envisions the realisation of a large KG collecting olfactory-related data from different sources. To do so, the design of a suitable representation model was presented in D4.1, relying upon and extending existing ontologies and introducing classes and relations specific to olfactory and perception in general. In addition, specialised vocabularies in machine-readable formats have been created, in which each term can group cross-language synonyms and has explicit logical relationships with other terms – e.g. a *neroli scent* can be part of the *fruity scent* family, and was historically believed to help *combat the plague*.

Such a KG can support the digital preservation of our olfactory heritage, organise the sparse information, capture the involved logical relationships, and enable their access, search, and reuse. In addition, the KG can provide specialised knowledge to feed AI applications, leveraging the data to automatically group similar odours, infer the dominant olfactory features of a particular time period or a place, answer smell-related questions, and improve the search of olfactory references in texts and images [Lisena et al., 2021].

This document presents the first version of the European Olfactory Knowledge Graph (EOKG), a graph of smell-related information coming from text and images from all Europe. The information in the EOKG is represented following the Odeuropa Data Model presented in the Deliverable D4.1. In particular, both the ontology expressed in OWL and the vocabularies expressed in SKOS are part of the RDF graph.

The EOKG infrastructure is made up of:

- a GraphDB triple store, described in Section 2;
- a SKOSMOS instance, described in Section 3;
- a Web API for vocabularies, described in Section 3;
- the `knowledge-graph` repository, containing all code and instructions for setting up the infrastructure, as well as the data dumps, described in Section 4.

In Section 5 we discuss the first data uploaded in the EOKG.

Figure 1 gives a general overview of the infrastructure. The repository includes the scripts for setting up the triplestore. Moreover, it contains the code (converters) for parsing the output of WP2 (image processing) and WP3 (text processing) into RDF format according to the Odeuropa Data model. SKOSMOS and the API sit in front of the triplestore, querying it when required.

To keep the infrastructure setup simple and replicable, we largely rely on Docker[2] images, creating a container for each required software component.

[1] Classes and relations in a KG are formally defined and conceptualised in machine-readable schemas called *ontologies*.

[2] https://www.docker.com/

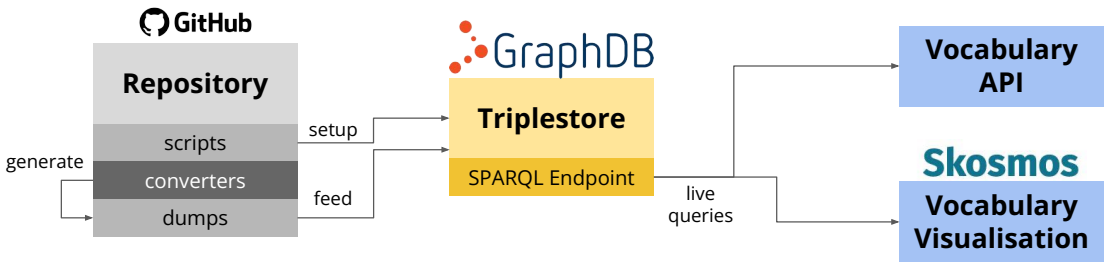

Figure 1: The EOKG infrastructure

2 The GraphDB triplestore

The first version of the EOKG is hosted on a GraphDB instance on EURECOM servers, accessible at http://data.odeuropa.eu/.

We chose GraphDB[3] [Güting, 1994], one of the most popular graph databases available on the market, for the following reasons:

- it offers a free-to-use solution providing all the features required for the project purposes;
- it supports RDF* [Hartig, 2019], which we intend to use in the data model (see D4.1);
- it exposes a web server which can be used for exploring/querying the database and dereferencing URIs;
- it provides a few nice visual interfaces for aggregated data.

All data are accessible through the SPARQL endpoint, exposed in the UI at http://data.odeuropa.eu/sparql (Figure 2) and via client API accepting SPARQL queries at http://data.odeuropa.eu/repositories/odeuropa. The endpoint can provide data in different formats, including JSON, XML, CSV and TSV.

URI dereferencing was put in place using internal URI rewriting functions based on the ProxyPass feature of the Apache Server. The GraphDB front-end was customised to appropriately respond to the requested URI with the list of its incoming and outcoming properties.

For example, it is possible to see dereferencing at work by navigating with a browser to http://data.odeuropa.eu/vocabulary/olfactory-objects/269, representing the *Incense* concept in the Olfactory Object vocabulary (Figure 3). Additionally, an interactive visualisation of the graph can be triggered from each entity, enabling to click and expand the graph iteratively, as shown in Figure 4.

3 SKOSMOS and RESTful API

In addition to the main triple store, two other services can be used to access Odeuropa data: 1) a vocabulary visualiser based on SKOSMOS and 2) a vocabulary API.

The vocabulary visualiser based on SKOSMOS [Suominen et al., 2015] is exposed at http://vocab.odeuropa.eu/. This tool enables to browse the vocabularies in one of the supported languages, showing labels, relationships, matches with external datasets such as ICONCLASS.

An example is shown in Figure 5. On the right-hand side, the detail about the *Incense Burner* concept is shown, displaying the concept's definition, labels in various languages, categories it belongs to (*smell source* and *artifact*), links to the related concept (*Incense*), and the link to the related ICONCLASS entity. On the left-hand side, it is possible to browse the vocabulary using

[3] https://graphdb.ontotext.com/

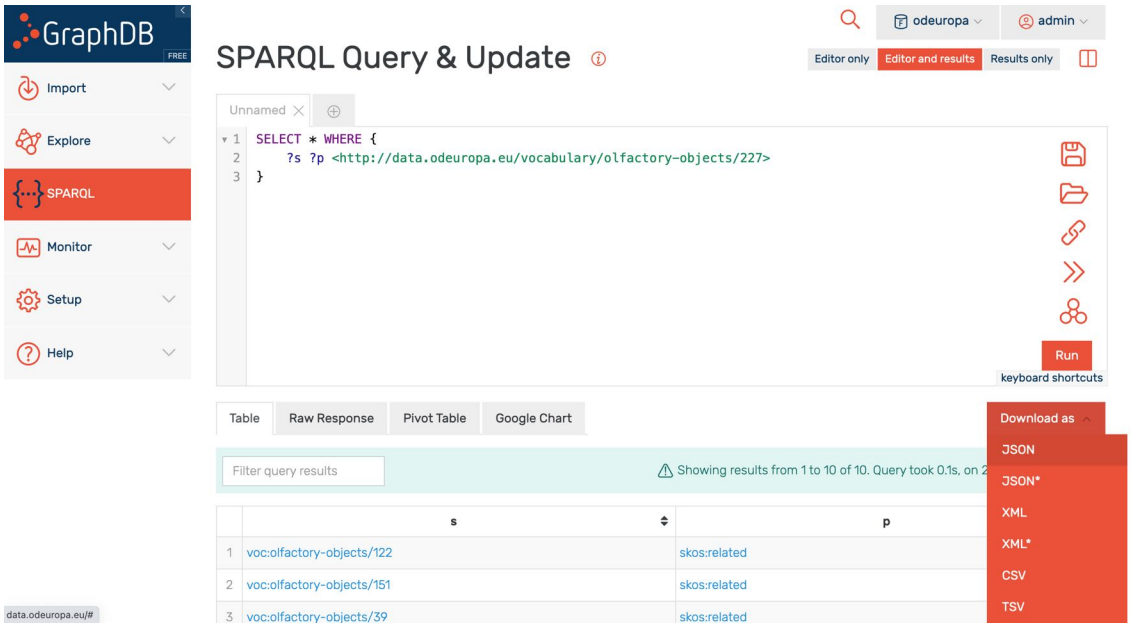

Figure 2: The SPARQL endpoint of data.odeuropa.eu, showing query results and offering the download of results in several formats

]

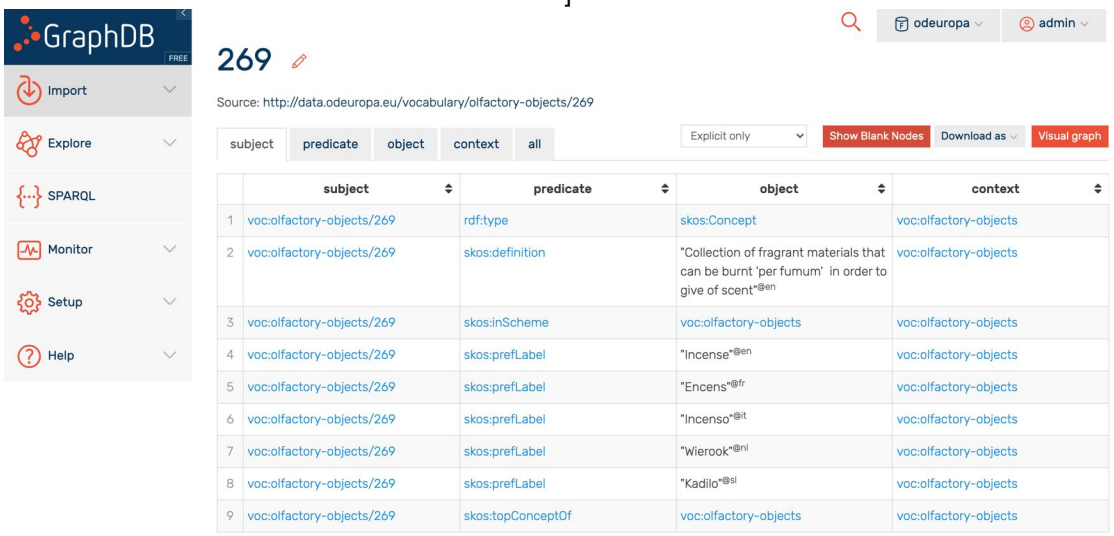

Figure 3: Browsing to http://data.odeuropa.eu/vocabulary/olfactory-objects/269, the server responds with the list of outcoming and incoming (under the *object* tab) properties

.

the alphabetical sorting of the concept names, by expanding the hierarchy and by looking at elements belonging to groups ((*smell source*, *carrier* and *artifact*).

In addition, SKOSMOS offers an overview page for each vocabulary reporting the metadata of the vocabulary – author, name, date of creation and editing – and some stats such as the total number of concepts and the number of labels for each language.

A vocabulary API is deployed at http://data.odeuropa.eu/api/vocabulary. The API is intended to provide a way to list all elements which are part of a vocabulary (e.g. http://data.odeuropa.eu/

]

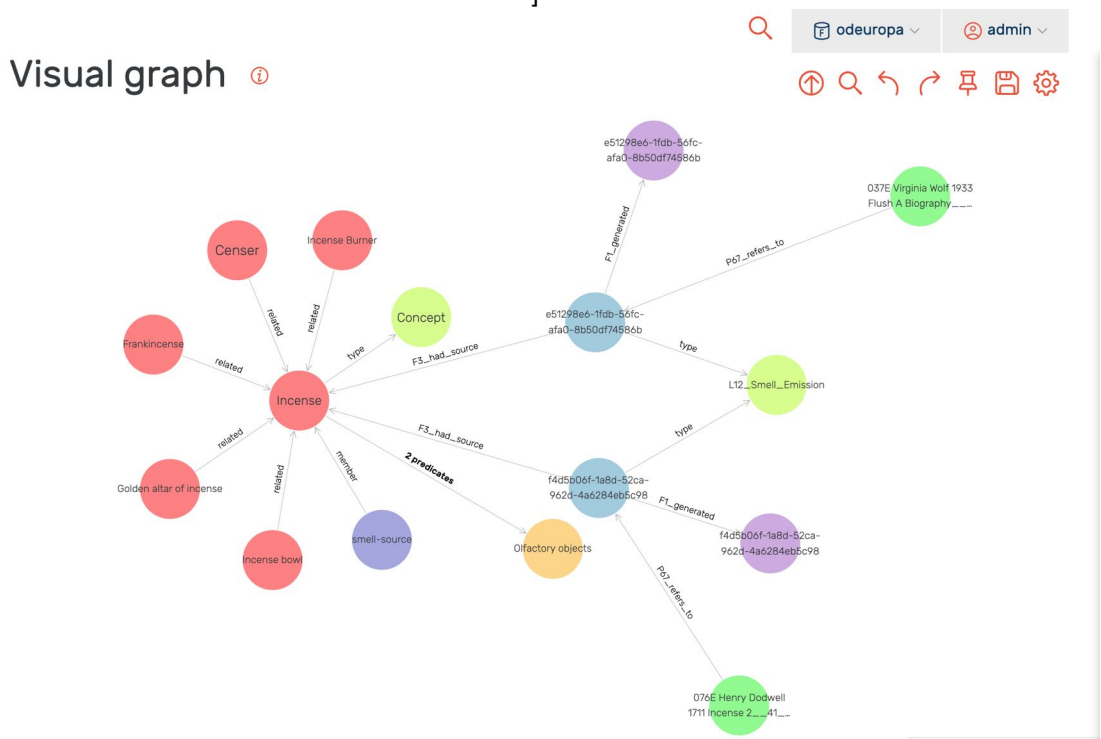

Figure 4: The visual graph of the *Incense* concept, which appears to be the odour source in two different Smell Emissions

.

api/vocabulary/olfactory-objects) and to provide search capabilities which are tolerant to spelling mistakes.

The API exposes the following paths:

- http://data.odeuropa.eu/api/vocabulary/ returns a list of vocabulary names to be used in the following request;
- http://data.odeuropa.eu/api/vocabulary/:voc_name allows to get all concept contained in the vocabulary called `voc_name`, or a selection of them, depending on the chosen query parameters.

The possible query parameters are:

- `lang`: Set the language preference through the `Accept-Language` field. e.g. `en;q=1, it;q=0.7 *;q=0.1`;
- `format`: Output format among `json` (default) and `json-ld`;
- `q`: Allows to search for a particular string. The results are ordered according to a similarity measure based on the Levenshtein distance. The search is multilingual. In case `lang` is set, the specified language are prioritized in the measure;
- autocomplete The search accept only labels that exactly includes the `q` value. The goal is not to find the best matches, but to give good suggestions to the user in an autocomplete scenario.

Some examples:

https://odeuropa.eu

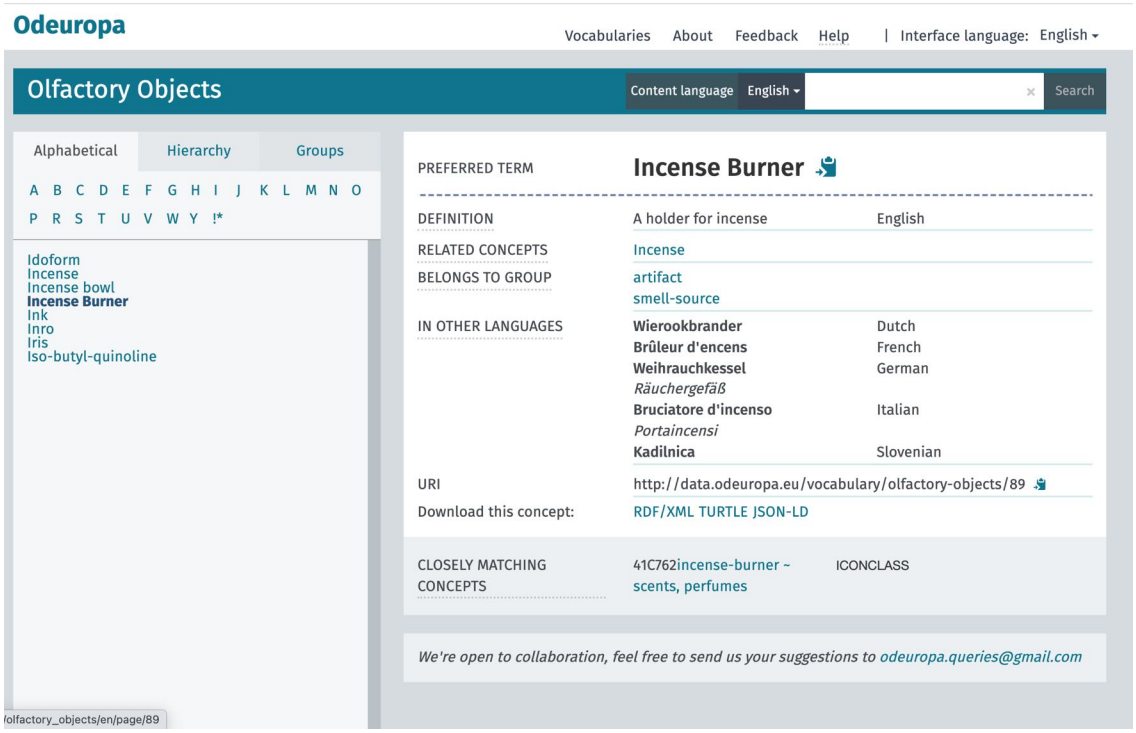

Figure 5: The *Incense Burner* concept in SKOSMOS

```
http://data.odeuropa.eu/api/vocabulary/wheel?q=fresh
http://data.odeuropa.eu/api/vocabulary/wheel?q=fresh&autocomplete
http://data.odeuropa.eu/api/vocabulary/olfactory-objects?q=inceso&lang=it
```

In particular, the last query example contains an intentional spelling error and will return the results more similar to the word ”inceno” in Italian, the first of which is the intended concept ”Incenso”.[4] The output is a JSON array in which each element is presented with its id, a label in the requested language and (in the case of a search) the relevance score. This score is computed as the maximum similarity measure (inverse Levenshtein distance) among the different available labels for that id (case insensitive). Weights are configured for reducing the scores in case the match targets a `skos:altLabel`, a label in another language, or a label not starting with the searched text.

The full documentation of this API is available at https://github.com/D2KLab/vocabulary-api.

4 The `knowledge-graph` repository

A public *git* repository containing all the necessary code and instructions for setting up the Odeuropa KG, together with the dump of the olfactory data is available at https://github.com/Odeuropa/knowledge-graph.

It includes the following directories:

- `scripts` contains the scripts to be run for setting up the EOKG infrastructure and loading the data dump into the triple store, following the instructions in the *README.md* file;
- `graphdb` contains some required files for creating the Docker image of GraphDB using the code in `scripts`;

[4] http://data.odeuropa.eu/vocabulary/olfactory-objects/269

- `skosmos` contains the configuration file for creating the SKOSMOS instance using the code in `scripts`;
- `dump`, contains all files (in RDF-XML and Turtle formats) to be uploaded in the EOKG. Each sub-directory will result in the generation of a distinct named graph in the KG with id `http://data.odeuropa.eu/[directory-name]`, with the exception of the `vocabulary` sub-directory which will generate as many graphs as vocabularies;[5];
- `populate` contains the first versions of the software for converting the annotations into RDF, following the Odeuropa model. The software has been developed in Python, making use of the `rdflib` library for manipulating RDF.[6]

5 Initial data in the EOKG

A first selection of data has been uploaded to the EOKG. In particular, we worked on the 84 annotated text documents in English coming from the Odeuropa benchmark (Deliverable D3.2), exported in tabular format. Apart from converting the data in RDF, we performed the following operations:

- Assigning the content of each field to the right class / property of the Odeuropa data model;
- Parsing dates using regular expressions. This parsing should be extended to durations (e.g. *for several days*), while a strategy for including generic expressions (e.g. *sometimes*, *at various times*) should be defined;
- Interlinking of places with Geonames.[7] 41 of 158 identified locations have been interlinked so far, while most of the others are not necessarily referring to a real place;
- Interlinking of elements of the graph with the controlled vocabularies (Deliverable D4.1). 32 labels have been correctly linked so far. There is much room for improvement in the future, e.g. first lemmatising the terms before attempting the matching.

The current version of the EOKG contains 22,850 distinct triples. Table 1 provides some statistics about the most represented classes.

Class	Number of instances
od:L11_Smell	1,530
od:L13_Olfactory_Experience	1,530
od:L12_Smell_Emission	1,530
crm:E13_Attribute_Assignment	1,084
crm:E33_Linguistic_Object	84
time:TemporalEntity	4
crm:E53_Place	177

Table 1: Statistics about classes represented in the database (see live query at http://data.odeuropa.eu/sparql?savedQueryName=Count%20classes&owner=admin)

[5] This is a requirement of the SKOSMOS software
[6] https://rdflib.readthedocs.io/
[7] http://geonames.org/

6 Conclusion

Table 2 summarizes the links to all the resources related to this deliverable.

Future work involving the EOKG includes:

- parse and include annotations made in other languages, enabling the parsing (e.g. dates) and interlinking of relevant labels;
- implement a strategy for including the data coming from image recognition in WP2;
- improve the parsing and interlinking for all terms, increasing the correct matches and evaluating it appropriately;
- implement a strategy for including in the KG the elements annotated as *Circumstances* (which may include co-occurring events, procedures, etc.) and as *Effects* (which may include gestures, motivations, memories, etc.). We intend to exploit the vocabularies in order to recognise these elements and insert them in the right part of the model;
- collaborate with WP5 to provide content to the storylines and to include them in the EOKG, finding a proper representation in RDF;
- provide a proper end-user access to the EOKG, through the demonstrators to be developed in the Deliverable D4.6.

Resource	**URL**
KG Home	http://data.odeuropa.eu/
SPARQL endpoint (UI)	http://data.odeuropa.eu/sparql
SPARQL endpoint (Software)	http://data.odeuropa.eu/repositories/odeuropa
Vocabularies (SKOSMOS)	http://vocab.odeuropa.eu/
Vocabularies (API)	http://data.odeuropa.eu/api/vocabulary
Repository	https://github.com/Odeuropa/knowledge-graph

Table 2: Resources table

References

[Achichi et al., 2018] Achichi, M., Lisena, P., Todorov, K., Troncy, R., and Delahousse, J. (2018). DOREMUS: A Graph of Linked Musical Works. In *International Semantic Web Conference (ISWC)*, pages 3–19. Springer.

[Carriero et al., 2019] Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., and Veninata, C. (2019). Arco: The italian cultural heritage knowledge graph. In *International Semantic Web Conference (ISWC)*, pages 36–52. Springer.

[Dijkshoorn et al., 2018] Dijkshoorn, C., Jongma, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., ter Weele, W., and Wielemaker, J. (2018). The Rijksmuseum collection as Linked Data. *Semantic Web Journal*, 9:221–230.

[Güting, 1994] Güting, R. H. (1994). Graphdb: Modeling and querying graphs in databases. In *20th International Conference on Very Large Data Bases (VLDB)*, pages 297—-308, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Hartig, 2019] Hartig, O. (2019). The RDF* and SPARQL* Approach to Annotate Statements in RDF and to Reconcile RDF and Property Graphs. In *W3C Workshop on Web Standardization for Graph Data*, Berlin, Germany.

[Isaac and Haslhofer, 2013] Isaac, A. and Haslhofer, B. (2013). Europeana Linked Open Data – data.europeana.eu. *Semantic Web Journal*, 4:291–297.

[Koho et al., 2021] Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., and Hyvönen, E. (2021). WarSampo knowledge graph: Finland in the Second World War as Linked Open Data. *Semantic Web Journal*, 12:265–278.

[Lisena et al., 2021] Lisena, P., van Erp, M., Bembibre, C., and Leemans, I. (2021). Data Mining and Knowledge Graphs as a Backbone for Advanced Olfactory Experiences. In *STT21: Smell, Taste, and Temperature Interfaces workshop*, Yokohama, Japan.

[Suominen et al., 2015] Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., and Retterath, A. (2015). Publishing skos vocabularies with skosmos. *Manuscript submitted for review*.