European Olfactory Knowledge Graph version 2

Deliverable D4.3

Version DRAFT

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 D4.3
Status & version	FINAL
Contractual date of delivery	30 June 2023
Date of delivery	30 June 2023
Type	Other
Security (distribution level)	Public
Number of pages	13
WP contributing to the deliverable	WP4
WP responsible	WP4
EC Project Officer	Hinano Spreafico
Authors: Pasquale Lisena,[1] Raphael Troncy[1]	
Internal reviewers: Arno Bosse,[2] Victoria-Anne Michel,[3]	
Affiliations (1) EURECOM, (2) KNAW, (3) ARU	
Keywords: Knowledge Graph, Semantic Web, Data, Data Access	
Abstract: This document presents the second version of the European Olfactory Knowledge Graph (EOKG). This resource is composed by a set of Controlled Vocabularies, tools for converting the data coming from WP2 (Image Recognition) and WP3 (Text Mining) into the Odeuropa Data model, the data and the infrastructure hosting them, and the tools for accessing the data.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	23/06/2023	First Complete Draft	Authors	All
0.1	27/06/2023	Internal review	Arno Bosse & Victoria-Anne Michel	All
0.2	28/06/2023	Revision after Review	Pasquale Lisena	All
1.0	30/06/2023	Final check and approval by project manager and PI	Marieke van Erp, Inger Leemans	-

Executive Summary

Summary table	
Challenges	The main challenge of this deliverable has been the setup of a complete pipeline for the interpretation, parsing, interlinking and conversion of the data coming from WP2 (Images) and WP3 (Texts).
Barriers	A barrier has been the heterogeneity of different data sources, which differ in the amount of data, language, available metadata, and internal organisation (for which we included some *ad hoc* portions of code).
Practices	In the realisation of these results, we adopted the main Semantic Web principles and state-of-art methods for interlinking and accessing Knowledge Graph (KG) data in a cultural heritage context.
Guidelines	We believe that the complete infrastructure and the proposed method for converting data can be considered as guideline for future work. Both resources can be easily adapted for reuse in different contexts and applications.

Layman's Summary

Deliverable D4.3 'European Olfactory Knowledge Graph version 2' is a set of resources which compose a Knowledge Graph of olfactory information derived by the data extracted from historical images (paintings, prints, etc., processed in WP2) and texts (novels, scientific texts, journals, diaries, etc., processed in WP3). The different resources include the tools for creating, storing and providing data to the final user, together with the data itself. The EOKG is now cmoposed of more than 700k smell experiences extracted from 62k textual resources (books) and 23k images (paintings).

List of Acronyms

AI. Artificial Intelligence
API. Application Programming Interface
CAS. Chemical Abstracts Service
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
VDI. Verein Deutscher Ingenieure

Contents

1 Introduction

This deliverable presents the second version (v2) of the European Olfactory Knowledge Graph (later in this document referred as EOKG or simply KG), the complimentary resources and the methods and software used for building, populating and accessing the data. The first version (v1) of the EOKG was presented in Deliverable D4.2 [Lisena et al., 2021b], which is superseded by this deliverable. Moreover, an extension and improvement to Deliverable D4.1 [Lisena et al., 2021a] is also proposed in this deliverable.

The EOKG infrastructure is made up of:

- a GraphDB triple store, described in Section 2 of D4.2;
- a SKOSmos[1] [Suominen et al., 2015] instance, described in Section 3 of D4.2;
- a Web AP for vocabularies, described in Section 3 of D4.2;
- **[NEW]** a Web API for the KG, described in Section 5.1;
- **[NEW]** the `knowledge-graph` repository, containing all code and instructions for setting up the infrastructure, as well as the data dumps, described in Section 4 of D4.2.

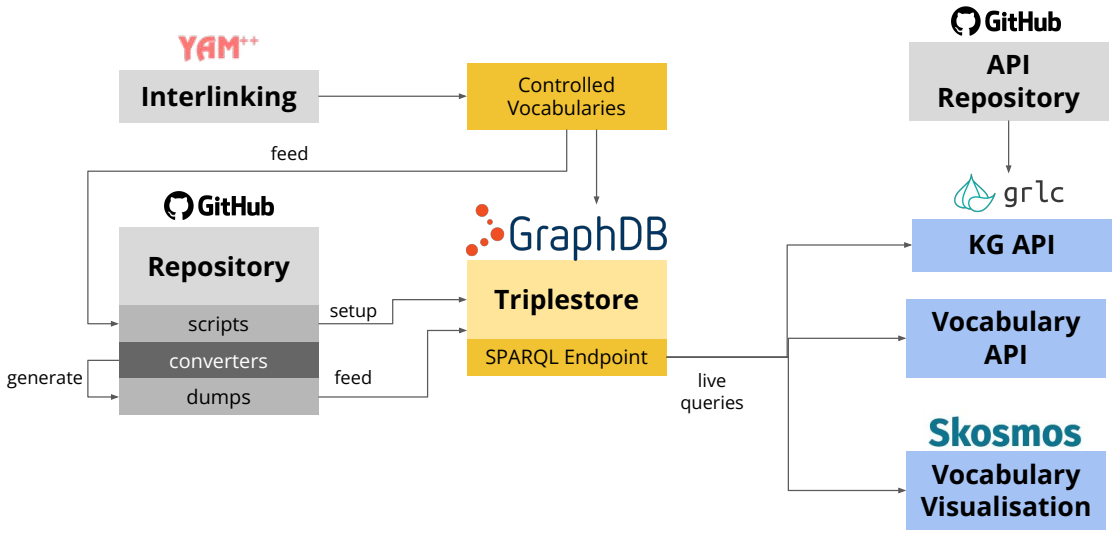

Figure 1: The EOKG infrastructure

Figure 1 gives a general overview of the infrastructure. The repository includes the scripts for setting up the triple store. Moreover, it contains the code (converters) for parsing the output of WP2 (image processing) and WP3 (text processing) into RDF according to the Odeuropa Data Model [Lisena et al., 2022]. SKOSMOS and the two API interfaces sit in front of the triple store, querying it when required. In addition, we materialised a proper interlinking between controlled vocabularies, as detailed in Section 2. The vocabularies are used in the conversion process, for directly linking them to the converted data.

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

[1] https://skosmos.org/
[2] https://www.docker.com/

2 Controlled Vocabularies

In [Lisena et al., 2021a], we introduced some controlled vocabularies realised during the project, for entity disambiguation and synonym management. In this version 2, we introduced two kinds of improvements.

First, we extended the vocabularies to include new ones, both among the already existing taxonomies that we converted in SKOS [Miles and Pérez-Agüera, 2007], and the ones realised in the context of the project. The result of this effort is a set of 13 converted taxonomies and 4 new vocabularies, respectively listed in Tables 1 and 2, in which the newly included vocabularies are marked in gray.

In particular, the following vocabularies have been added:

- Dutch Historical Smell Vocabulary (DHSV), a collection of Dutch smell words collected from historical dictionaries and through questionnaires by Inger Leemans for the Neuswijzer project [Verbeek and Leemans, 2023];
- From Dravnieks' Atlas of Odor Character Profiles [Dravnieks, 1985], we extracted the list of descriptors and odorants, in two different vocabularies;
- Flavornet's odorants [Arn and Acree, 1998] (the descriptors were already included in v1);
- Plutchick's Wheel of Emotions, which is not an odor-specific vocabulary but has a role in the Odeuropa Data Model [Lisena et al., 2021a];
- Two vocabularies from the standard VDI for Olfactometry: Part 1 for Determination of Odour intensity [VDI, 1992]; Part 2 for Determination of Hedonic Odour Tone [VDI, 1994].

Vocabulary	Type	Levels	Top Level Concepts	Total Concepts
Drom's fragrance circle	Odour wheel	2	16	77
Michael Edwards' scent wheel	Odour wheel	2	4	18
Odour wheel of historical books	Odour wheel	2	8	43
Nose-first classification of iconographies	Classification	1+1	25	168
Flavornet and human odour space	Classification	1+1	25	495
Zwaardemaker smell system	Classification	1+1	9	9
Dutch Historical Smell Vocabulary (DHSV)	Smell Word	1	800	800
Dravnieks descriptors	Classification	1	146	146
Dravnieks odorants	Compounds	1	800	800
Flavornet odorants	Compounds	1	740	740
Plutchik's Wheel of Emotions	Emotions	3	17	35
VDI 1 - Odour intensity	Intensity	1	7	7
VDI 2 - Hedonic Odour Tone	Hedonic	1	58	58

Table 1: Vocabularies converted in SKOS. Some classification systems have a second level which consists of smell sources rather than smell classes (reported as 1+1).

The inclusion of Dravnieks' and Flavornet odorants makes it possible to connect the EOKG with the chemical perfumery domain. Those two vocabularies have the Chemical Abstracts Service (CAS) number, a widely adopted identifier system for chemical substances. The CAS number allows an easy interlink with the entities representing these substances in Wikidata and can potentially be a bridge to other specialised datasets.

Among the vocabularies developed during the project, we see the inclusion of the vocabulary of Noses (smell experts) – which includes relevant professions such as *Apothecary* or particular animals like *Dog*, among other terms – and the extension of the other vocabularies both in coverage and languages. A wide restructuring of the Olfactory Object vocabulary has been realised, with

the rationalisation in 14 super-categories – e.g. Animal Product, Flora, Artifact, etc. – and the inclusion of new terms, passing from 417 (v1) to 574 (v2) distinct concepts.

Vocabulary	Total Concepts	EN	DE	FR	IT	NL	SL
Fragrant Spaces	115	115	10	111	109	111	110
Olfactory Gestures	35	35	0	33	32	16	31
Olfactory Objects	574	567	239	550	540	542	568
Noses	35	35	0	0	0	0	0

Table 2: Multilingual vocabularies for English (EN), German (DE), French (FR), Italian (IT), Dutch (NL), and Slovene (SL)

An interlinking strategy has been put in place to better connect the different vocabularies. Following the experience in [Achichi et al., 2018], we used YAM++[3] [Bellahsene et al., 2017] for performing a string matching.[4] This string matching is a combination of the Levenshtein distances computed on the tokens composing the labels (removing stopwords and rewarding partial overlaps). We pick the best match for each term having a confidence score ≥ 0.7 (on a range between 0 and 1) and instantiate for each match a `skos:exactMatch` link between the original and matching term. A total of 606 `skos:exactMatch` links (of which some repeated in the two directions of the edge) between vocabulary terms have been added, with the detail of the count of alignments between each vocabulary reported in Table 3. We believe that these links (making the graph more densely interconnected) will improve the computation over the graph, e.g. the generation of graph embeddings.

	dhsv sw	dravnieks odorants	dravnieks descriptors	drom	edwards	flavornet descriptors	historic book	historic scent	zwaardemaker	TOTAL
dhsv sw			6	2	1	4		6		**19**
dravnieks odorants	6				5		15			**26**
dravnieks descriptors	6			12	5	54	15	26	2	**120**
drom	2	12	12		5	18	3	18		**70**
edwards	2		4	6		5	4	4	1	**26**
flavornet descriptors	6		58	19	7		18	42	1	**151**
historic book			15	3	4	17		10	1	**50**
historic scent	6	29	31	19	4	35	13			**137**
zwaardemaker		2	2		1	1	1			**7**
TOTAL	**28**	**43**	**128**	**61**	**32**	**134**	**69**	**106**	**5**	**606**

Table 3: Exact matches instantiated between each couple of vocabularies.

As future work, we would like to use the same method to interlink these vocabularies with the olfactory objects, instantiating `skos:related` links between odor classification terms (e.g. *woody*) and smell sources (e.g. *wood*).

[3] https://gite.lirmm.fr/opendata/yampp-ls

[4] We decided to not include any alignment involving structural analysis (even if available in YAM++) because the different structures of odor classifications are intended to represent the different approaches and points of view of their authors. We want to preserve this difference and find matches between these vocabularies beyond the different structures.

3 Ingesting Datasets

The KG population is realised using a centralised software (lately called *converter*) developed in Python, relying on widely used libraries such as rdflib, spaCy and SPARQL Transformer [Lisena et al., 2019]. The converter is organised with the following modules.

Two **entry-point files** for converting images (`convert_img.py`) and text (`convert_text.py`), which require parameters such as the language (for text) and the folder name. This folder is expected to include 3 files:

- a metadata file, including information such as the author, the title and the external link of the resource;
- the annotation file, in a tabular format;
- a mapping file, which reports the equivalences between the identifiers of the first two files.

After reading these files, the software starts parsing and interpreting them, assigning to each piece of information one of the **Entity subclasses**. A subclass is instantiated for each of the entities of the Odeuropa Data Model (e.g. Olfactory Experience or Gesture), in order to define in a unique place the attributes of each entity, how they interact each other, and some common behavior – e.g. when an entity is instantiated, an URI is automatically assigned under the `http://data.odeuropa.eu/` namespace. The URIs are generated in a deterministic way, according to the rules reported at https://github.com/Odeuropa/knowledge-graph/blob/master/URI-patterns.md,

Finally, some **special modules** handle some specific complex activities such as:

- Parsing of time;
- Parsing and interlinking of places to GeoNames;
- Interlinking of several entities with our controlled vocabularies.

The complete documentation of the software is available in the github repository.

Apart from the Odeuropa benchmark (already present in v1), all the ingested graph are coming from the annotation effort of WP2 and WP3, namely:

- the ODOR dataset [Zinnen et al., 2022], including both manual and automatic annotations;
- the automatic extraction with the text processing system version 1 [Menini et al., 2022];
- the automatic extraction with the text processing system version 2 [Hurriyetoglu et al., 2023] (limited for now to EEBO, British Library, the English part of Gutenberg, Pulse, ECCO).

4 European Olfactory Knowledge Graph (EOKG) Overview

The current version of the EOKG contains 32,435,839 distinct triples, describing 718,375 smell instances. It includes 694,505 sentences from 62,727 textual resources (books, documents, letters, legal articles, etc.) and 23,870 images.

In Table 4, the number of the smell instances are reported for each named graph representing a precise data source (image or text corpus).

Table 5 provides some statistics about the most represented classes, as well as how many times we were able to interlink them with our controlled vocabularies or with external knowledge bases (Geonames, Wikidata, data.bnf.fr). Please note that for places successfully interlinked with Geonames, we instantiate a single instance for all occurrences of the same place (e.g. London), although this place can involve multiple smells. Similarly, for people, we generate the same URI (so the same entity) if the name and surname of the person are the same.

Graph	N. of smell instances	Type	Language
http://data.odeuropa.eu/british-library	266667	Text	EN
http://data.odeuropa.eu/gallica	261502	Text	FR
http://data.odeuropa.eu/gutenberg	33456	Text	EN, IT
http://data.odeuropa.eu/dbnl	32878	Text	NL
http://data.odeuropa.eu/image-annotation	23870	Image	–
http://data.odeuropa.eu/dlib	22270	Text	SL
http://data.odeuropa.eu/liberliber	21741	Text	IT
http://data.odeuropa.eu/dta	14844	Text	DE
http://data.odeuropa.eu/wikisource	12970	Text	IT
http://data.odeuropa.eu/ecco	8435	Text	EN
http://data.odeuropa.eu/text-annotation	7125	Text	All
http://data.odeuropa.eu/pulse	6562	Text	EN
http://data.odeuropa.eu/royal-society-corpus	4632	Text	EN
http://data.odeuropa.eu/old-bailey-corpus	998	Text	EN
http://data.odeuropa.eu/bibbleue	385	Text	FR
http://data.odeuropa.eu/grimm	40	Text	FR

Table 4: Count of smell instances per graph.

Class	N. instances	Vocabularies	External sameAs
crmsci:S10_Material_Substantial	1048284	483342	
crm:E33_Linguistic_Object	881441		62985
od:L12_Smell_Emission	718375		
od:L11_Smell	718375		
od:L13_Olfactory_Experience	718375		
crm:E13_Attribute_Assignment	590285		
crm:E21_Person	83601	66	8709
crm:E53_Place	53512	8215	1181
crm:E36_Visual_Item	51486		2380
crm:E90_Symbolic_Object	47794		
crm:E39_Actor	40113		
crm:E22_Human-Made_Object	30165	30165	
time:TemporalEntity	21744		
prov:Activity	7241		
readit:REO21	494	491	
od:L7_Gesture	483	483	
od:L6_Animal	74	74	

Table 5: Statistics about classes represented in the database (see live query at https://data.odeuropa.eu/sparql?savedQueryName=Count%20classes%20(full)&owner=admin&execute)

5 Data Access

There are several several ways to access the data of the EOKG. Next to the SPARQL Endpoint, SKOSmos and the Vocabulary API (described in [Lisena et al., 2021b]), we include two new access points in this second version of the KG.

5.1 EOKG API

The EOKG API is another way to retrieve information from the KG, specifically intended for web developers without specific knowledge about SPARQL. The API has been realised with the combination of two technologies [Lisena et al., 2019]:

(1) *grlc* is a platform which reads the queries stored in a repository and creates a running RESTful API around them. In addition, some syntactic sugar allows one to define query parameters. *grlc* is also automatically generating the documentation of the API itself, relying on OpenAPI standards;

(2) *SPARQL Transformer* is a library (integrated in *grlc*) which allows one to write queries to a SPARQL Endpoint and receive responses in the form of a filled JSON template. The advantage of this system is that the query results are not returned as a list of solutions of the query (with repetitions, fields to parse, etc.) but in a convenient JSON structure decided by the developer.

The realised API includes ways to search for specific smells, obtaining all information available for a given smell, retrieving labels for specific concepts, etc.

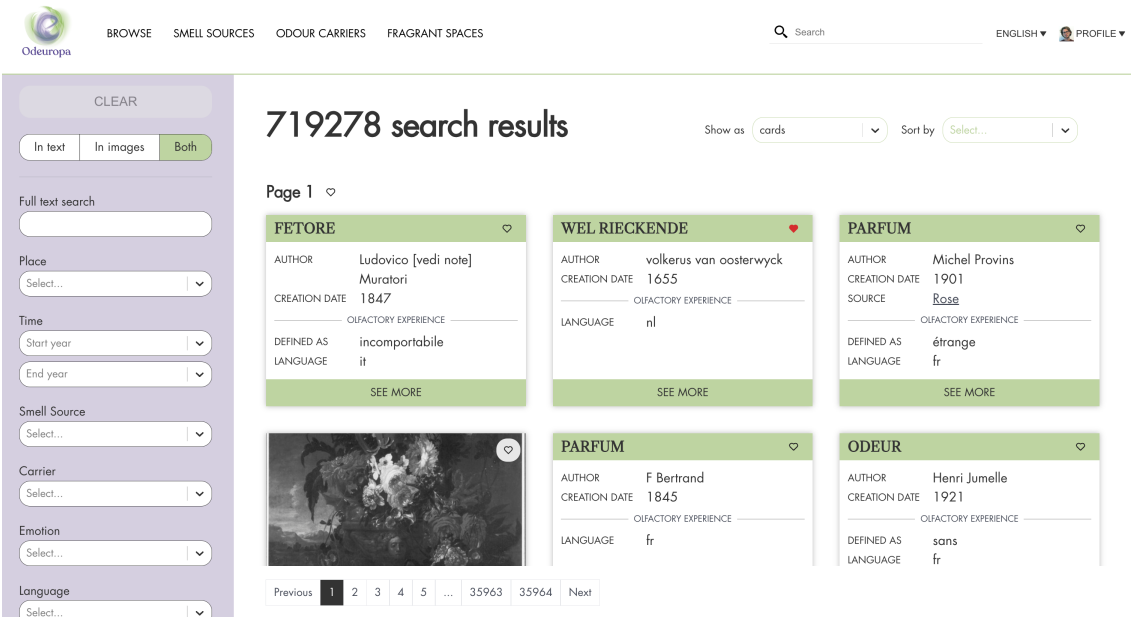

Figure 2: The search page in Odeuropa Smell Explorer

5.2 Smell Explorer

All the EOKG data are also accessible via an online user-friendly web application called the Odeuropa Smell Explorer, https://explorer.odeuropa.eu/. The tool, based on KG Explorer [Ehrhart et al., 2021], does not require any programming by the user, and has two main features:

- An advanced search function for smell references, with the possibility to select textual or visual references (or both), perform a full text search or select one of the terms of the multilingual vocabularies for filtering the relevant results in a multilingual fashion (Figure 2);
- The possibility to look at the list of smell sources and fragrant spaces and see, for each of them, aggregated results such as timelines (to see the evolution of a smell in time), maps (to identify areas in which a smell is particularly present) and word clouds (to understand how this smell has been perceived and described by people).

More details about the Smell Explorer will be given in D4.6 *Demonstrators for Olfactory Cultural Heritage* (due in M36).

6 Conclusion

Table 6 summarizes the links to all the resources related to this deliverable.

Resource	**URL**
KG	http://data.odeuropa.eu/
- (SPARQL endpoint, UI)	http://data.odeuropa.eu/sparql
- (SPARQL endpoint, Software)	http://data.odeuropa.eu/repositories/odeuropa
- (source code + dumps)	https://github.com/Odeuropa/knowledge-graph
Vocabularies (SKOSMOS)	http://vocab.odeuropa.eu/
- (API)	http://data.odeuropa.eu/api/vocabulary
- (source code)	https://github.com/Odeuropa/vocabularies
EOKG API	http://grlc.eurecom.fr/api/Odeuropa/kg-api/
- (source code)	https://github.com/Odeuropa/kg-api
Smell Explorer	https://explorer.odeuropa.eu/
- (source code)	https://github.com/Odeuropa/explorer

Table 6: Resources table

Future work regarding the EOKG includes:

- integrate the results coming from the text processing system v2 for all languages [Hurriyetoglu et al., 2023];
- implement a strategy for including in the KG contextual information [Novalija et al., 2023] and the results of emotion analysis [Odeuropa team, 2023];
- include the results of WP5 to provide access to the storylines and encyclopedia content;
- provide abetter end-user access to the EOKG via demonstrator apps being reported in Deliverable D4.6.

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 *The Semantic Web – ISWC 2018*, pages 3–19.

[Arn and Acree, 1998] Arn, H. and Acree, T. (1998). Flavornet: a database of aroma compounds based on odor potency in natural products. *Developments in food science*, 40:27–28.

[Bellahsene et al., 2017] Bellahsene, Z., Emonet, V., Ngo, D., and Todorov, K. (2017). YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation. In Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., and Hartig, O., editors, *The Semantic Web: ESWC 2017 Satellite Events*, pages 137–142.

[Dravnieks, 1985] Dravnieks, A. (1985). *Atlas of odor character profiles*. American Society for Testing and Materials, Philadelphia, USA.

[Ehrhart et al., 2021] Ehrhart, T., Lisena, P., and Troncy, R. (2021). KG Explorer: a Customisable Exploration Tool for Knowledge Graphs. In *Visualization and Interaction for Ontologies and Linked Data (VOILA)*.

[Hurriyetoglu et al., 2023] Hurriyetoglu, A., Tonelli, S., Menini, S., Brank, J., Tullet, W., and van Erp, M. (2023). Odeuropa text processing system with historical corpus analysis version 2. Technical report, Odeuropa. Deliverable D3.4.

[Lisena et al., 2019] Lisena, P., Meroño-Peñuela, A., Kuhn, T., and Troncy, R. (2019). Easy Web API Development with SPARQL Transformer. In *18$ˆth$ International Semantic Web Conference (ISWC), In-Use Track*, Auckland, New Zealand.

[Lisena et al., 2021a] Lisena, P., Schwabe, D., van Erp, M., Troncy, R., Tullet, W., Leemans, I., Marx, L., and Ehrich, S. (2021a). Ontology and Vocabularies for Olfactory Information. Technical report, Odeuropa. Deliverable D4.1.

[Lisena et al., 2022] Lisena, P., Schwabe, D., van Erp, M., Troncy, R., Tullett, W., Leemans, I., Marx, L., and Ehrich, S. C. (2022). Capturing the semantics of smell: The odeuropa data model for olfactory heritage information. In *The Semantic Web: 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29–June 2, 2022, Proceedings*, pages 387–405.

[Lisena et al., 2021b] Lisena, P., Troncy, R., and Ehrhart, T. (2021b). European Olfactory Knowledge Graph, version 1. Technical report, Odeuropa. Deliverable D4.3.

[Menini et al., 2022] Menini, S., Sinem Tekiroğlu, S., and Tonelli, S. (2022). Odeuropa text processing system version 1. Technical report, Odeuropa. Deliverable D3.3.

[Miles and Pérez-Agüera, 2007] Miles, A. and Pérez-Agüera, J. R. (2007). SKOS: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3-4):69–83.

[Novalija et al., 2023] Novalija, I., Troncy, R., Mladenić, D., Rei, L., Lisena, P., Menini, S., and Rebboud, Y. (2023). Odeuropa text processing system with historical corpus analysis version 2. Technical report, Odeuropa. Deliverable D4.5.

[Odeuropa team, 2023] Odeuropa team (2023). Final Odeuropa text processing system with emotion analysis version 3. Technical report, Odeuropa. Deliverable D3.5, foreseen in November.

[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*.

[VDI, 1992] VDI (1992). Olfactometry-Determination of Odour Intensity, Part 1. Technical report, VDI, Düsseldorf, Germany. VDI-guideline 3882/1.

[VDI, 1994] VDI (1994). Olfactometry-Determination of Hedonic Odour Tone, Part 2. Technical report, VDI, Düsseldorf, Germany. VDI-guideline 3882/2.

[Verbeek and Leemans, 2023] Verbeek, C. and Leemans, I. (2023). *NeusWijzer. Geuratlas van de Lage Landen*. Boom uitgevers Amsterdam.

[Zinnen et al., 2022] Zinnen, M., Madhu, P., Kosti, R., Bell, P., Maier, A., and Christlein, V. (2022). Odeuropa dataset of smell-related objects.