When AI and Dialect Data meet:
crossing-borders between dialectology and data science:
an exploration for the Southern Dutch Dialects
(Short Paper)

Krishna Kumar Thirukokaranam Chandrasekar, IDLab & Ghent Centre for Digital Humanities,
Sally Chambers, Ghent Centre for Digital Humanities, Ghent University,
Veronique De Tier, Jesse de Does, Katrien Depuydt, Dutch Language Institute

Keywords: CLARIAH, CLARIN, DiSSCo, herbaria, cross-collection linking, dialectology, FAIR data, language atlases, lexicography, linguistic data, geo-visualisation, Open Science

1. Introduction

The Database of the Southern Dutch Dialects (DSDD) is the result of bringing together the datasets of three large dialect dictionaries (The Dictionary of the Flemish, Brabantic and Limburgian dialects) in a harmonised dataset of concepts (Van den Heuvel, et al., 2016), a user-friendly search engine and a geo-visualisation tool. The application backend provides an Application Programming Interface (API) to export subsets of the data for analysis using existing digital research tools.

At the previous DH Benelux conferences the DSDD team introduced the project (2017), explored the cartographic tools (2018), demonstrated the prototype (2019) and explored the potential for interdisciplinary synergies (2021). Currently the database consists of 29.981 concepts and 530.605 different dialect words. The team will now present a case study that illustrates the role of dialect data in enabling cross-domain studies such as efficiently searching for plants in global herbaria collections using their local dialect names.

2. Interdisciplinary Methodological Approach

The DSDD was conceived as an interdisciplinary project, bringing together researchers from four core disciplinary areas:

a) dialectologists[1] for a thorough understanding of the dialect words (Van Keymeulen, 2004), and the lexicographical construction of the dictionaries and **computational linguists**[2] for the modelling of dialect data (De Vriend, 2012) and the conceptual understanding of the linguistic make-up of the dictionaries,

b) cartographers[3] to identify appropriate web-mapping technologies currently used beyond the field of linguistics which could be adapted for geo-visualisation of dialect data,

[1] Ghent University, Department of Linguistics; Radboud University, Nijmegen and the Meertens Institute

[2] Dutch Language Institute; Catholic University of Leuven, University of Groningen

[3] CartoGis Research Group, Department of Geography, Ghent University

c) digital humanities experts[4] to understand how to manage research data sustainably and to design research scenarios for the analysis of the integrated dataset using existing digital research tools[5] and
d) computer and data scientists[6] to understand how to manage and inter-link the linguistic concepts and dialect words using linked open data technologies, plus provide access to data via an API.

In this paper we will focus on the collaboration with computer and data scientists and the application of AI with regards to natural heritage collections.

3. DSDD Platform and Integrated Dataset

The DSDD Platform consists of a user-friendly search engine and a geo-visualisation tool[7]. In the DSDD researchers can search for dialect words for specific concepts (e.g. 'dandelion') or for the specific meaning of a dialect word. Furthermore they can filter the results according to their needs, by country, theme or location (see Figure 1). The geographical distribution of the results can be visualised on dialect maps using colours and symbols, which can be customised by the user. The researcher can select the various colours, symbols and size him or herself and adjust the legend to a certain extent as s/he sees fit. (see Figure 2)

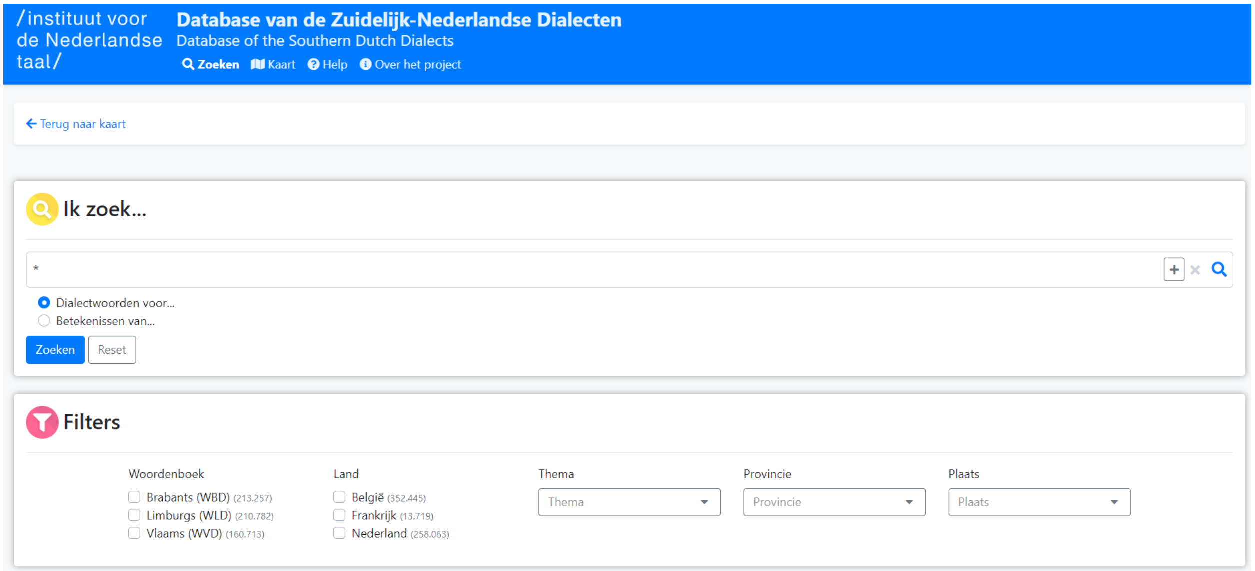


Figure 1. DSDD Search Engine

[4] Ghent Centre for Digital Humanities
[5] See for example: Barbot, L.; Fischer, F.; Moranville, Y and Pozdniakov, I (2019) *Which DH tools are actually used in research?* [Blogpost] weltliteratur.net - A Black Market for the Digital Humanities: https://weltliteratur.net/dh-tools-used-in-research/
[6] IDLab - Internet Technology and Data Science Lab, Ghent University
[7] The application consists of a Solr-based backend implementing an API on the data, and a user interface developed in Vue.js, with cartographical components based on the Leaflet library.

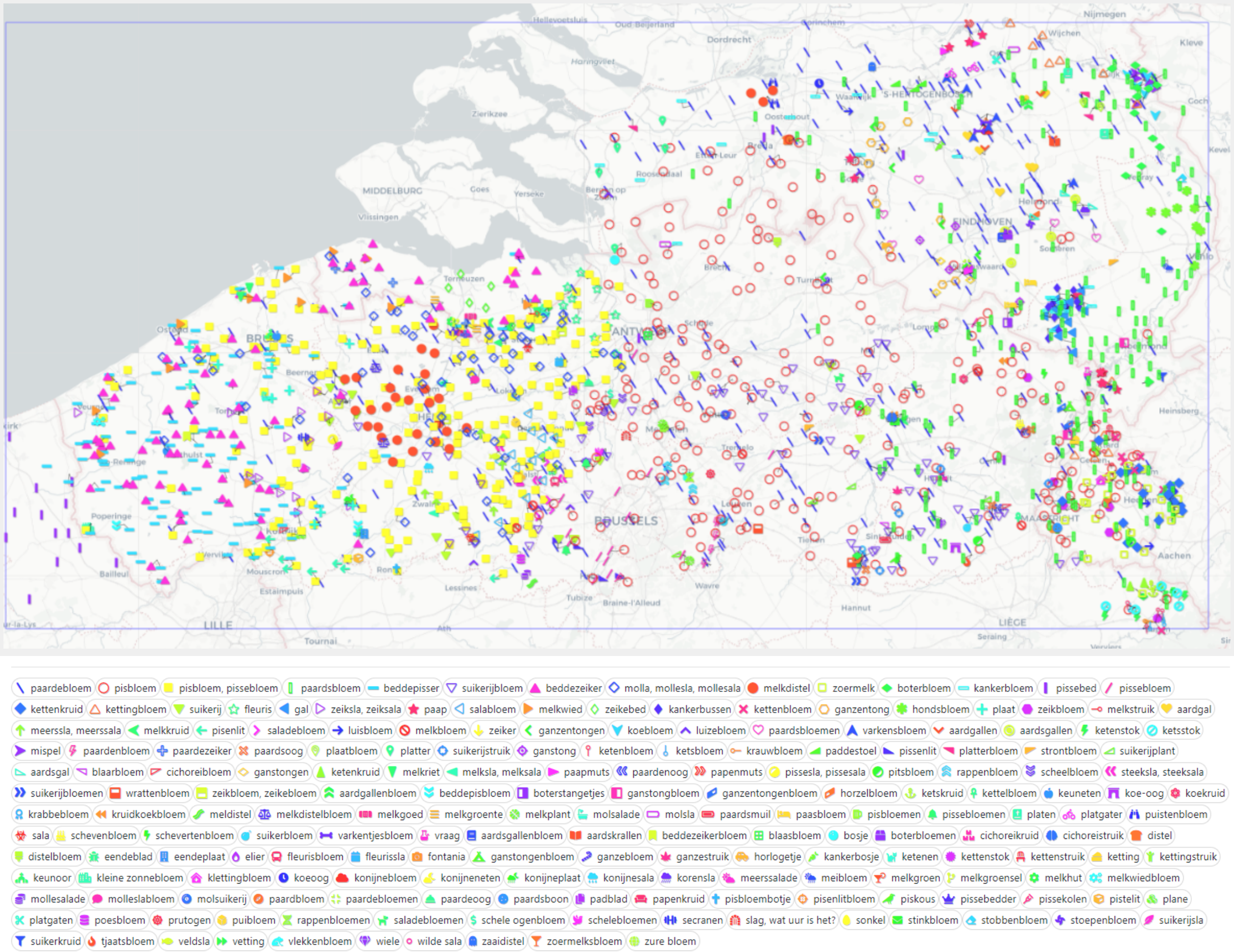
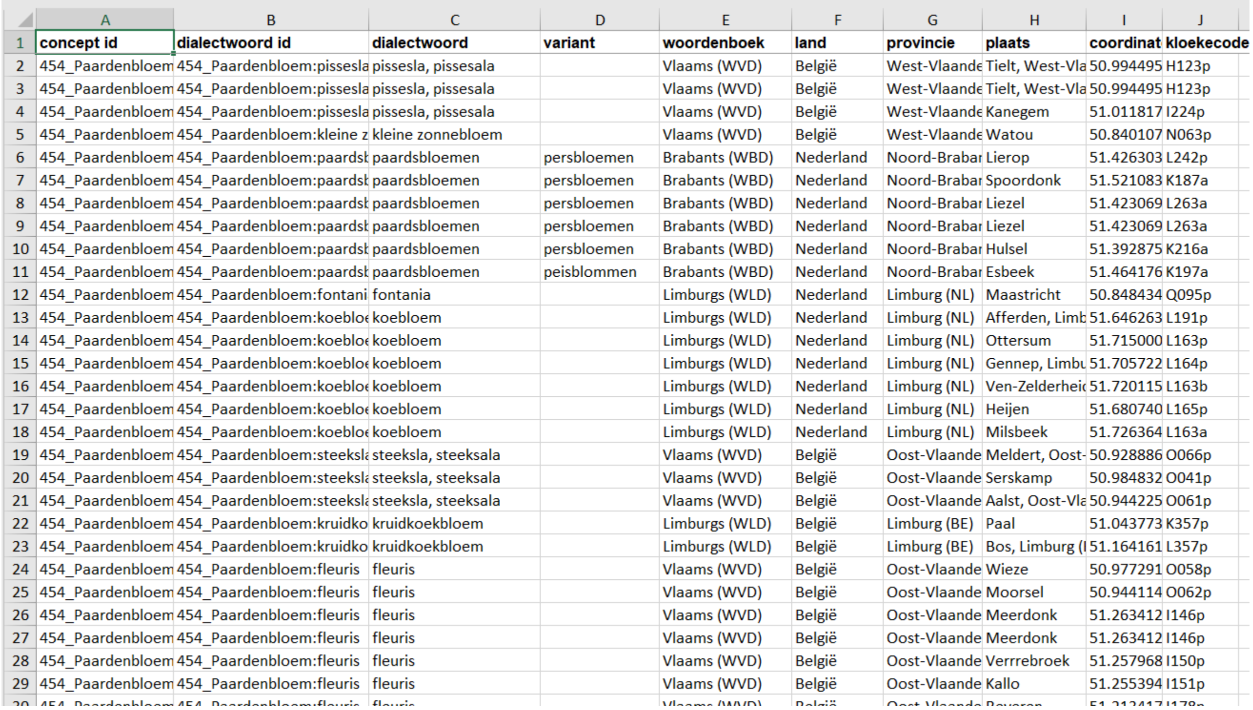


Figure 2. Geo-visualisation of the dialect words for 'dandelion'

Until now, it had not been possible to gain data-level access to the integrated dataset. Within the DSDD platform, researchers are able to export subsets of the data, such as Excel, tsv, csv and xml files (see Figure 3) for analysis by using existing digital research tools. This functionality is a first application of the DSDD Application Programming Interface (API), which is also accessed by the user interface, for access to larger datasets.



	A	B	C	D	E	F	G	H	I	J
1	concept id	dialectwoord id	dialectwoord	variant	woordenboek	land	provincie	plaats	coordinat	kloekecode
2	454_Paardenbloem	454_Paardenbloem:pissesla	pissesla, pissesala		Vlaams (WVD)	België	West-Vlaande	Tielt, West-Vla	50.994495	H123p
3	454_Paardenbloem	454_Paardenbloem:pissesla	pissesla, pissesala		Vlaams (WVD)	België	West-Vlaande	Tielt, West-Vla	50.994495	H123p
4	454_Paardenbloem	454_Paardenbloem:pissesla	pissesla, pissesala		Vlaams (WVD)	België	West-Vlaande	Kanegem	51.011817	I224p
5	454_Paardenbloem	454_Paardenbloem:kleine z	kleine zonnebloem		Vlaams (WVD)	België	West-Vlaande	Watou	50.840107	N063p
6	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	persbloemen	Brabants (WBD)	Nederland	Noord-Brabar	Lierop	51.426303	L242p
7	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	persbloemen	Brabants (WBD)	Nederland	Noord-Brabar	Spoordonk	51.521083	K187a
8	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	persbloemen	Brabants (WBD)	Nederland	Noord-Brabar	Liezel	51.423069	L263a
9	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	persbloemen	Brabants (WBD)	Nederland	Noord-Brabar	Liezel	51.423069	L263a
10	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	persbloemen	Brabants (WBD)	Nederland	Noord-Brabar	Hulsel	51.392875	K216a
11	454_Paardenbloem	454_Paardenbloem:paardsb	paardsbloemen	peisblommen	Brabants (WBD)	Nederland	Noord-Brabar	Esbeek	51.464176	K197a
12	454_Paardenbloem	454_Paardenbloem:fontani	fontania		Limburgs (WLD)	Nederland	Limburg (NL)	Maastricht	50.848434	Q095p
13	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Afferden, Limb	51.646263	L191p
14	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Ottersum	51.715000	L163p
15	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Gennep, Limbu	51.705722	L164p
16	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Ven-Zelderheid	51.720115	L163b
17	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Heijen	51.680740	L165p
18	454_Paardenbloem	454_Paardenbloem:koebloe	koebloem		Limburgs (WLD)	Nederland	Limburg (NL)	Milsbeek	51.726364	L163a
19	454_Paardenbloem	454_Paardenbloem:steeksla	steeksla, steeksala		Vlaams (WVD)	België	Oost-Vlaande	Meldert, Oost-	50.928886	O066p
20	454_Paardenbloem	454_Paardenbloem:steeksla	steeksla, steeksala		Vlaams (WVD)	België	Oost-Vlaande	Serskamp	50.984832	O041p
21	454_Paardenbloem	454_Paardenbloem:steeksla	steeksla, steeksala		Vlaams (WVD)	België	Oost-Vlaande	Aalst, Oost-Vla	50.944225	O061p
22	454_Paardenbloem	454_Paardenbloem:kruidko	kruidkoekbloem		Limburgs (WLD)	België	Limburg (BE)	Paal	51.043773	K357p
23	454_Paardenbloem	454_Paardenbloem:kruidko	kruidkoekbloem		Limburgs (WLD)	België	Limburg (BE)	Bos, Limburg (I	51.164161	L357p
24	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Wieze	50.977291	O058p
25	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Moorsel	50.944114	O062p
26	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Meerdonk	51.263412	I146p
27	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Meerdonk	51.263412	I146p
28	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Verrrebroek	51.257968	I150p
29	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Kallo	51.255394	I151p
30	454_Paardenbloem	454_Paardenbloem:fleuris	fleuris		Vlaams (WVD)	België	Oost-Vlaande	Beveren	51.213417	I178p

Figure 3. Data export of the dialect words for paardenbloem 'dandelion'

4. Case Study: Searching Herbaria

Herbaria are physical repositories of preserved plant collections that are usually in the form of dried plant specimens mounted on a sheet of paper or book. A herbarium specimen consists of a pressed and dried plant sample that is permanently glued and/or strapped to a sheet of paper along with a documentation label. Herbaria act as time capsules transporting important biodiversity information across time.

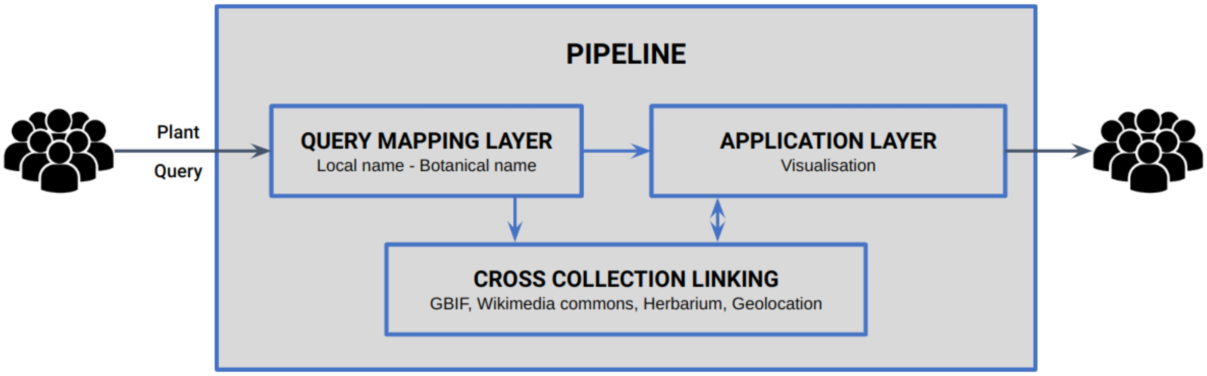
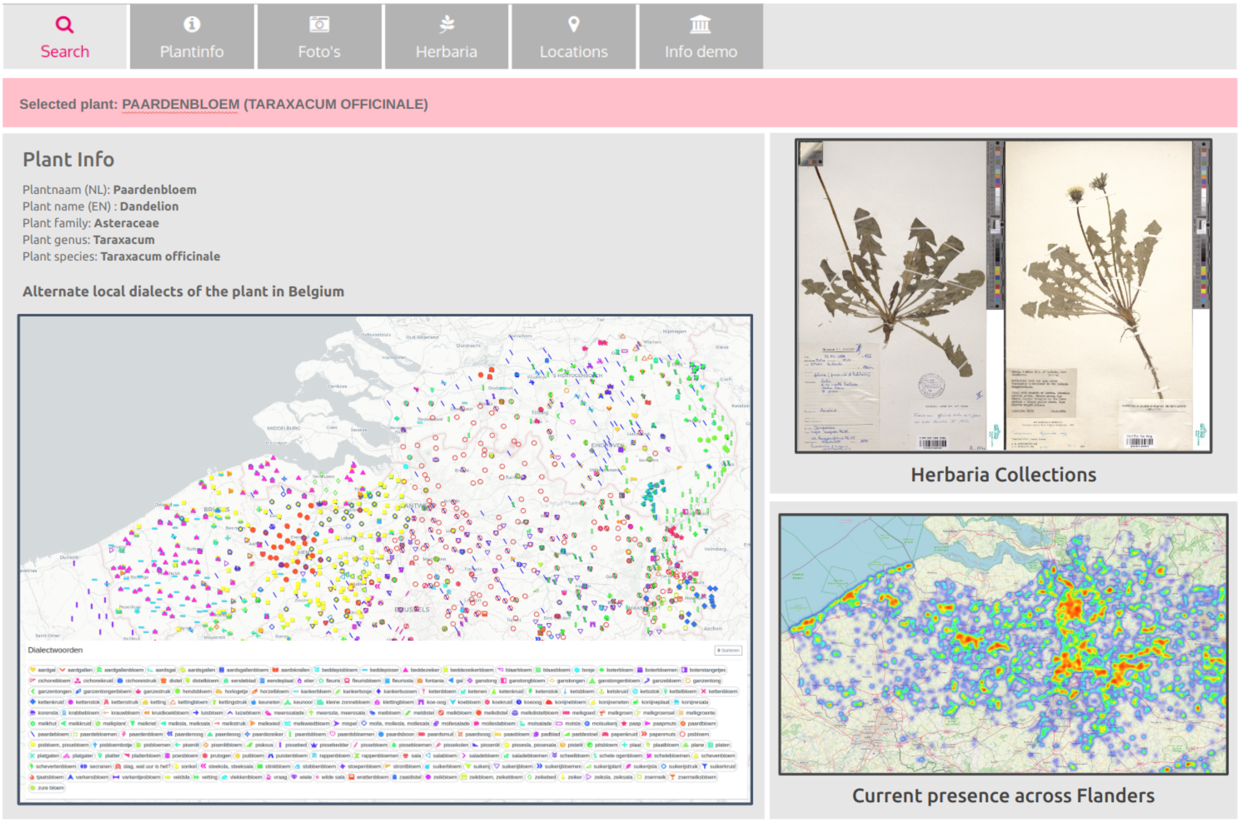


Figure 4. Pipeline with a query mapping layer using DSDD

With the advent of large infrastructure projects such as DiSSCo, there has been an exponential growth in the usage of computerised data information systems to record and access digitised plant specimen collections worldwide. However, the usability of these collections are largely limited since the plants in these collections are indexed using their botanical names that are not often familiar to the general public. Therefore, to improve the reach and usability of such collections, an exploratory study was performed using the dialect data from DSDD to search for plants using local names within Belgium. Figure 4 shows the adapted overall pipeline with a query mapping layer built using DSDD.



Selected plant: PAARDENBLOEM (TARAXACUM OFFICINALE)

Plant Info

Plantnaam (NL): **Paardenbloem**
Plant name (EN) : **Dandelion**
Plant family: **Asteraceae**
Plant genus: **Taraxacum**
Plant species: **Taraxacum officinale**

Alternate local dialects of the plant in Belgium

Figure 5. Result of the pipeline for query word paardenbloem 'dandelion'

As end users will not always be familiar with the botanical name of a plant, a link has been made to the dialect names (almost two hundred different names) of these plants over different regions in Belgium. As such, people can query for plants in their own dialect. The mappings between the original plant names and their dialect alternatives are based on the DSDD. Finally, in order to maximise usability/retrievability, a fuzzy string matching algorithm matches the query input to the plant names in the dataset. This is in turn mapped to the botanical name of the plant. As explained in (Thirukokaranam Chandrasekar et. al, 2021), other existing plant collections such as GBIF[8] and PLANTCOL[9] are also linked to the herbaria that further maximises the usability of these collections. Figure 5 shows a sample outcome of the pipeline.

[8] GBIF - https://www.gbif.org/occurrence/charts?dataset_key=bfc6fe18-77c7-4ede-a555-9207d60d1d86
[9] PLANTCOL Database - https://www.plantcol.be/

5. Future Directions

Future work on the DSDD will consist of adding semasiological dialect dictionary data, starting with *Woordenboek der Zeeuwse Dialecten*. The infrastructure will also be extended in the context of the CLARIAH-PLUS project. By opening up and integrating this unique dialect dataset, a significant milestone towards realising a lexicographical dialect data infrastructure covering the entire Dutch language area has been achieved.

As learnt from the case study, the DSDD layer opens up the possibility to not only link collections across different dimensions but also makes it more usable to the general public. As such, this could be further extended to other digital heritage collections that could result in the creation of multiple cross-domain applications. All the more, DSDD could be considered as a pilot study that could give rise to more such dialect studies across different regions of the world.

Acknowledgement

The initial part of the case study was performed at IDLab - Ghent University under the supervision of Prof. Steven Verstockt and was funded by projects - Floredgand and DiSSCo.

Bibliography

Barbot, L.; Fischer, F.; Moranville, Y and Pozdniakov, I (2019) *Which DH tools are actually used in research?* [Blogpost] weltliteratur.net - A Black Market for the Digital Humanities: https://weltliteratur.net/dh-tools-used-in-research/

De Vriend, F. & L. Boves, H. van den Heuvel, R. van Hout, J. Kruijsen & J. Swanenberg, Jos. (2006). A unified structure for dutch dialect dictionary data. in: *Proceedings of The fifth international conference on Language Resources and Evaluation* (LREC 2006).

De Vriend, F. (2012), *Tools for Computational Analyses of Dialect Geography Data*. PhD Radboud University Nijmegen.

Kruijsen, J. and Sijs, N. van der (2010). Mapping Dutch and Flemish. In: Lameli, Alfred, Kehrein, Roland and Rabanus, Stefan (eds). *Language and Space: An International Handbook of Linguistic Variation: Language Mapping, Handbooks of linguistics and communication science*; 30.2, pages 180–202. De Gruyter Mouton.

Thirukokaranam Chandrasekar, K. K., Deman, E., & Verstockt, S. (2021). Cross-collection linking of botanical imagery in Ghent altarpiece to learn more about van Eyck's masterpiece and to explore a region's plant richness and diversity over time. ACM JOURNAL ON COMPUTING AND CULTURAL HERITAGE, 14(3)

Van den Heuvel, H, E. Sanders en N. van der Sijs (2016), Curation of Dutch Dialect Dictionaries In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 16).

Van Hout, R, N. van der Sijs, E. Komen en H. van den Heuvel (2018), A fast and flexible web interface for dialect research in the Low Countries. In: *Proceedings of The fifth international conference on Language Resources and Evaluation* (LREC 18).

Van Keymeulen, J. (2004), Trefwoorden en lexicale varianten in de grote regionale woordenboeken van het zuidelijke Nederlands (WBD, WLD, WVD). In: De Caluwe J, G. De Schutter, M. Devos en J. Van Keymeulen*, Taeldeman, man van de taal, schatbewaarder van de taal.* Vakgroep Nederlandse Taalkunde UGent – Academia Press, Gent (2004); 897-908.

Van Keymeulen, J., V. De Tier, R. Vandenberghe & S. Chambers (2019), The dictionary of the Southern Dutch Dialects (DSDD): designing a virtual research environment for digital lexicological research. in: *Dialectologia. Special issue*, 8 (2019), 93-115.