/instituut voor
de Nederlandse
taal/

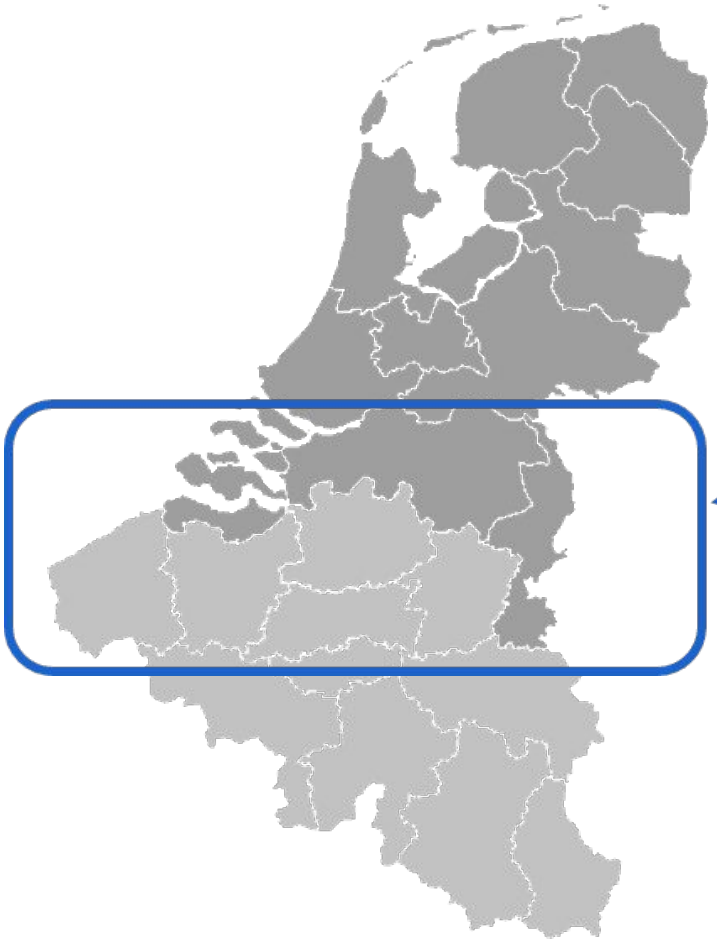
digital humanities Benelux
Leiden 2021

UNIVERSITEIT
GENT

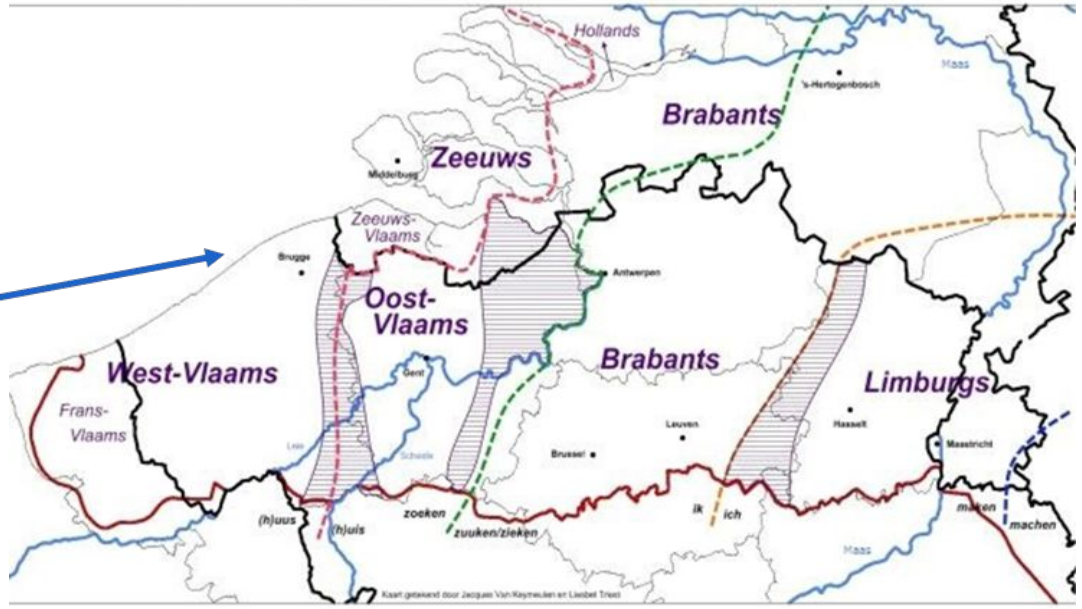
Interdisciplinary synergies: opening up new perspectives for data-driven research of the Southern Dutch Dialects

Veronique De Tier, Jesse de Does, Katrien Depuydt,
Tanneke Schoonheim (†), Koen Mertens, **Dutch Language Institute**,
Jacques Van Keymeulen, Roxane Vandenberghe, Lien Hellebaut,
Department of Linguistics, Ghent University and Sally Chambers,
Ghent Centre for Digital Humanities, Ghent University

Southern Dutch Dialects

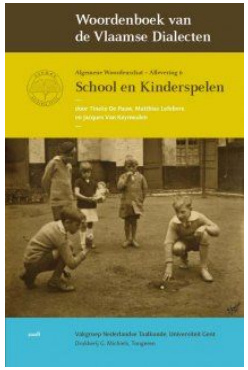


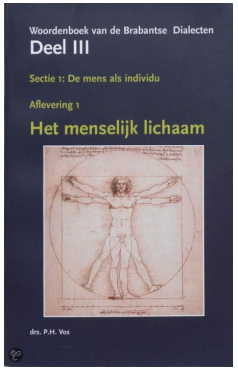
= **four dialect groups**: Flemish, Zeelandic, Brabantic, Limburgian

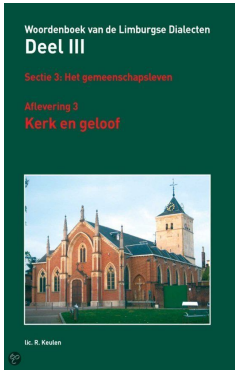


Spoken in the Dutch speaking part of Belgium, the southern part of the Netherlands and the northern part of France

Onomasiological (thematic) dictionaries



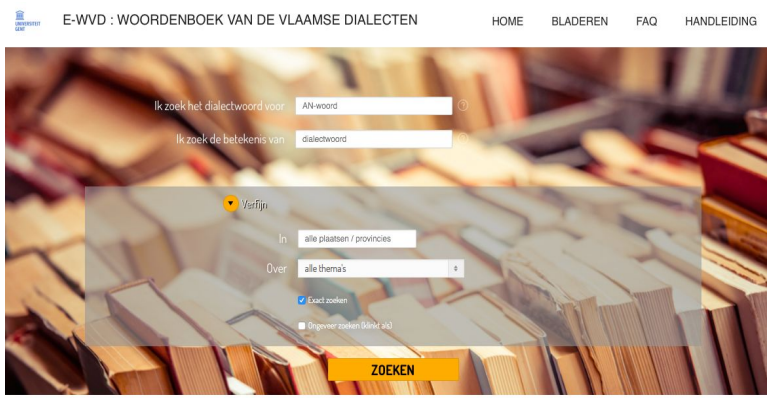




THEMATIC

Zeelandic dictionary:
ALPHABETICAL

own database + website



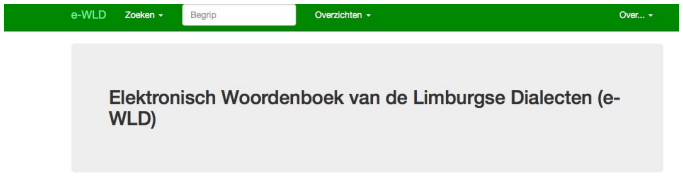
Het e-WVD is de online database van het Woordenboek van de Vlaamse Dialecten (WVD), een woordenboekproject dat in 1972 aan de Universiteit Gent werd opgestart en in 2018 werd afgerond. Het WVD verzamelt de dialectwoordenschat uit Oost-Vlaanderen, West-Vlaanderen, Frans-Vlaanderen en Zeeuws-Vlaanderen. De dialectwoordenschat van Antwerpen, Vlaams-Brabant, Noord-Brabant en de beide Limburgen is te vinden in de (digitale) woordenboeken van de twee zusterprojecten: het Woordenboek van de Brabantse Dialecten (e-WBD) en het Woordenboek van de Limburgse Dialecten (e-WLD).

Het e-WBD bevat de inhoud van de drie delen van het *Woordenboek van de Brabantse dialecten* die in druk in 31 afleveringen zijn verschenen tussen 1967 e 2005 (zie tabblad Overzichten). Daarnaast zijn er in 1967 en 2000 twee aflevering gepubliceerd die geen taalmateriaal bevatten, maar inleidingen. In het e-WBD ku geïnteresseerden de Brabantse **trefwoorden** opzoeken voor bepaalde **begrippe** zoals 'appelstroop' of 'persstro'. Daarbij worden ook de verschillende fonetische varianten, de zogenoemde **dialectopgaven**, van de trefwoorden getoond.

Het is ook mogelijk alle trefwoorden van één bepaalde Brabantse **plaats** op te zoeken, hetzij door te zoeken op de plaatsnaam, hetzij door te zoeken op de **Kloekecode** van de bewuste plaatsnaam. De Kloekecodes zijn unieke nummeri die de dialectoloog G.G. Kloeke in 1926 heeft toegekend aan de plaatsen en gehuchten in Nederland, Vlaanderen, Frans-Vlaanderen en het noordwesten van Duitsland. Ze bestaan uit een hoofdletter gevolgd door drie cijfers met soms een kleine letter daaraan toegevoegd. Omdat de oorspronkelijke nummering in de lo van de tijd is uitgebreid, bestaan er een naast de oude Kloekecodes ook nieuwe Kloekecodes; in de zoekapplicatie worden beide versies getoond.

Het e-WBD bevat 15.794 begrippen, 140.091 trefwoorden en 1.704.116 dialectopgaven, verzameld in ruim 2950 Brabantse plaatsen ofwel dialecten (ied plaats vertegenwoordigt een eigen dialect).

De website met e-WBD is op 14 december 2017 feestelijk gelanceerd tijdens een dialectencongres in Den Bosch. De website vormt een aanvulling van de eind 2016 gelanceerde website e-WLD, waarop het Woordenboek van de Limburgse Dialecten doorzoekbaar is gemaakt.



Het e-WLD bevat de inhoud van de drie delen van het *Woordenboek van de Limburgse dialecten* die in druk in 39 afleveringen zijn verschenen tussen 1983 en 2008 (zie tabblad Overzichten). In het e-WLD kunnen geïnteresseerden de Limburgse **trefwoorden** opzoeken voor bepaalde **begrippen**, zoals 'appelstroop' of 'persstro'. Daarbij worden ook de verschillende fonetische varianten, de zogenoemde **dialectopgaven**, van de trefwoorden getoond.

Het is ook mogelijk alle trefwoorden van één bepaalde Limburgse **plaats** op te zoeken, hetzij door te zoeken op de plaatsnaam, hetzij door te zoeken op de **Kloekecode** van de bewuste plaatsnaam. De Kloekecodes zijn unieke nummeringen die de dialectoloog G.G. Kloeke in 1926 heeft toegekend aan de plaatsen en gehuchten in Nederland, Vlaanderen, Frans-Vlaanderen en het noordwesten van Duitsland. Ze bestaan uit een hoofdletter gevolgd door drie cijfers met soms een kleine letter daaraan toegevoegd. Omdat de oorspronkelijke nummering in de loop van de tijd is uitgebreid, bestaan er een naast de oude Kloekecodes ook nieuwe Kloekecodes; in de zoekapplicatie worden beide versies getoond.

Het e-WLD bevat 17.539 begrippen, 137.231 trefwoorden en 1.759.090 dialectopgaven, verzameld in ruim 1000 Limburgse plaatsen ofwel dialecten (iedere plaats vertegenwoordigt een eigen dialect). Uniek voor het Limburgs is de mijnbouw, waaraan aflevering II.5 is gewijd. In de negentien **mijnen** (12 in Nederlands Limburg en 7 in Belgisch Limburg) werd een bijzondere vorm van het dialect gesproken, en daarmee vormen die mijnen een uitbreiding van het Limburgse plaatsennetwerk. Het is mogelijk te zoeken op de gegevens van deze mijnen; dat levert dan alleen gegevens op afkomstig uit aflevering II.5.

De website met e-WLD is op 17 december 2016 feestelijk gelanceerd tijdens een dialectencongres in Venlo.

e-wvd.be / e-wbd.nl / e-wld.nl

Aims of the DSDD

- **Database of Southern Dutch Dialects (DSDD):** to aggregate and standardise three existing dialect databases of the Flemish, Brabantic and Limburgian dialects into one integrated dataset

- Provide access to the **integrated DSDD dataset** via a **user-friendly search engine** and a **geo-visualisation tool**

- Enable **export subsets of the data** via an **Application Programming Interface (API)** using **existing digital research tools**

To achieve this it was essential to bring together researchers from several disciplinary areas ...

Interdisciplinary Methodological Approach

The consortium consisted of researchers from four core disciplinary areas:

- **Dialectologists** for a thorough understanding of the dialect words and the lexicographical construction of the dictionaries and **computational linguists** for the modelling of dialect data and the conceptual understanding of the linguistic make-up of the dictionaries
- **Cartographers** to identify appropriate web-mapping technologies
- **Computer and data scientists** to understand how to manage and inter-link the various linguistic concepts and dialect words using linked open data technologies, plus provide access to the data via an API
- **Digital humanities experts** 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 or collaborative notebooks.

Interdisciplinary Methodological Approach

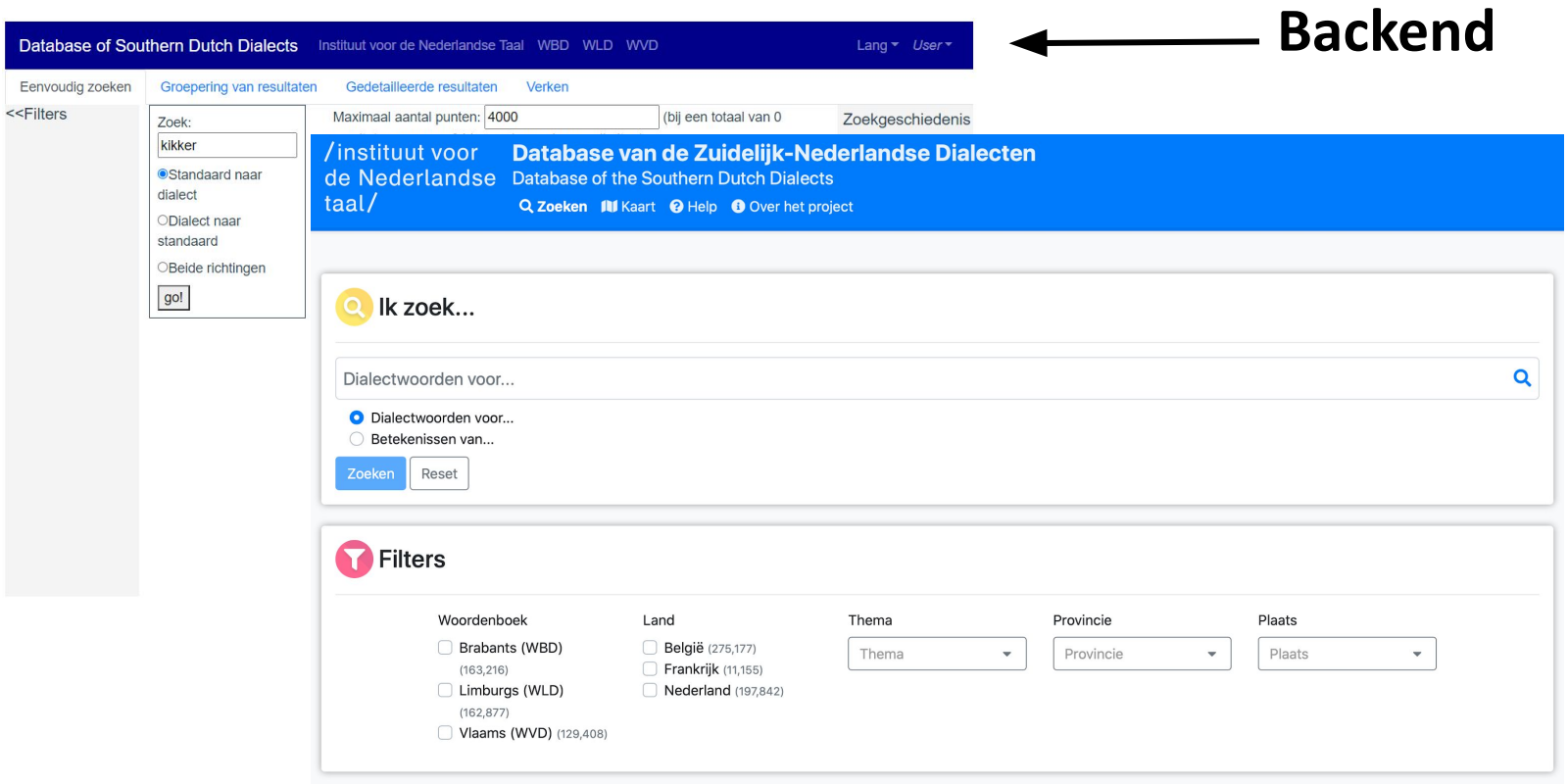
- **Interdisciplinary collaboration was essential**: due to the skills and expertise needed. Carrying out this project by any of the individual research groups alone would not have been possible.

- **Combining the strengths of the different research teams:** resulted in new approaches to geo-visualisation and inter-linking of dialect data

- **Digital humanities:** the inherently collaborative nature was the catalyst for reaching out beyond our disciplinary boundaries, on the whole this occurred naturally

- **Challenges**: achieving a high-enough level of domain understanding to enable common discussions.

DSDD Search Engine (and backend)

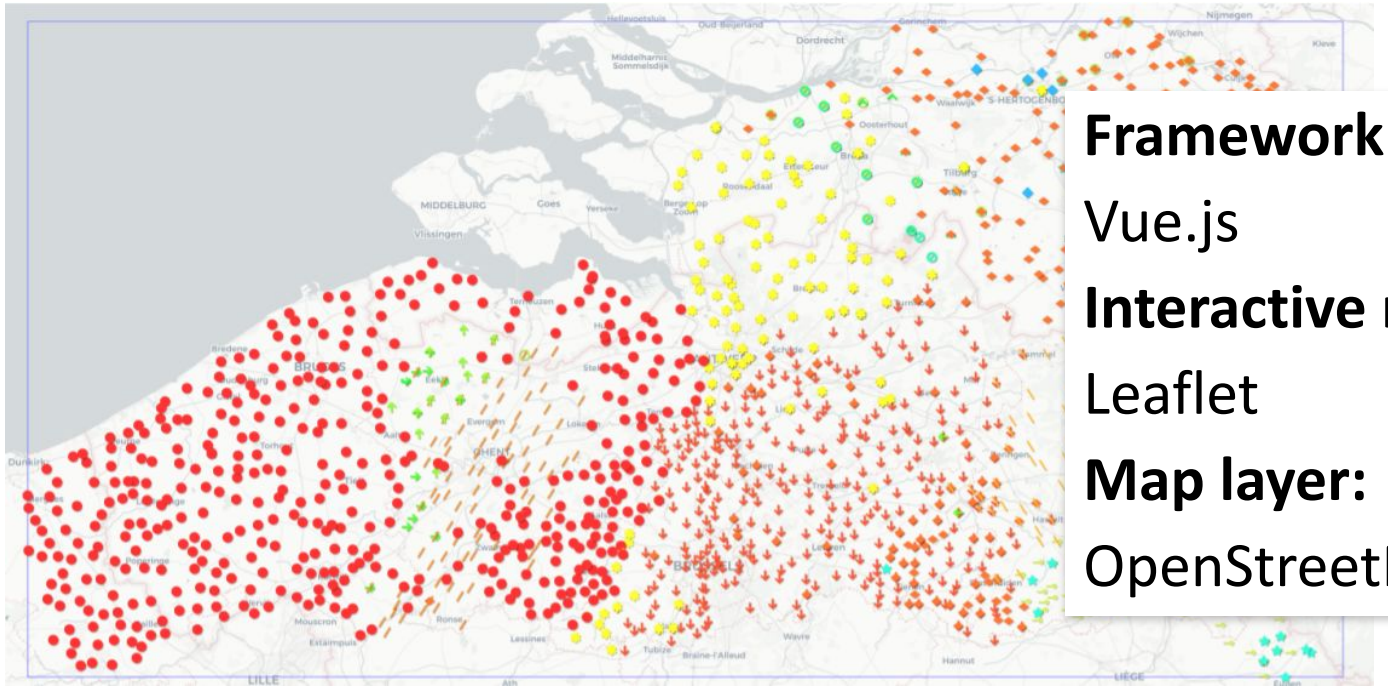


Backend

https://dsdd.ivdnt.org/DSDD/

Powered by Solr

Geo-visualisation of dialect words for ‘kikker’ (frog)



Framework:

Vue.js

Interactive map:

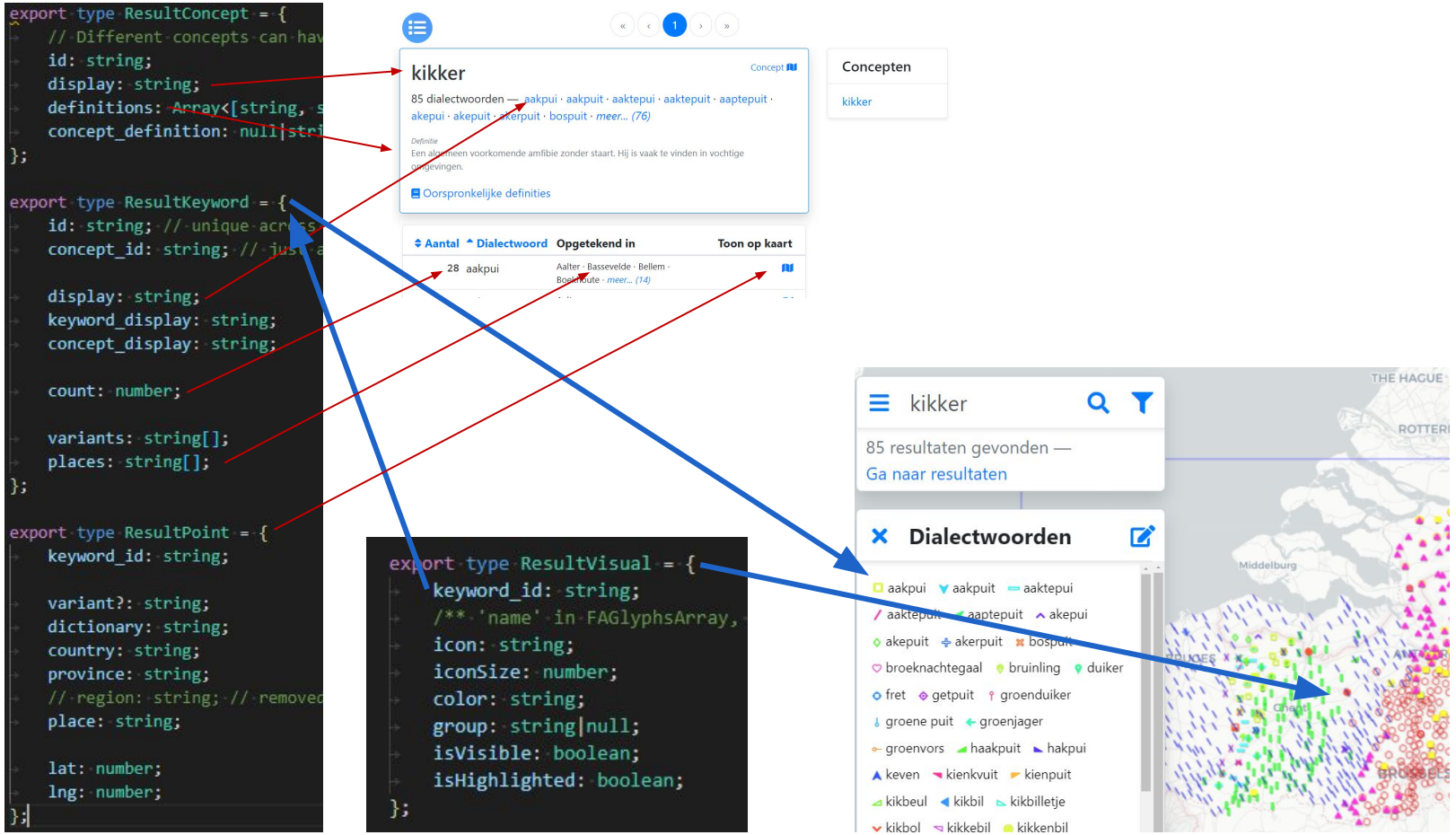
Leaflet

Map layer:

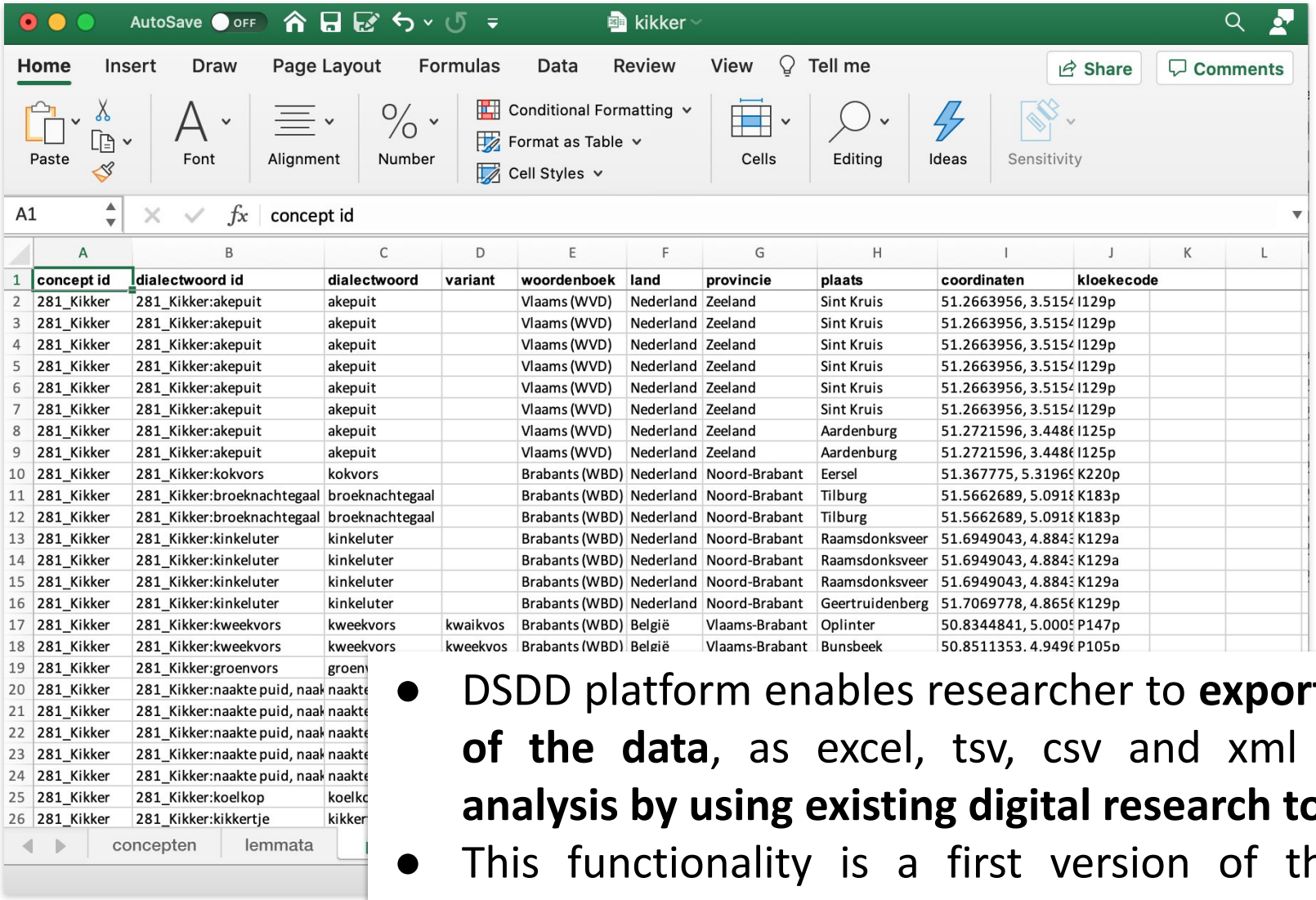
OpenStreetMap

puid, puit | vors | kikvors | pui | kwakvors | kwakkerd | puit | kwakvros | kikker | aakpui | kwakker | aaktepui | kinkvors | aaktepuit
kwekkerd | paddenmoek | kinkpuit | vros | kikbil | kikmauw | kikvorsje | kruts | kwekker | akepuit | kwak | puiden, puiten | puitje | kikkerd
kinkeluut | kinkenduut | naakte puid, naakte puit | kinkeluter | naaktepui | aaptepuit | groenjager | kikpuit | kikvros | lookvors | pad | vorsje
aakpuit | akepui | broeknachtegaal | keven | kikbol | kikkenbil | kink | kweekvors | puider | weerkwakkerd | akerpuit | bospuit | bruinling
duiker | fret | getpuit | groenduiker | groene puit | groenvors | haakpuit | hakpui | kienkvuit | kienpuit | kikbeul | kikbilletje | kikkebil
kikkenduut | kikkertje | kinkerd | klabotskop | koelekop | koelkop | kokkerdil | kokvors | kroddel | kwaker | kwek | loker | meun | naakpui
naakpuit | puilodder | sprinkhaan | vek | vrosje

Backend database powers both the DSDD search engine & cartographic tool



Data-level access to the integrated dataset

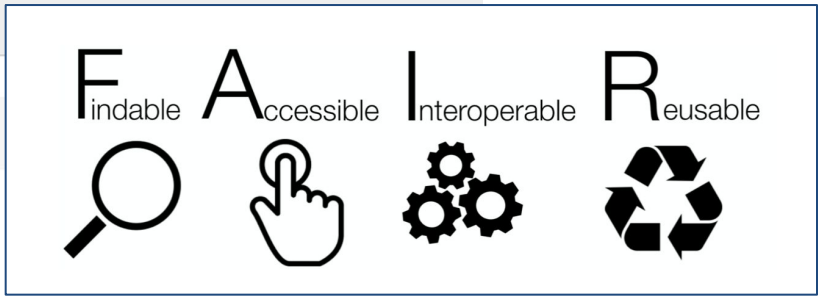


concept id	dialectwoord id	dialectwoord	variant	woordenboek	land	provincie	plaats	coordinaten	kloekecode
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Sint Kruis	51.2663956, 3.5154	I129p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Aardenburg	51.2721596, 3.4486	I125p
281_Kikker	281_Kikker:akepuit	akepuit		Vlaams (WVD)	Nederland	Zeeland	Aardenburg	51.2721596, 3.4486	I125p
281_Kikker	281_Kikker:kokvors	kokvors		Brabants (WBD)	Nederland	Noord-Brabant	Eersel	51.367775, 5.3196	K220p
281_Kikker	281_Kikker:broeknachtegaal	broeknachtegaal		Brabants (WBD)	Nederland	Noord-Brabant	Tilburg	51.5662689, 5.0918	K183p
281_Kikker	281_Kikker:broeknachtegaal	broeknachtegaal		Brabants (WBD)	Nederland	Noord-Brabant	Tilburg	51.5662689, 5.0918	K183p
281_Kikker	281_Kikker:kinkeluter	kinkeluter		Brabants (WBD)	Nederland	Noord-Brabant	Raamsdonksveer	51.6949043, 4.8843	K129a
281_Kikker	281_Kikker:kinkeluter	kinkeluter		Brabants (WBD)	Nederland	Noord-Brabant	Raamsdonksveer	51.6949043, 4.8843	K129a
281_Kikker	281_Kikker:kinkeluter	kinkeluter		Brabants (WBD)	Nederland	Noord-Brabant	Raamsdonksveer	51.6949043, 4.8843	K129a
281_Kikker	281_Kikker:kinkeluter	kinkeluter		Brabants (WBD)	Nederland	Noord-Brabant	Geertruidenberg	51.7069778, 4.8656	K129p
281_Kikker	281_Kikker:kweekvors	kweekvors	kwaikvos	Brabants (WBD)	België	Vlaams-Brabant	Oplinter	50.8344841, 5.0005	P147p
281_Kikker	281_Kikker:kweekvors	kweekvors	kweekvos	Brabants (WBD)	België	Vlaams-Brabant	Bunsbeek	50.8511353, 4.9496	P105p
281_Kikker	281_Kikker:groenvors	groenv							
281_Kikker	281_Kikker:naakte puid, naak	naakte							
281_Kikker	281_Kikker:naakte puid, naak	naakte							
281_Kikker	281_Kikker:naakte puid, naak	naakte							
281_Kikker	281_Kikker:naakte puid, naak	naakte							
281_Kikker	281_Kikker:naakte puid, naak	naakte							
281_Kikker	281_Kikker:koelkop	koelko							
281_Kikker	281_Kikker:kikkertje	kikker							

- DSDD platform enables researcher to **export subsets of the data**, as excel, tsv, csv and xml files for **analysis by using existing digital research tools**.
- This functionality is a first version of the **DSDD Application Programming Interface (API)** for access to larger datasets.

Reusability of the DSDD Dataset

CLARIN | About | Participants | Services | Knowledge Base | Funding | Events



Home > Knowledge Base > CLARIN for Researchers

Resource families

We currently offer overviews of 11 corpora families, 5 families of lexical resources, and 4 tool families:

Corpora
- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

www.clarin.eu/resource-families

Future Directions

- **Next steps:** integrating the remaining data of the three dictionaries and fine-tuning the concept layer of the DSDD database

- **Future work:** 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.

- This unique dataset paves the way for **comparative research and geo-visualisation of dialect data and linguistic variation**, for example, **across Germanic languages.**

- Opening up and integrating this **unique dataset** is a significant milestone towards realising the vision of a **lexicographical dialect data infrastructure covering the entire Dutch language area**.

/instituut voor
de Nederlandse
taal/

digital humanities Benelux
Leiden 2021

UNIVERSITEIT
GENT

Thanks for your attention!

Veronique De Tier, Jesse de Does, Katrien Depuydt,
Tanneke Schoonheim (†), Koen Mertens, **Dutch Language Institute**,
Jacques Van Keymeulen, Roxane Vandenberghe, Lien Hellebaut,
Department of Linguistics, Ghent University and Sally Chambers,
Ghent Centre for Digital Humanities, Ghent University

/instituut voor
de Nederlandse
taal/

digital humanities Benelux
Leiden 2021

UNIVERSITEIT
GENT

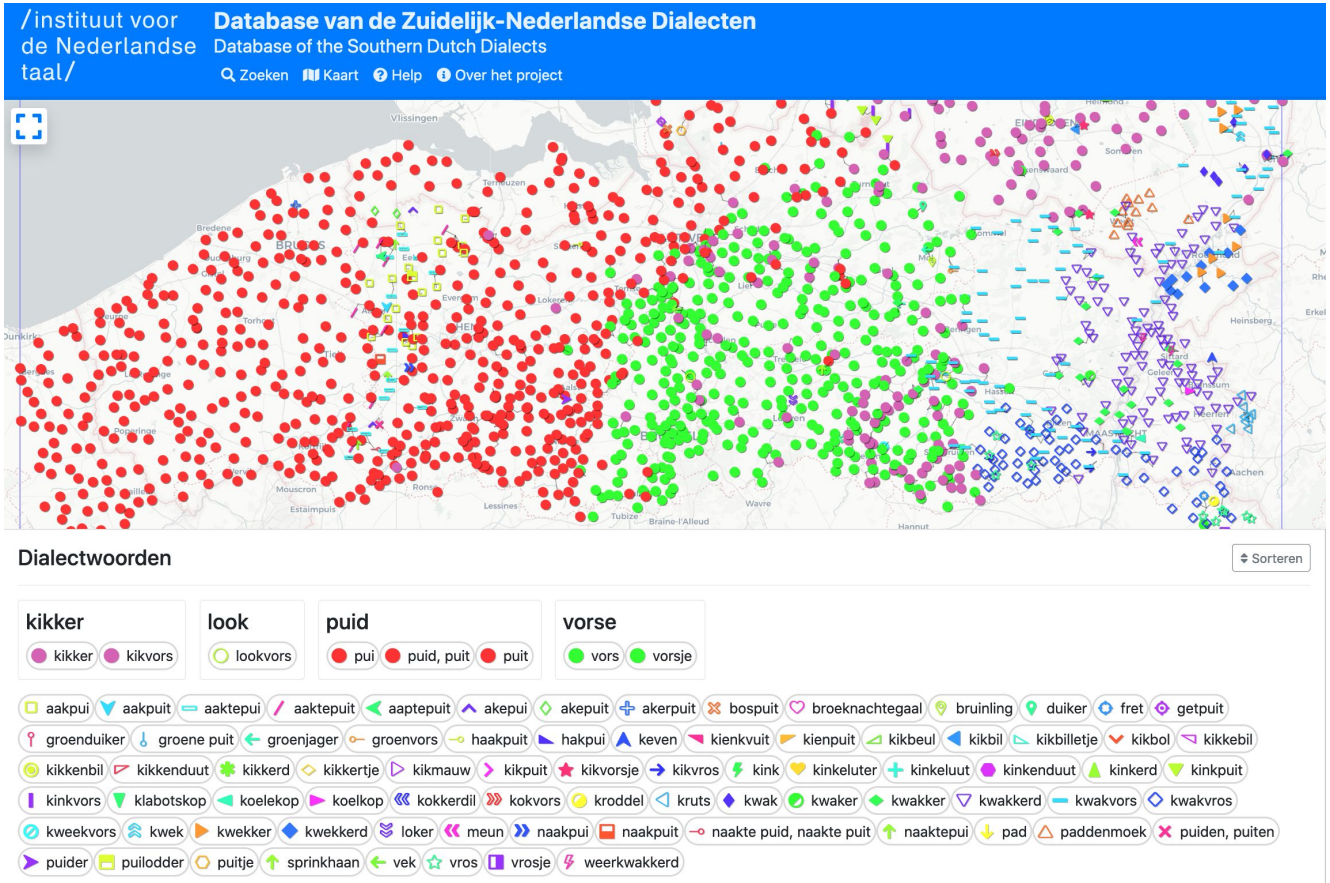
Interdisciplinary synergies: opening up new perspectives for data-driven research of the Southern Dutch Dialects

Veronique De Tier, Jesse de Does, Katrien Depuydt,
Tanneke Schoonheim (†), Koen Mertens, **Dutch Language Institute**,
Jacques Van Keymeulen, Roxane Vandenberghe, Lien Hellebaut,
Department of Linguistics, Ghent University and Sally Chambers,
Ghent Centre for Digital Humanities, Ghent University

Interdisciplinary synergies: opening up new perspectives for data-driven research of the Southern Dutch Dialects

- **Database of Southern Dutch Dialects (DSDD):** three existing dialect databases of the Flemish, Brabantic and Limburgian dialects into one integrated dataset
- **Interdisciplinary Methodological Approach** (dialectologists, cartographers, computer and data scientists, digital humanities experts)
- **Combining the strengths of the different research teams** resulted in new approaches to geo-visualisation and inter-linking of dialect data
- Resulted in a **harmonised dataset of concepts** and the **DSDD Platform**: a user-friendly search engine, geo-visualisation tool and an API
- Opening up and integrating this **unique dataset** is a significant milestone towards realising the vision of a **lexicographical dialect data infrastructure covering the entire Dutch language area**.

Interdisciplinary synergies: opening up new perspectives for data-driven research of the Southern Dutch Dialects



https://dsdd.ivdnt.org/DSDD/