

Christina Mutter | Aleksander Wiatr

# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function

<http://www.verba-alpina.gwi.uni-muenchen.de>

Journée d'étude « Picard et ressources numérique »  
18th september 2018, Maison de la Recherche, Lille





## Outline

### 1) Project description

- team
- area under investigation
- research aims
- data and methodology

### 2) Lexicographic function

- transcription (analogue/digital data)
- tokenization
- typification
- data access (interactive map/database)



### 3) Approaches to sustainability

- versioning
- citability
- long-term archiving



## 1) Project description

- *VerbaAlpina. Der alpine Kulturreraum im Spiegel seiner Mehrsprachigkeit* (VerbaAlpina. The Alpine cultural region reflected through its multilingualism)
- Funded by the German Research Foundation (DFG)
- 1<sup>st</sup> term: 10/2014-10/2017, 2<sup>nd</sup> term: 11/2017-11/2020 (perspective until 2025)
- Investigation of the multilingual Alpine region
- Combination of (geo-)linguistics and digital humanities



## Team

### Project leaders

- Prof. Dr. Thomas Krefeld (Institute of Romance Studies)
- Dr. Stephan Lücke (LMU Center for Digital Humanities)

### Member of staff

- David Englmeier (computer science)
- Markus Kunzmann (German studies)
- Christina Mutter (scientific coordination, Romance studies)
- Aleksander Wiatr (Romance/Slovenian studies)
- Florian Zacherl (computer science)
- Alessia Brancatelli, Julie Defert, Monika Hausmann, Filip Hristov, Katharina Knapp, Marina Pantele, Daniela Warras (scientific assistants)

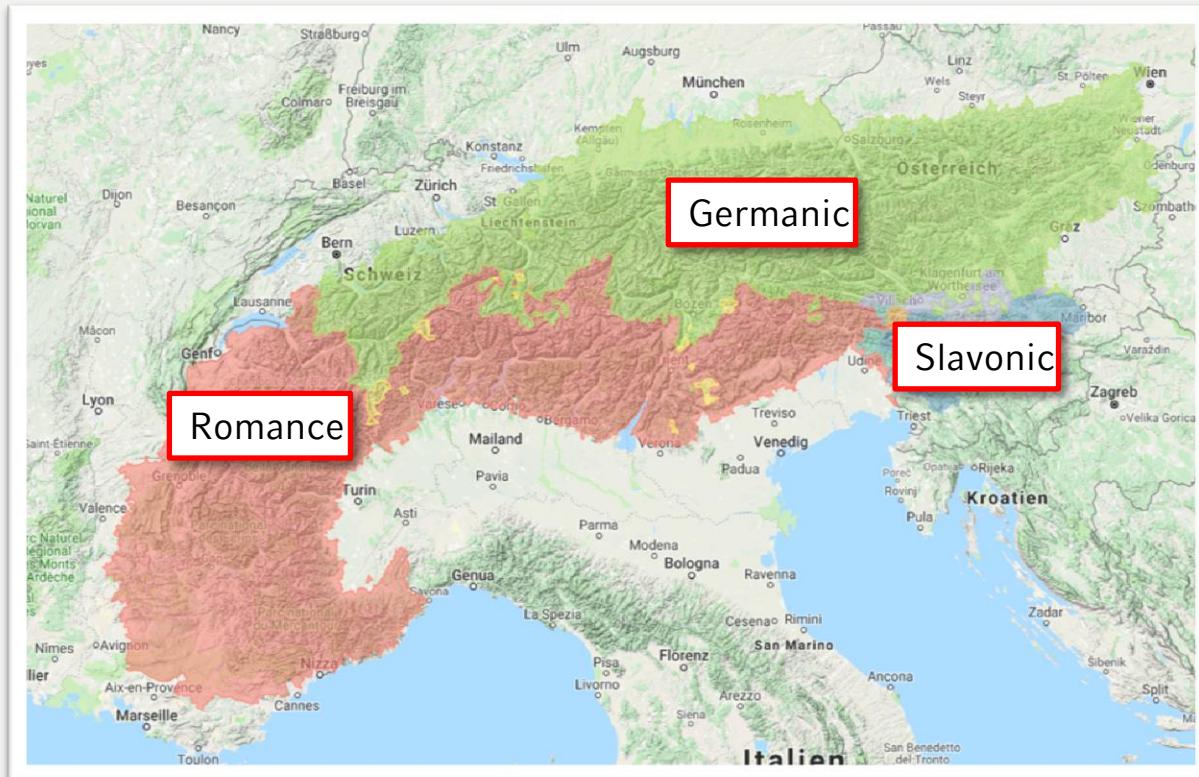


## Area under investigation: The Alpine region

- Area of investigation is limited to the territorial borders defined by the Alpine convention
- surface area of 190,600 km<sup>2</sup>, encompasses parts of six different countries (D, A, CH, I, F, SLO) and two entire countries (FL, MC)



- ethnographic and topographic homogeneity and strong linguistic heterogeneity → 3 language families (Germanic, Romance and Slavonic)





## Three conceptual domains

project years	1	2	3	4	5	6	7	8	9	
calendar year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
quarter	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv	i, ii, iii, iv
project phase	I			II			III			
focus	<b>culture</b> <ul style="list-style-type: none"> <li>• alpine pasture farming</li> <li>• milk processing</li> </ul>			<b>nature</b> <ul style="list-style-type: none"> <li>• landscape formations</li> <li>• weather</li> <li>• fauna</li> <li>• flora</li> </ul>			<b>modern life</b> <ul style="list-style-type: none"> <li>• ecology</li> <li>• tourism</li> </ul>			



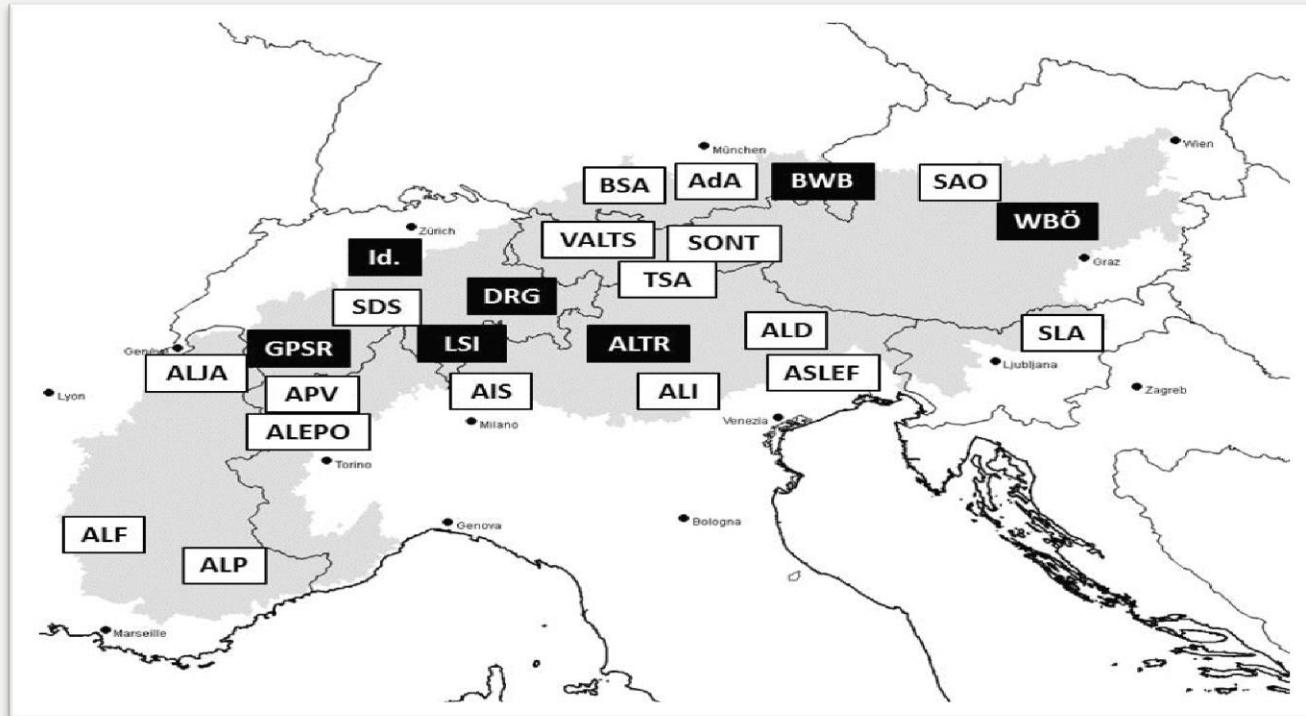
## Research Aims

- Selective and analytical investigation of the linguistically and dialectally highly fragmented alpine space in its historico-cultural and historical-linguistic unity
- Overcoming of the traditional limitation of geolinguistic investigation to nation-states
- recognition of connections regarding the etymology of the individual dialectical words
- Setting up a portal by using modern media technology: documentation, data collection, collaborative development
- cooperation with other projects is fundamental for VerbaAlpina



## Data and methodology

- Collection and analysis of data from linguistic atlases and from geo-referenced dictionaries from the past one hundred years





## Data and methodology

- **Online-Crowdsourcing:** to even out, complete and correct inhomogenous data stock
- Combination of three different approaches of digital geolinguistics:
  - digitally published atlases (data gathered through traditional methods, e.g. ALD)
  - atlases which document diverse languages and language families (e.g. WALS)
  - web-based atlases (e.g. AdA)



## Data and methodology

- challenge:

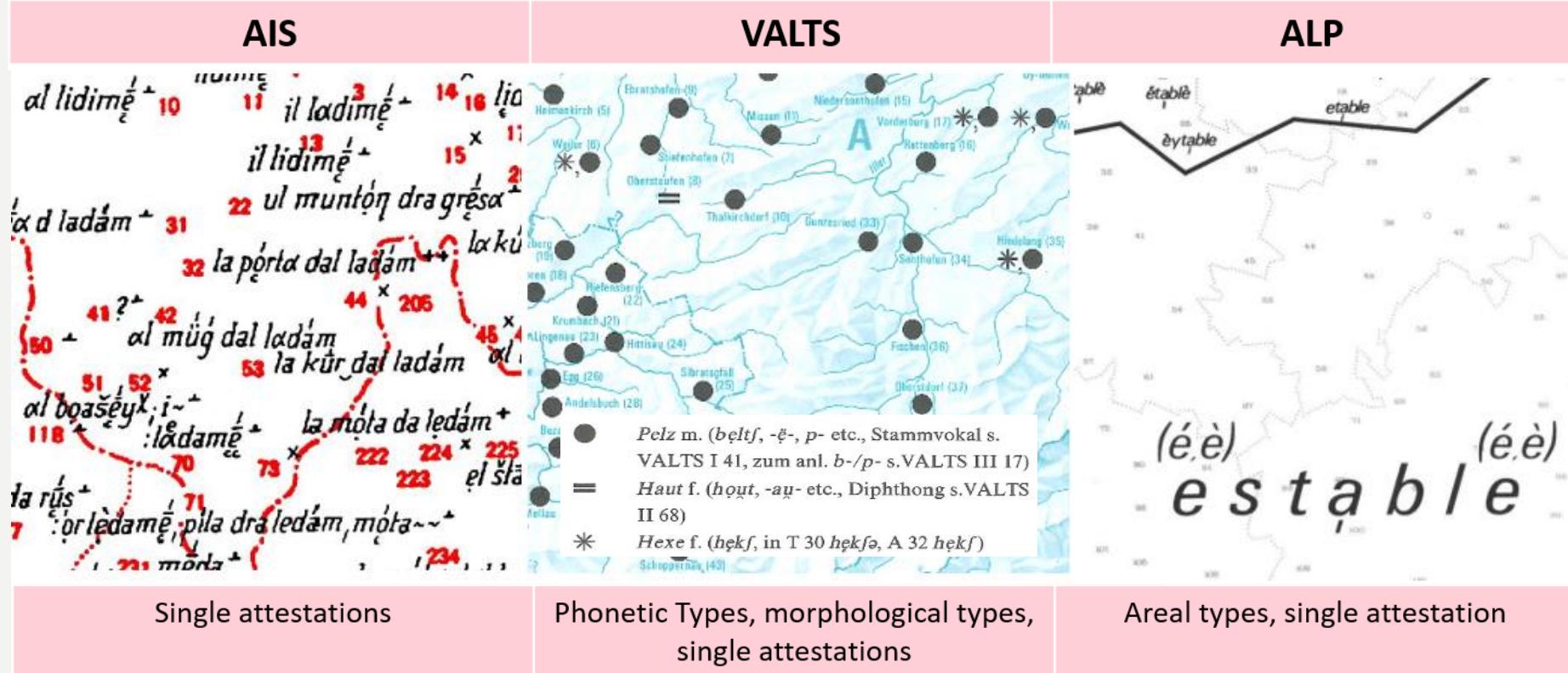
lack of uniformity of data from individual data sources

- unification of the different transcription systems
- process of systematic data processing:

- Transcription
- Tokenization
- Typification



## Lexicographic Process





## Lexicographic Process: Transcription Tool

Meine Websites VerbaAlpina 14 + Neu Willkommen, AWiatr

1 von 1 20% KARTE 1207

VI. BAND SPRACH- UND SACHATLAS ITALIENS UND DER SÜDSCHWEIZ.

**IL BURRO**  
BUTTER — BEURRE

Ga. Hib. 155—ALF 130 — Bloch I —  
Brun. 141  
28,5 — 79,13 — 7,4

**Legende:**

Die Butter wird in den Provinzen Italiens, wo Bl produziert oder wo das Öl beim Kochen verwendet wird, selten oder gar nicht gebraucht. In den Städten Mittelitaliens, seltener Südtirols, ist die Butter als neuerdings eingeführtes Produkt bereits bekannt, aber diese Verwendung in der Küche bleibt bei der bodenständigen Bevölkerung wenig üblich. Einem Einblick in die tatsächlichen Verhältnisse gewähren uns auch die X. 1200 (Gebet), 2005 (abenteuert), 2006 (Butterbrot).

Käsemantel mit Butterinhalt  
(burro rucioso in una forma di pane di ceci) Die Butter ist in einen breitflächigen Mantel eingeschlossen. Vgl. Strizz u. X.M., 2006 (ley)

221 a mandolino 222 a mandolino  
223 la baviera 224 a mandolino  
225 a mandolino (selten) 226 a mandolino, u baviera  
226 la mandolino 227 fritt. \* -drost 228 semantel mit But.  
228 la mandolino 229 kerinhalb, Butter-  
inhalt)  
für den Gebrauch 230 a mandolino

**AIS**

AIS#1207 3 (il pane di burro)

**AIS 1207 - il pane di burro - Informant\_Nr 1 (Brigels-Breil)**

Transkription  Eintragen <vacat> Problem Beleg

Zugewiesene Konzepte !

**BUTTERDALENT**

Konvention für die Transkription aller für VA relevanten Sprachadventen

Version vom 05.12.2016

Wir unterscheiden zwischen Basiszeichen und Diakritika. Alle Zeichen, die sich nicht auf der Grundlinie befinden, werden als Diakritika bezeichnet.

Basiszeichen, die in der ASCII-Tabelle vorhanden sind, werden beibehalten (= alle lateinischen Buchstaben; „nicht“ deutsche Umlaute)

Folgende Basiszeichen, die nicht in der ASCII-Tabelle enthalten sind, werden wie folgt transkribiert (in der hinteren Spalte steht der Sprachatlas, in dem das jeweilige Basiszeichen bzw. Diakritikum zum ersten Mal „entdeckt“ wurde)

	Griechisches Alpha	a1	AIS
α	spiegelverkehrtes a	a2	TSA
β	Griechisches Beta	b1	AIS
γ		b2	SLA
δ	Griechisches Gamma	g1	AIS
ε		g2	ALF
γ		g3	SLA
ζ		g4	SLA
η	Griechisches Delta	d1	AIS
θ		d2	SLA
ι		i1	SLA
λ		i2	SLA
ɔ		ii	SLA
ø	Griechisches Theta	t1	AIS
ø	geschweifte wa	t2	SLA



## Lexicographic Process: Beta Code

Diacritics

Beta Co

Base sign

Diacritics

- No |
- Inst |
- The |
- trans |
- For |
- allow |

	erhebung	beta	ipa	Hex_ipa	Art
AIS			'	&#x0020;	Trennzeichen
AIS	/		'	&#x2c8;	Akzent
AIS	//		'	&#x2c8;&#x2c8;	Akzent
AIS	\		.	&#x2cc;	Akzent
AIS	a		a	&#x0061	Vokal
AIS	a!		a	&#x0061;	Zeichen
AIS	a%		ä	&#x0061;&#x0068;&#x0306;	Zeichen
AIS	a%0		ä	&#x0061;&#x0068;&#x0306;	Zeichen
AIS	a&		e	&#x0250;	Zeichen
AIS	a(		a	&#x0251;	Zeichen
AIS	a(*)		a	&#x0251;	Zeichen
AIS	a(-		a:	&#x0251;&#x02d0;	Zeichen
AIS	a(-{i})		ä	&#x0251;&#x02d0;&#x0069;&#x0306;	Zeichen
AIS	a(-		e:	&#x0250;&#x02d0;	Zeichen
AIS	a(:		æ	&#x00e6;	Zeichen
AIS	a({i})-		ä	&#x0251;&#x02d0;&#x0069;&#x0306;	Zeichen
AIS	a( -		e:	&#x0250;&#x02d0;	Zeichen
AIS	a)		a	&#x0251;	Zeichen
AIS	a)-		a:	&#x0251;&#x02d0;	Zeichen
AIS	a*		a	&#x0061;	Zeichen
AIS	a*(		ä	&#x0251;	Zeichen
AIS	a*-		a:	&#x0061;&#x02d0;	Zeichen
AIS	a*?-		a:	&#x0061;&#x02d0;	Zeichen
AIS	a*:		æ	&#x00e6;	Zeichen
AIS	a*?-		a:	&#x0061;&#x02d0;	Zeichen
AIS	a -		a:	&#x0061;&#x02d0;	Zeichen
AIS	a( -		a:	&#x0251;&#x02d0;	Zeichen

?\\

ibed

e



## Lexicographic Process: Tokenization

Attestation in Beta Code	Attestation in IPA	Concept
una1 mu:g/a1 da1 va/c)/	una mydʒa da v'atç	HERD OF COWS
TOKENIZATION		
una1	una	ARTICLE
mu:g/a1	mydʒa	HERD
da1	da	PREPOSITION
va/c)/	v'atç	COW



## Lexicographic Process: Typification

- **Morpho-lexical type: Orthography, language family, POS, gender, affix + base type (lexical base)**

	<i>barga</i>	<i>barg</i>	<i>margin</i>	<i>bargun</i>
<b>language family</b>	roa	roa	roa	roa
<b>PoS</b>	noun	noun	noun	noun
<b>affix</b>	-	-	+	+
<b>gender</b>	f	m	m	M
<b>morpho-lexical type</b>	1	2	3	3



\**barga*/\**barca*



**706 LE FROMAGE**

Q. 117, 1  
ALF 613  
AIS 1209, 1215\*  
ALLY 399  
ALMG 1007  
943

942  
ādzó, tōmmé f +

+  
950  
rōmādzu, tūmà f ≠

nō f +

953  
frōmāzo, tōn

942  
ādzó, tōmmé f +

+  
950  
rōmādzu, tūmà f ≠

nō f +

869  
frōmādze, tūmà f +

88  
frōmādze, tūmà f + 879  
i, tūmò f +

f +

877

**S.**

**H.-ALPES**

**B.-ALPES**

1 frōmēj  
15 lē frōmēj fōr  
5  
6  
lū frōmēj<sup>3</sup>  
lū frōmāju<sup>2</sup>  
lū frōmēdzu<sup>1</sup>  
lē frōmēdzu dē gruyèr

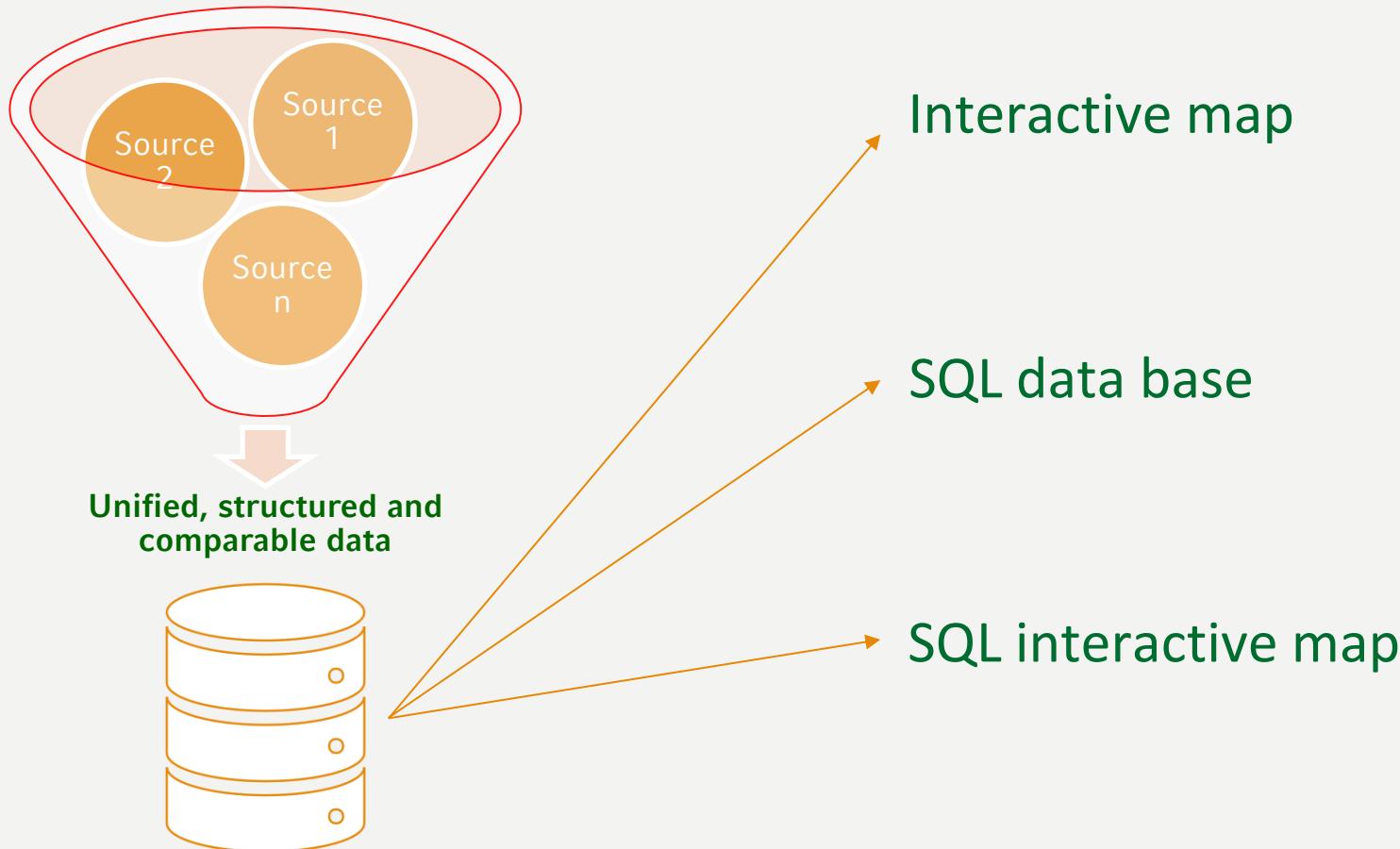
**fromage / formaggio fromage / formaggio**

Language family	roa	roa
PoS	noun	noun
Affix	Ø	Ø
Gender	m	Ø
Morph-lex Type	1	2

**Base type: FORMATICUS**



## Lexicographic Process





## Data access: interactive map

The screenshot shows the VerbaAlpina search interface. At the top, there is a search bar with placeholder text "Eine Suchanfrage eingeben..." and a dropdown menu titled "Kartographische Darstellung" with options like "Physique" and "Hexagonal". Below this is a sidebar with various filters:

- Sprachliche Kategorien: Basistypen, Morpho-lexikalische Typen, Phonetische Typen, Konzepte.
- Sprachkennzeichen Peripherie: informanter, Ergänzende Daten, Flächen und Regionen.
- Legende: Synoptische Karten.

The main area features a map of France with several regions highlighted in different colors (red, orange, yellow, green, blue, purple). Below the map are sections for "Données langagières" and "Données supplémentaires", each with dropdown menus for "Types de base", "Types morpho-lexicaux", "Types phonétiques", "Concepts", "Informateurs", "Données complémentaires", and "Surfaces et régions". At the bottom, there are links for "Légende" and "Plans synoptiques", and a button to "Fermer le menu".

## Two view modes:

- Physical & hexagonal

## Diverse filters:

- Onomasiologic
- Semasiologic
- Peripheral data

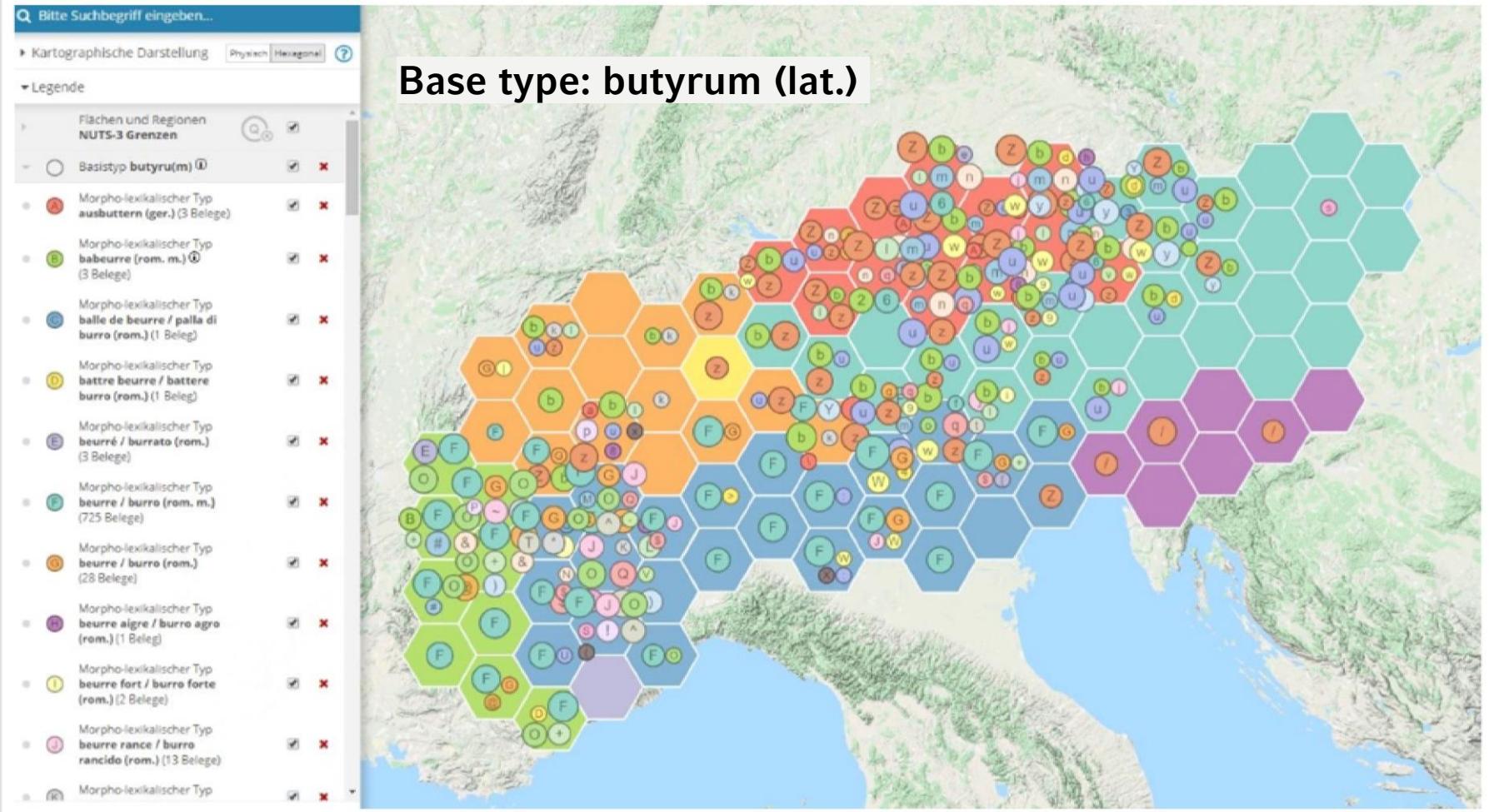


[https://www.verba-alpina.gwi.uni-muenchen.de/?page\\_id=133](https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133)

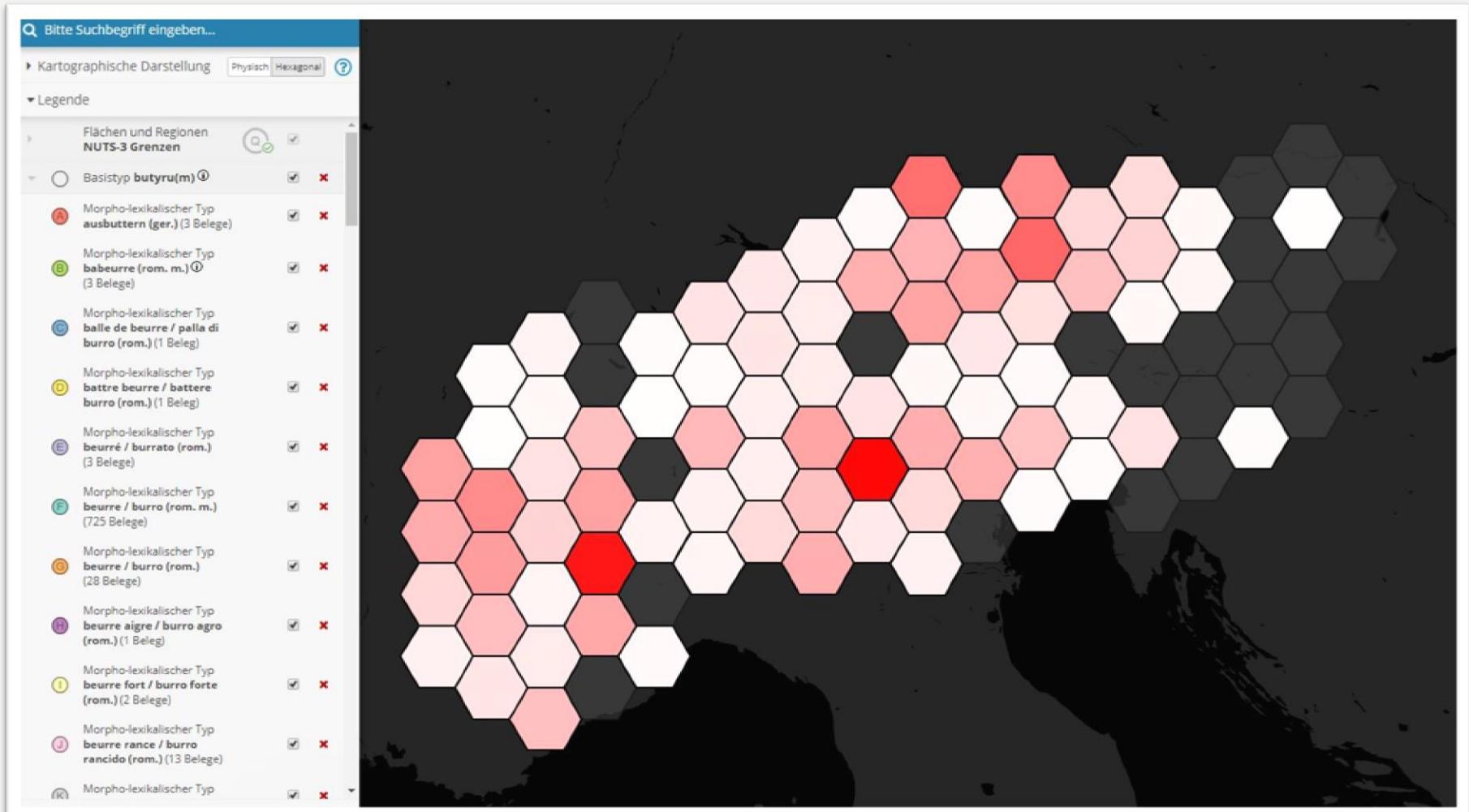
## Tools:

- Qu
- Sy





# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function





## Data access: SQL-Queries

Id_Beleg	Beleg	Quelle_Beleg	Name_Konzept	Beschreibung_Konzept	Breitengrad	Laengengrad	Gemeinde	Publikationsjahr	Erhebungsjahr	Sprache_Infor
281	mējrā	ALF#SUPP_37#14#889#Barcelonnette (B...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,650463	44,3863	Barcelonnette	1897-1900	(NULL)	rom
281	mējrā	ALF#SUPP_37#14#889#Barcelonnette (B...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,650463	44,3863	Barcelonnette	1897-1900	(NULL)	rom
282	moeðs	ALF#SUPP_37#14#956#Sax (Haute-Sav...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,838646	46,045357	Sax-Fer-à-Cheval	1897-1900	(NULL)	rom
283	moeðs	ALF#SUPP_37#14#967#Chamonix (Haut...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,869433	45,923697	Chamonix-Mont-Blanc	1897-1900	(NULL)	rom
284	pávλο:	ALF#SUPP_37#14#967#Chamonix (Haut...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,869433	45,923697	Chamonix-Mont-Blanc	1897-1900	(NULL)	rom
284	pávλο:	ALF#SUPP_37#14#967#Chamonix (Haut...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,869433	45,923697	Chamonix-Mont-Blanc	1897-1900	(NULL)	rom
285	säle	ALF#SUPP_37#14#967#Chamonix (Haut...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	6,869433	45,923697	Chamonix-Mont-Blanc	1897-1900	(NULL)	rom
286	mējro	ALF#SUPP_37#14#982#Maisette = Mais...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	7,1266143	44,9268317	Perrero	1897-1900	(NULL)	rom
286	mējro	ALF#SUPP_37#14#982#Maisette = Mais...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	7,1266143	44,9268317	Perrero	1897-1900	(NULL)	rom
287	rnjāndo	ALF#SUPP_37#14#982#Maisette = Mais...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	7,1266143	44,9268317	Perrero	1897-1900	(NULL)	rom
287	rnjāndo	ALF#SUPP_37#14#982#Maisette = Mais...	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	7,1266143	44,9268317	Perrero	1897-1900	(NULL)	rom
289	tēdzo	AIS#1192#1#1#Brigels-Brel	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,049995	46,766663	Brel/Brigels	1928-1940	(NULL)	rom
289	tēdzo	AIS#1192#1#1#Brigels-Brel	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,049995	46,766663	Brel/Brigels	1928-1940	(NULL)	rom
291	tēdzo	AIS#1192#1#3#Ptasch	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,216672	46,716663	Duvin	1928-1940	(NULL)	rom
291	tēdzo	AIS#1192#1#3#Ptasch	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,216672	46,716663	Duvin	1928-1940	(NULL)	rom
293	tēdzo	AIS#1192#1#5#Ems-Domat	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,449999	46,833329	Domat/Ems	1928-1940	(NULL)	rom
293	tēdzo	AIS#1192#1#5#Ems-Domat	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,449999	46,833329	Domat/Ems	1928-1940	(NULL)	rom
295	tēa	AIS#1192#1#7#Ardez	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	10,200002	46,766666	Ardez	1928-1940	(NULL)	rom
295	tēa	AIS#1192#1#7#Ardez	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	10,200002	46,766666	Ardez	1928-1940	(NULL)	rom
297	tēdzo	AIS#1192#1#11#Surrhain (Somvix)	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	8,933332	46,71666	Sumvitg	1928-1940	(NULL)	rom
297	tēdzo	AIS#1192#1#11#Surrhain (Somvix)	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	8,933332	46,71666	Sumvitg	1928-1940	(NULL)	rom
299	tēdzo	AIS#1192#1#13#Vrin	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,083334	46,649994	Vrin	1928-1940	(NULL)	rom
299	tēdzo	AIS#1192#1#13#Vrin	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,083334	46,649994	Vrin	1928-1940	(NULL)	rom
301	tēdzo	AIS#1192#1#14#Dahn (Prätz)	SENNHÜTTE	GEBAUDE, EINFACH, AUF DER ALM ZUR...	9,400005	46,733328	Sam	1928-1940	(NULL)	rom

- Direct access to the data base: structured & comparable data
- Textual mode
- Filtered and prepared data sets can be downloaded for further evaluation



## Approaches to sustainability

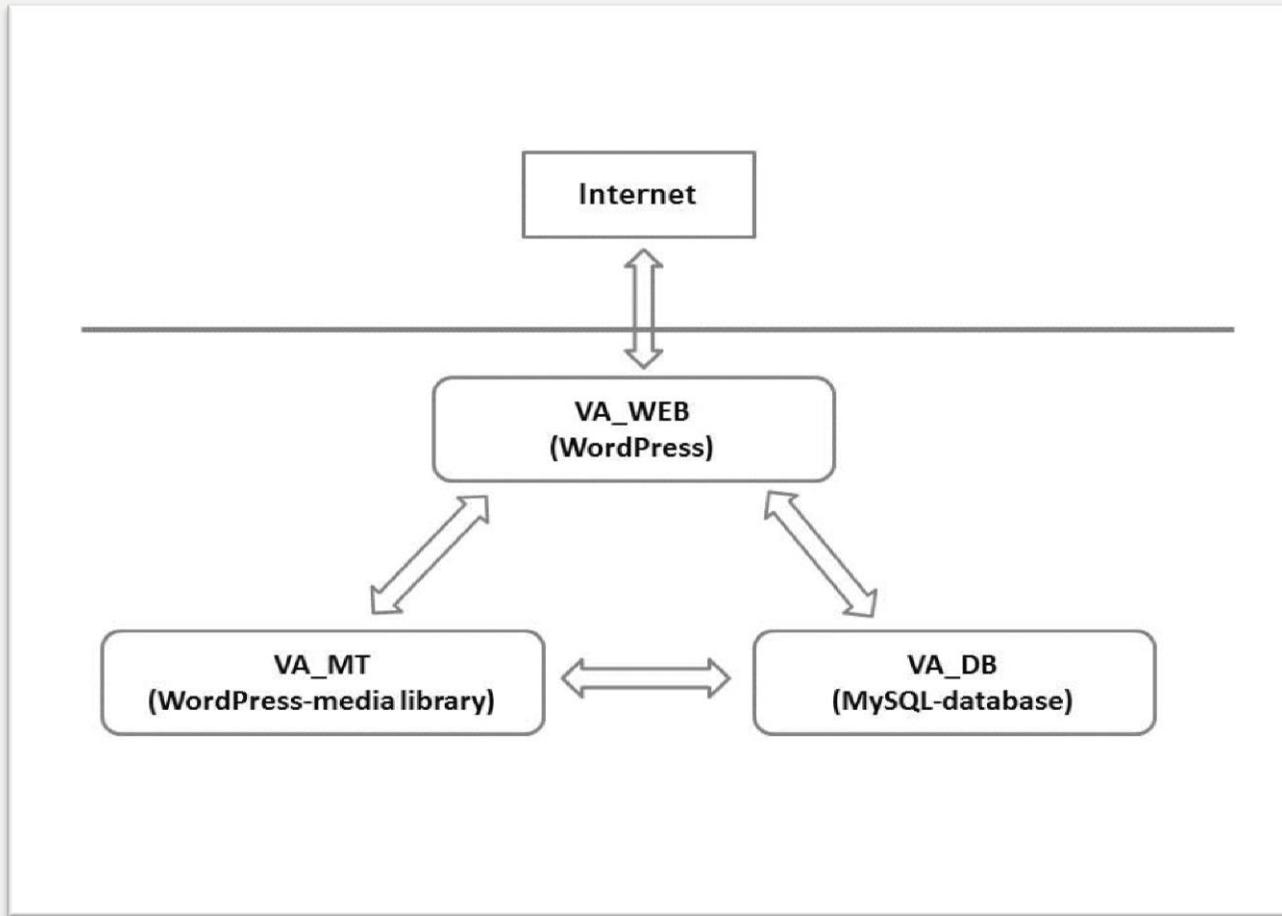
**1) Versioning**

**2) Citability**

**3) long-term archiving**



## 1) Versioning: The modules of VerbaAlpina





## 1) Versioning: The modules of VerbaAlpina

- VA\_DB and VA\_WEB are “frozen” every six months (15/06 and 15/12)
- versioning numbers for frozen copies (scheme: [year]/[sequence number], e.g. 18/1)
- every productive VA version is named XXX
- Possibility to switch between “productive” and “frozen” versions



## 2) Citability

- made possible by the versioning process
- date of last access is not necessary anymore  
(cited versions are stable)
- cite in the following way:

VerbaAlpina (VA), <http://www.verba-alpina.gwi.uni-muenchen.de>, [version]  
eg.: VerbaAlpina (VA), <http://www.verba-alpina.gwi.uni-muenchen.de>, 15/1

- graphic contents may also be cited: individual URLs for  
pages and pop-up windows



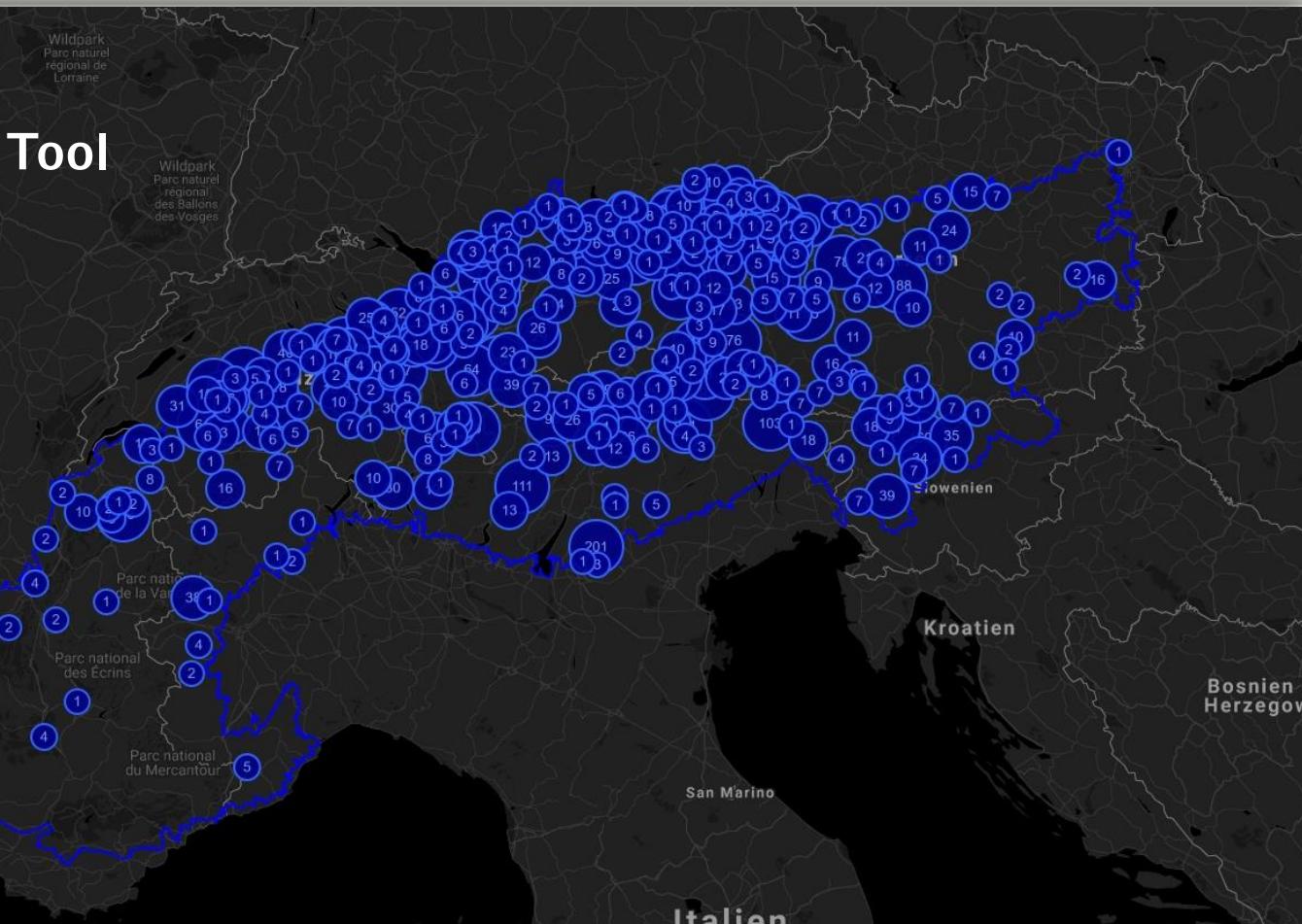
### 3) Long-term archiving

- Multiple copies of the project data archived by several institutions
  - currently: IT-Gruppe Geisteswissenschaften of Munich University (LMU Center for Digital Humanities) and archive.org
  - In the long term: University Library of Munich
- Documentation: data structuring, logical relationships between data and data categories, character encoding
- conversion from Google Maps to Leaflet is planned at the latest until the end of 2019

Wildpark  
Parc Naturel  
Régional du PerchePark  
naturel régional  
Anjou-Touraine

## Crowdsourcing Tool

Frankreich

Wildpark  
Parc Naturel  
Régional  
Périgord  
LimousinWildpark  
Parc Naturel  
Régional de  
Millévaches  
en LimousinParc Naturel  
Régional du MorvanWildpark  
Parc naturel  
régional de  
LorraineWildpark  
Parc naturel  
régional des  
Ballons  
des VosgesParc Naturel  
Régional des  
Monts d'ArdècheParc national  
des CévennesWildpark  
Parc Naturel  
Régional du  
Haut LanguedocFrancoprovençal  
Parc naturel  
régional des  
Alpes-de-Haute-Provence

Comment est-ce qu'on dit pour concept à commune ? Votre réponse





# Thanks for your attention!

