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 Kulturraum im Spiegel seiner Mehrsprachigkeit* (VerbaAlpina. The Alpine cultural region reflected through its multilingualism)
- Funded by the German Research Foundation (DFG)
- 1st term: 10/2014-10/2017, 2nd 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)

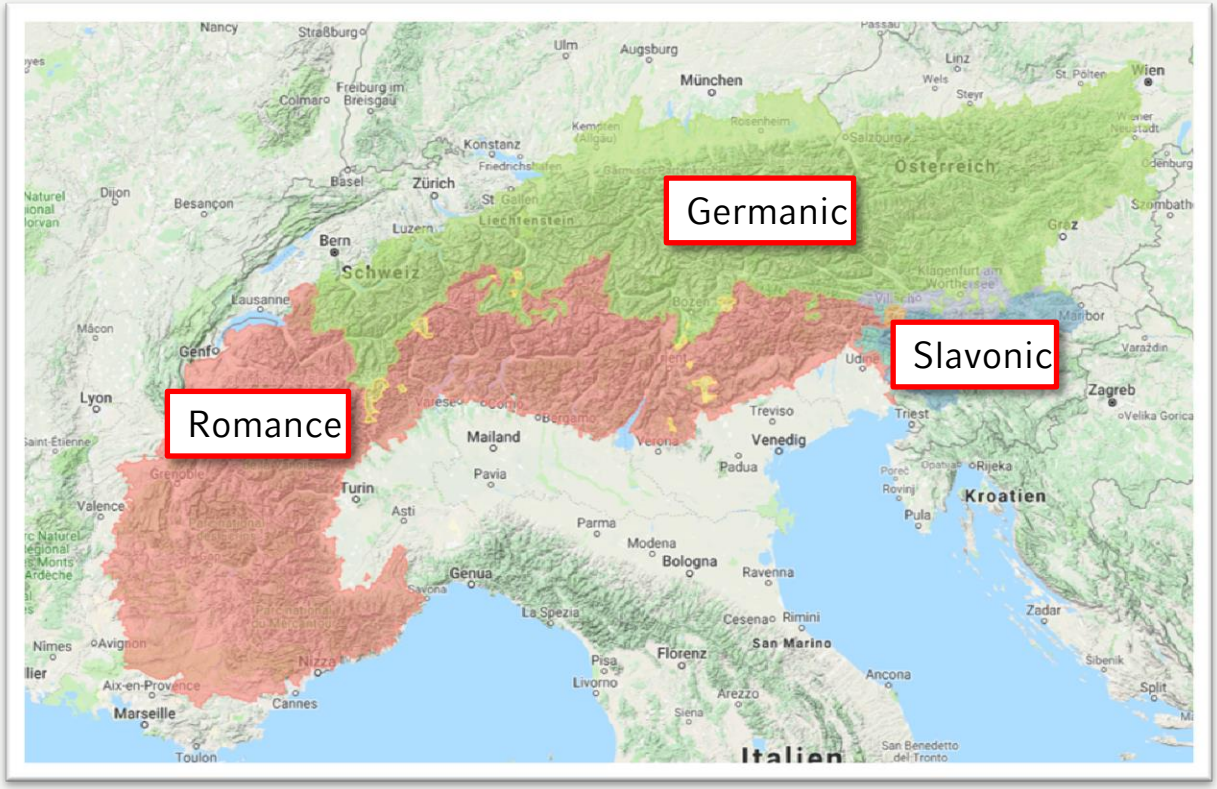
Member of staff

- 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$^2$, 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 |

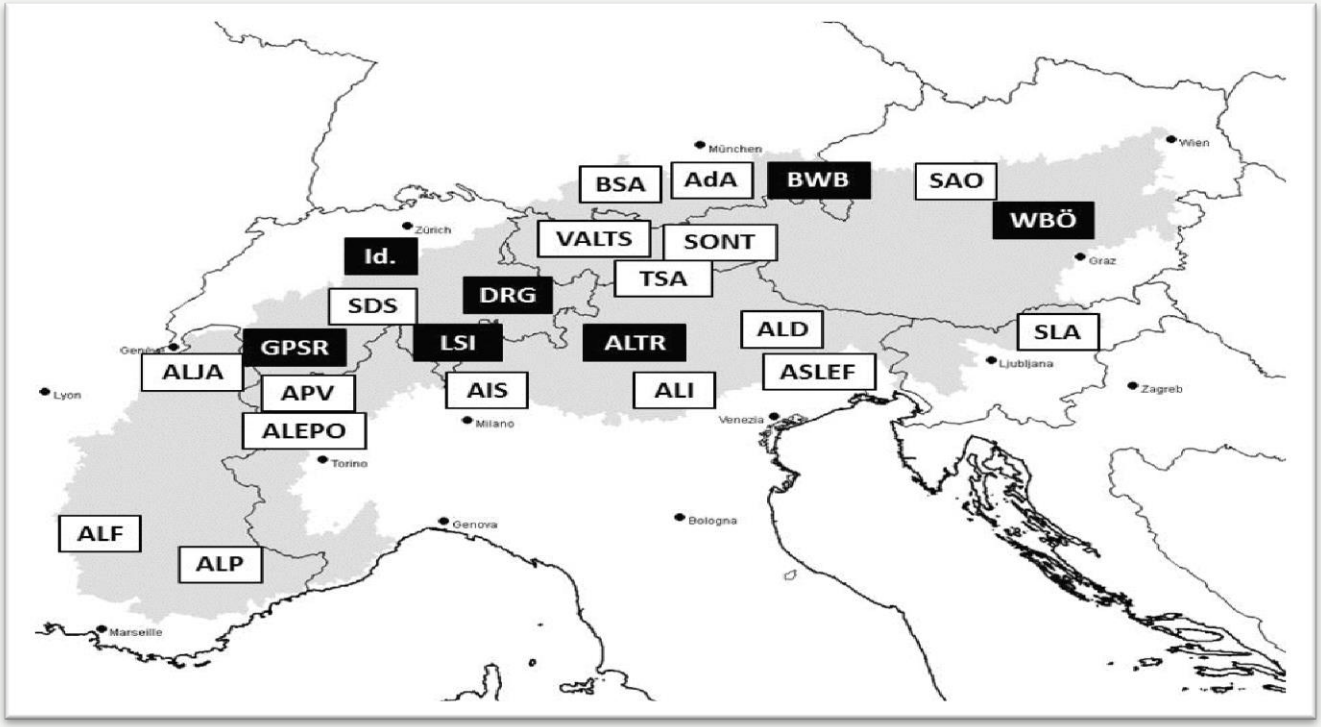
| project phase | I | II | III |
|---|---|---|---|
| focus | **culture**<br>• alpine pasture farming<br>• milk processing | **nature**<br>• landscape formations<br>• weather<br>• fauna<br>• flora | **modern life**<br>• ecology<br>• tourism |

## 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

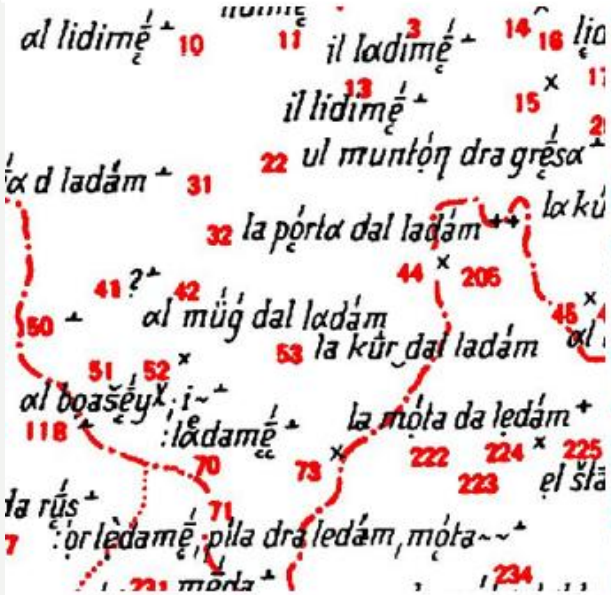
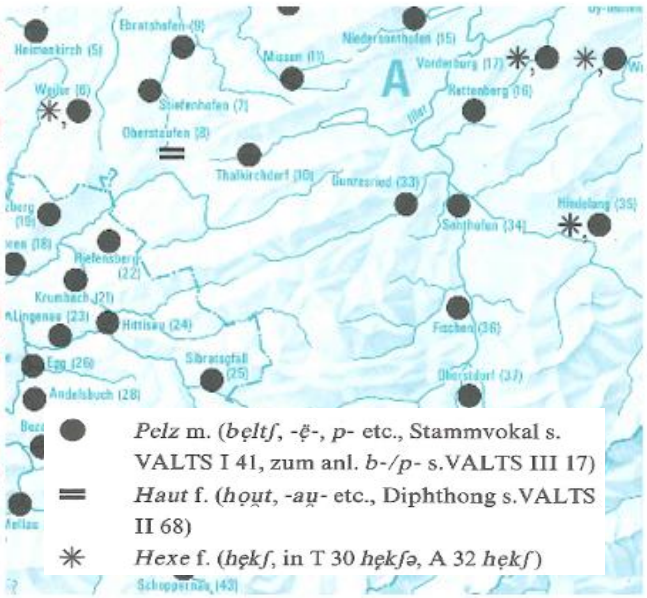
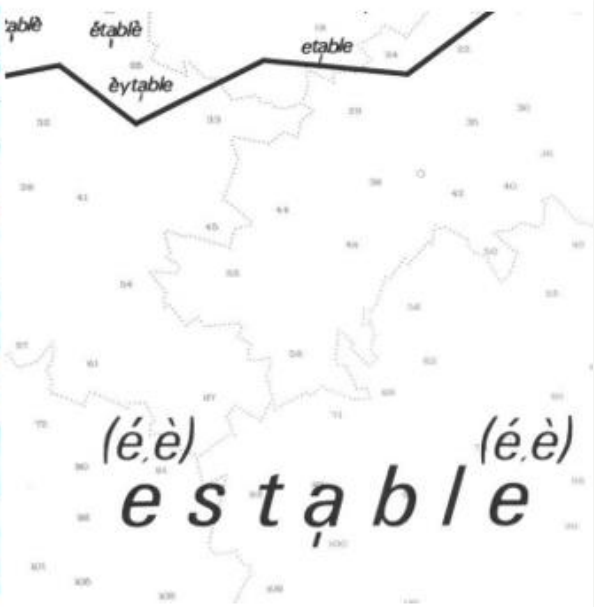
→ process of systematic data processing:

**- Transcription**

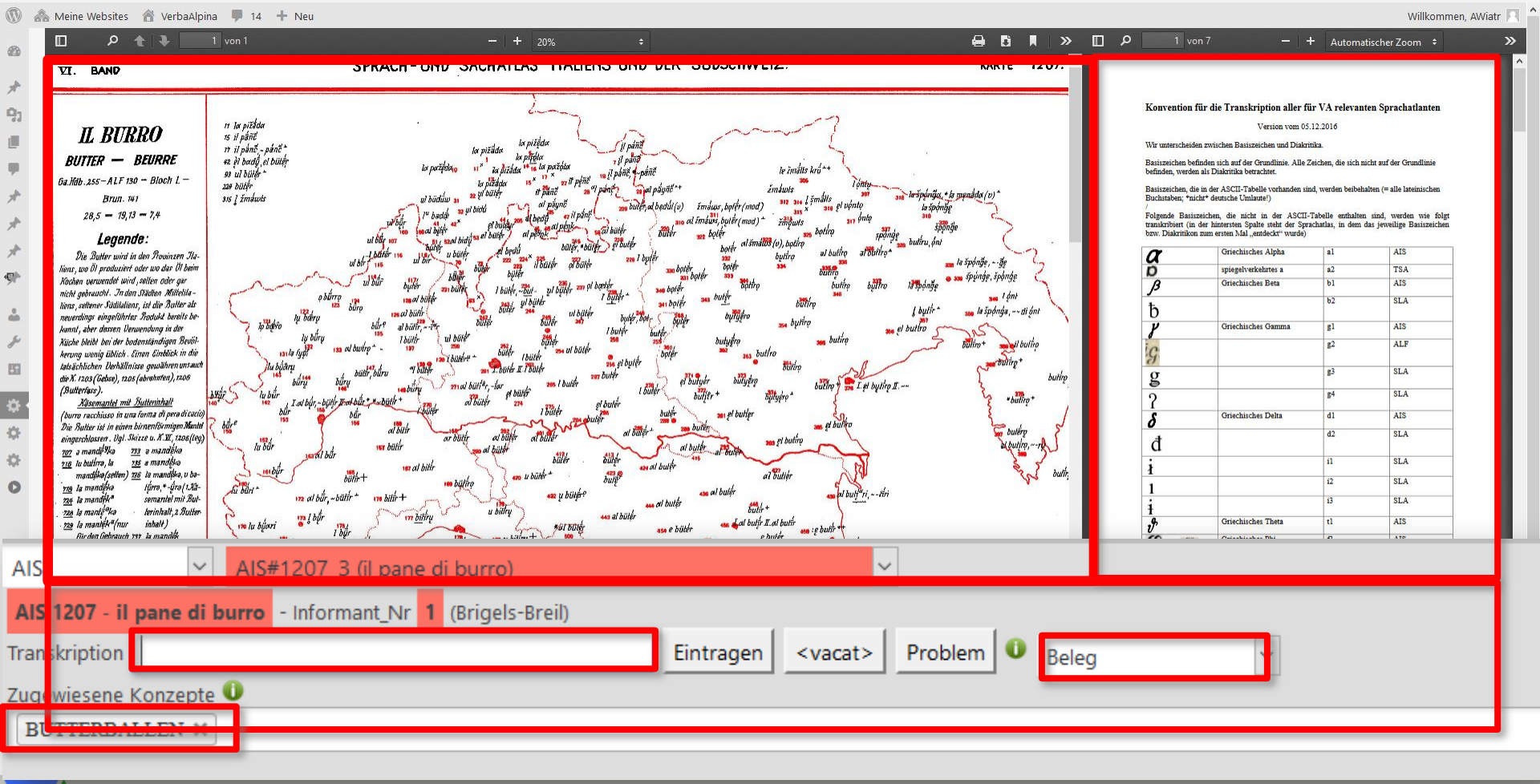
**- Tokenization**

**- Typification**

# Lexicographic Process

| AIS | VALTS | ALP |
|---|---|---|
|  |  |  |
| Single attestations | Phonetic Types, morphological types, single attestations | Areal types, single attestation |

# Lexicographic Process: Transcription Tool



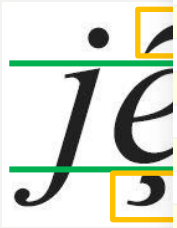
## Lexicographic Process: Beta Code

Beta Co

Diacritics

Base sign

Diacritics



?\\

- No l
- Inst
- The tran
- For allo

ibed

e

| erhebung | beta | ipa | Hex_ipa | Art |
|---|---|---|---|---|
| AIS | | |   | Trennzeichen |
| AIS | / | ˈ | ˈ | Akzent |
| AIS | // | ˈ | ˈˈ | Akzent |
| AIS | \ | ˌ | ˌ | Akzent |
| AIS | a | a | &#x0061 | Vokal |
| AIS | a! | a | a | Zeichen |
| AIS | a% | aȟ | ah̆ | Zeichen |
| AIS | a%0 | aȟ | ah̆ | Zeichen |
| AIS | a& | ɐ | ɐ | Zeichen |
| AIS | a( | ɑ | ɑ | Zeichen |
| AIS | a(* | ɑ | ɑ | Zeichen |
| AIS | a(- | ɑː | ɑː | Zeichen |
| AIS | a(-{i} | ɑi̯ | ɑːĭ | Zeichen |
| AIS | a(-\| | ɐː | ɐː | Zeichen |
| AIS | a(: | æ | æ | Zeichen |
| AIS | a({i}- | ɑi̯ | ɑːĭ | Zeichen |
| AIS | a(\|- | ɐː | ɐː | Zeichen |
| AIS | a) | ɑ | ɑ | Zeichen |
| AIS | a)- | ɑː | ɑː | Zeichen |
| AIS | a* | a | a | Zeichen |
| AIS | a*( | ɑ | ɑ | Zeichen |
| AIS | a*- | aː | aː | Zeichen |
| AIS | a*-? | aː | aː | Zeichen |
| AIS | a*: | æ | æ | Zeichen |
| AIS | a*?- | aː | aː | Zeichen |
| AIS | a*\| | ɐ | ɐ | Zeichen |
| AIS | a- | aː | aː | Zeichen |
| AIS | a-( | ɑː | ɑː | Zeichen |

# Lexicographic Process: Tokenization

| Attestation in Beta Code | Attestation in IPA | Concept |
|---|---|---|
| una1 mu:g/a1 da1 va/c)/ | unɑ mydʒɑ dɑ v'atɕ | HERD OF COWS |
| TOKENIZATION | | |
| una1 | unɑ | ARTICLE |
| mu:g/a1 | mydʒɑ | HERD |
| da1 | dɑ | PREPOSITION |
| va/c)/ | v'atɕ | COW |

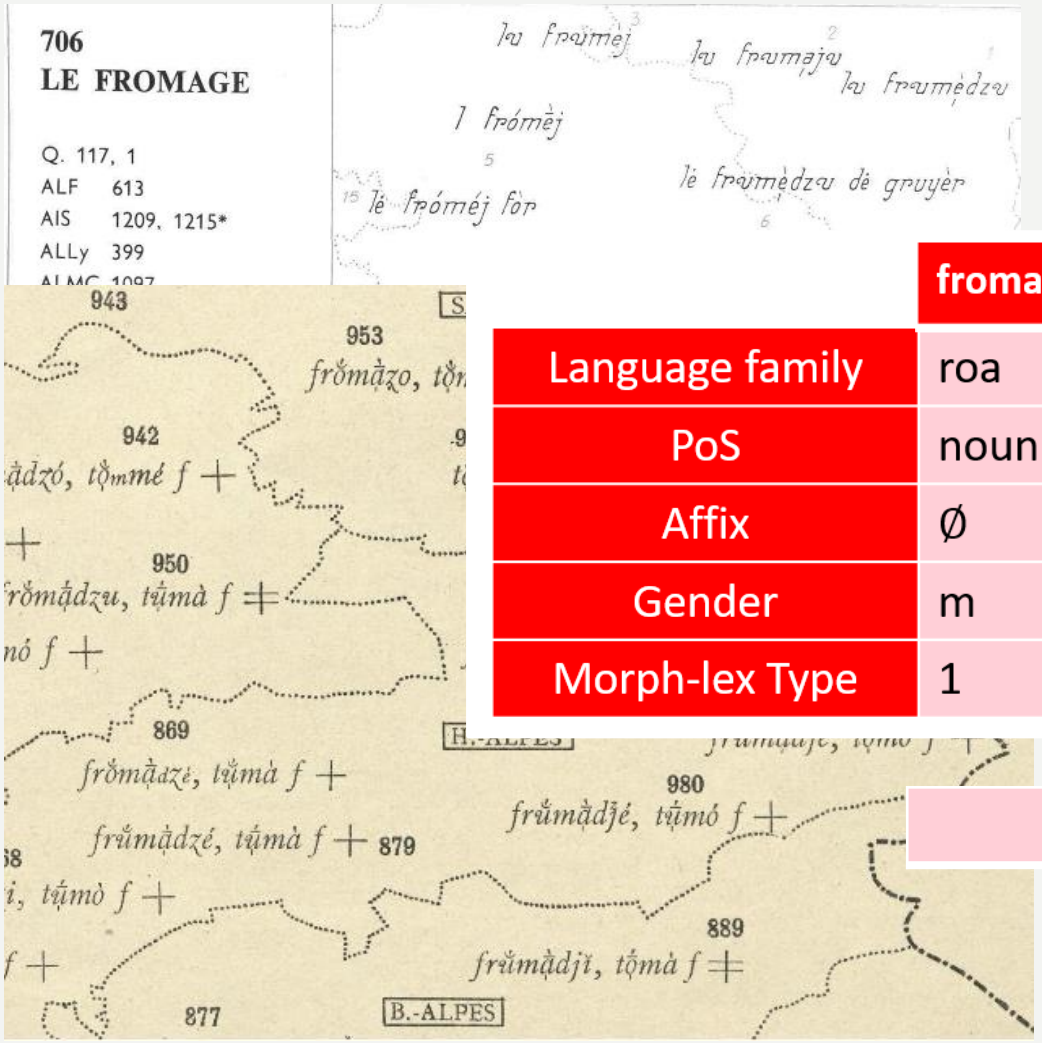
# Lexicographic Process: Typification

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

| | *barga* | *barg* | *margun* | *bargun* |
|---|---|---|---|---|
| **language family** | roa | roa | roa | roa |
| **PoS** | noun | noun | noun | noun |
| **affix** | - | - | + | + |
| **gender** | f | m | m | M |
| **morpho-lexical type** | **1** | **2** | **3** | **3** |

*barga/*barca

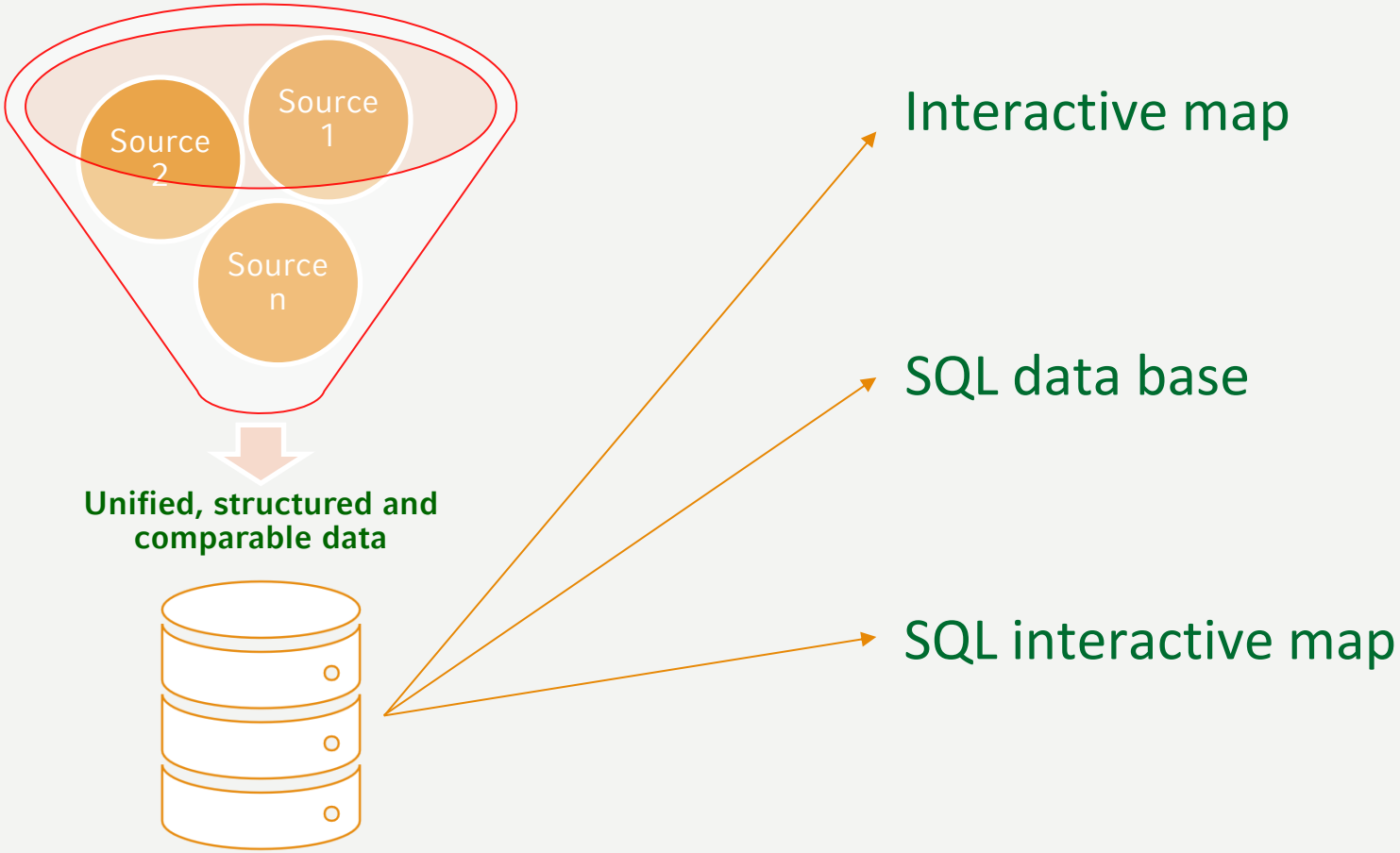
# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function




| | fromage / formaggio | fromage / formaggio |
|---|---|---|
| Language family | roa | roa |
| PoS | noun | noun |
| Affix | Ø | Ø |
| Gender | m | Ø |
| Morph-lex Type | 1 | 2 |

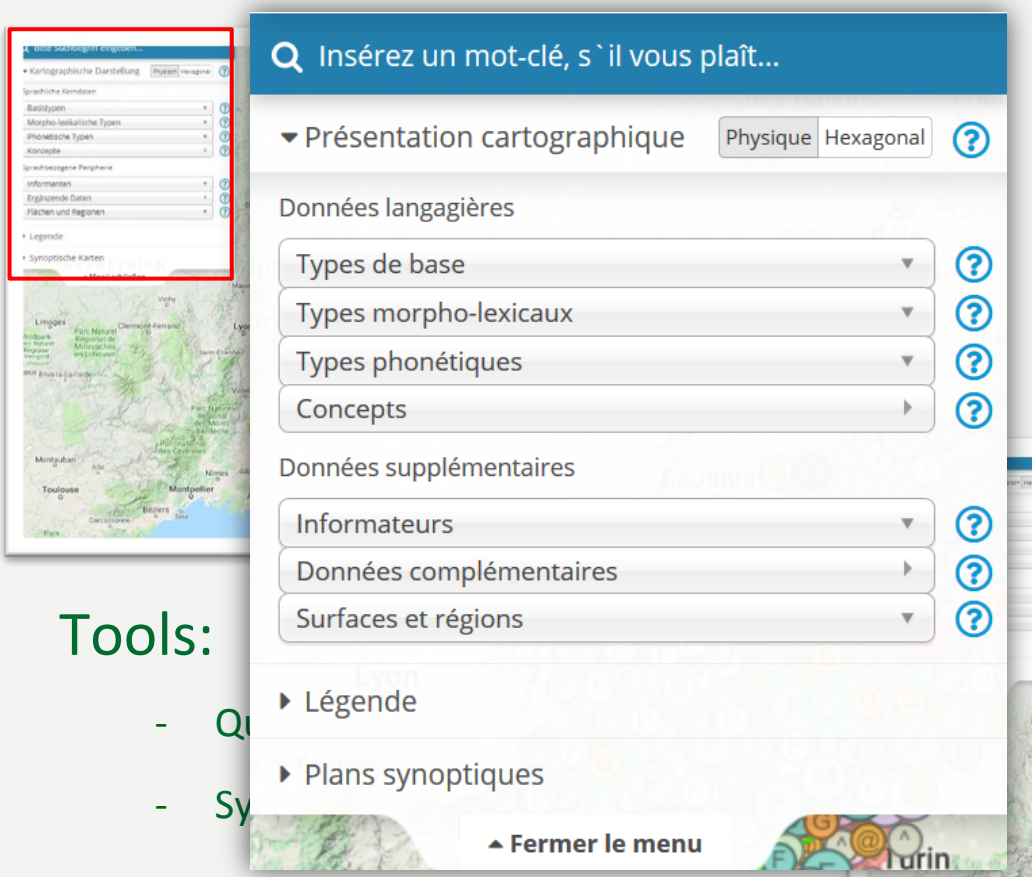
**Base type: FORMATICUS**

# Lexicographic Process



# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function

## Data access: interactive map




## Two view modes:

- Physical & hexagonal

## Diverse filters:

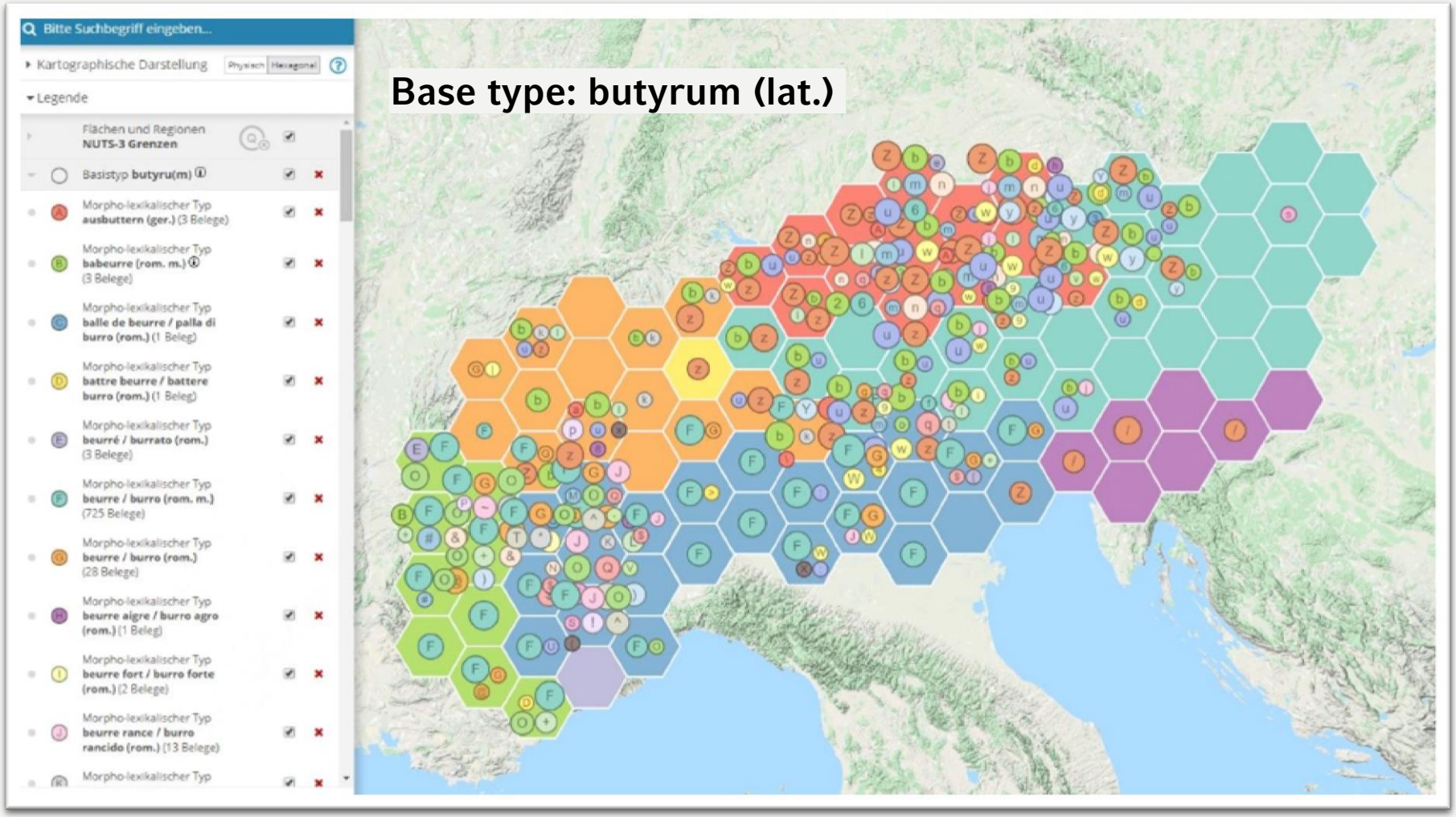
- Onomasiologic
- Semasiologic
- Peripherical data

## Tools:

- Qu
- Sy

https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133

# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function

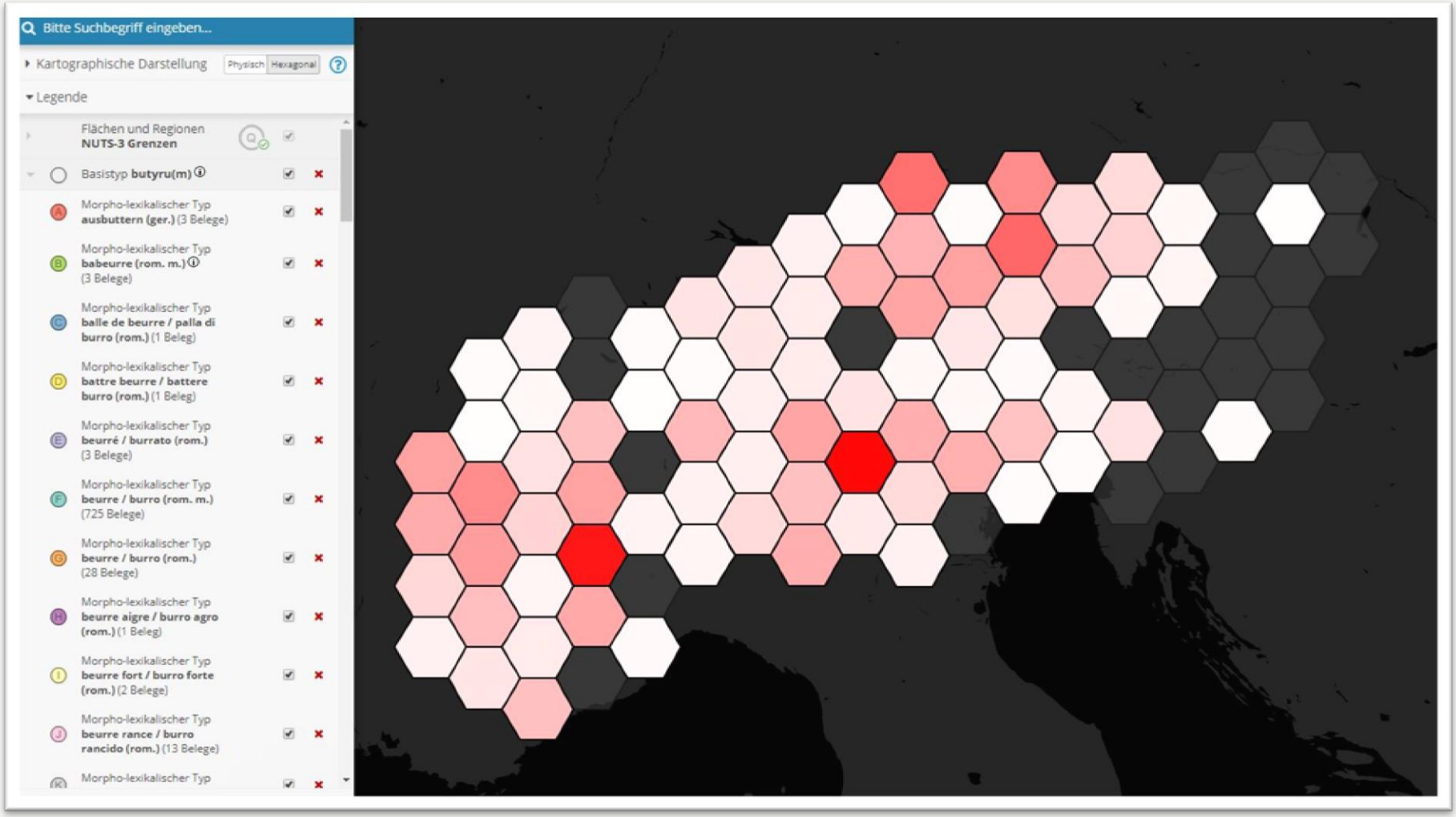





Christina Mutter | Aleksander Wiatr

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function






Christina Mutter | Aleksander Wiatr

# Data access: SQL-Queries

| Id_Beleg | Beleg | Quelle_Beleg | Name_Konzept | Beschreibung_Konzept | Breitengrad | Laengengrad | Gemeinde | Publikationsjahr | Erhebungsjahr | Sprache_Infor |
|---|---|---|---|---|---|---|---|---|---|---|
| 281 | mĕjră | ALF#SUPP_37#14#889#Barcelonette (B... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,650463 | 44,3863 | Barcelonnette | 1897-1900 | (NULL) | rom |
| 281 | mĕjră | ALF#SUPP_37#14#889#Barcelonette (B... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,650463 | 44,3863 | Barcelonnette | 1897-1900 | (NULL) | rom |
| 282 | mœ̆s | ALF#SUPP_37#14#956#Sixt (Haute-Sav... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,838646 | 46,045357 | Sixt-Fer-à-Cheval | 1897-1900 | (NULL) | rom |
| 283 | mœ̆s | ALF#SUPP_37#14#967#Chamonix (Haut... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,869433 | 45,923697 | Chamonix-Mont-Blanc | 1897-1900 | (NULL) | rom |
| 284 | pǎvʌo: | ALF#SUPP_37#14#967#Chamonix (Haut... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,869433 | 45,923697 | Chamonix-Mont-Blanc | 1897-1900 | (NULL) | rom |
| 284 | pǎvʌo: | ALF#SUPP_37#14#967#Chamonix (Haut... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,869433 | 45,923697 | Chamonix-Mont-Blanc | 1897-1900 | (NULL) | rom |
| 285 | sâĕ | ALF#SUPP_37#14#967#Chamonix (Haut... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 6,869433 | 45,923697 | Chamonix-Mont-Blanc | 1897-1900 | (NULL) | rom |
| 286 | mĕjro | ALF#SUPP_37#14#982#Maisette = Mais... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 7,1266143 | 44,9268317 | Perrero | 1897-1900 | (NULL) | rom |
| 286 | mĕjro | ALF#SUPP_37#14#982#Maisette = Mais... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 7,1266143 | 44,9268317 | Perrero | 1897-1900 | (NULL) | rom |
| 287 | mjâñdo | ALF#SUPP_37#14#982#Maisette = Mais... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 7,1266143 | 44,9268317 | Perrero | 1897-1900 | (NULL) | rom |
| 287 | mjâñdo | ALF#SUPP_37#14#982#Maisette = Mais... | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 7,1266143 | 44,9268317 | Perrero | 1897-1900 | (NULL) | rom |
| 289 | tẽdzo | AIS#1192#1#1#Brigels-Breil | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,049995 | 46,766663 | Breil/Brigels | 1928-1940 | (NULL) | rom |
| 289 | tẽdzo | AIS#1192#1#1#Brigels-Breil | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,049995 | 46,766663 | Breil/Brigels | 1928-1940 | (NULL) | rom |
| 291 | tẽdzo | AIS#1192#1#3#Pitasch | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,216672 | 46,716663 | Duvin | 1928-1940 | (NULL) | rom |
| 291 | tẽdzo | AIS#1192#1#3#Pitasch | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,216672 | 46,716663 | Duvin | 1928-1940 | (NULL) | rom |
| 293 | tẽdzo | AIS#1192#1#5#Ems-Domat | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,449999 | 46,833329 | Domat/Ems | 1928-1940 | (NULL) | rom |
| 293 | tẽdzo | AIS#1192#1#5#Ems-Domat | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,449999 | 46,833329 | Domat/Ems | 1928-1940 | (NULL) | rom |
| 295 | tẽo | AIS#1192#1#7#Ardez | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 10,200002 | 46,766666 | Ardez | 1928-1940 | (NULL) | rom |
| 295 | tẽo | AIS#1192#1#7#Ardez | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 10,200002 | 46,766666 | Ardez | 1928-1940 | (NULL) | rom |
| 297 | tẽdzo | AIS#1192#1#11#Surrhein (Somvix) | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 8,933332 | 46,71666 | Sumvitg | 1928-1940 | (NULL) | rom |
| 297 | tẽdzo | AIS#1192#1#11#Surrhein (Somvix) | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 8,933332 | 46,71666 | Sumvitg | 1928-1940 | (NULL) | rom |
| 299 | tẽdzo | AIS#1192#1#13#Vrin | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,083334 | 46,649994 | Vrin | 1928-1940 | (NULL) | rom |
| 299 | tẽdzo | AIS#1192#1#13#Vrin | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,083334 | 46,649994 | Vrin | 1928-1940 | (NULL) | rom |
| 301 | tẽdzo | AIS#1192#1#14#Dahn (Präz) | SENNHÜTTE | GEBÄUDE, EINFACH, AUF DER ALM ZUR... | 9,400005 | 46,733328 | Sarn | 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

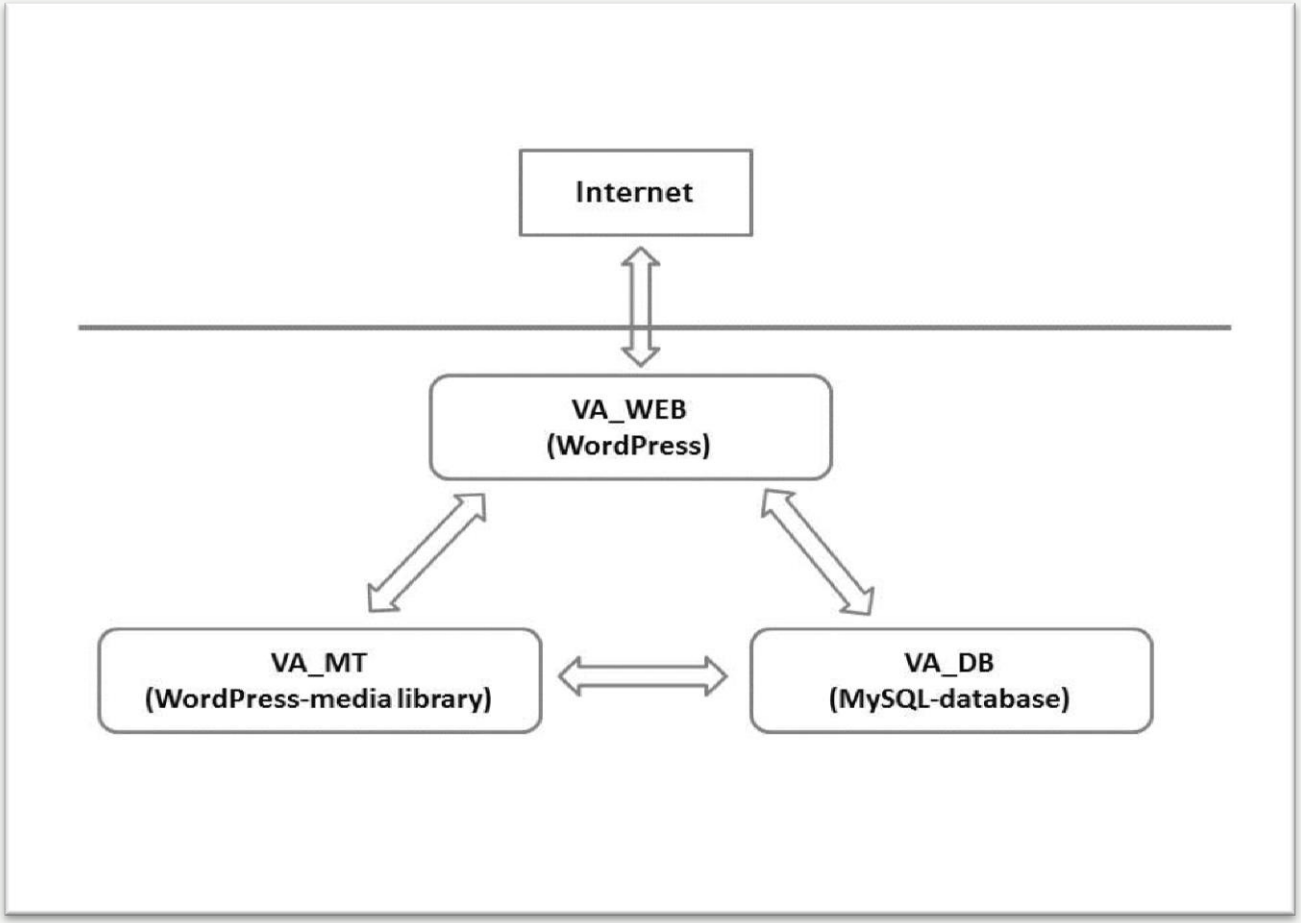
## Approaches to sustainability

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

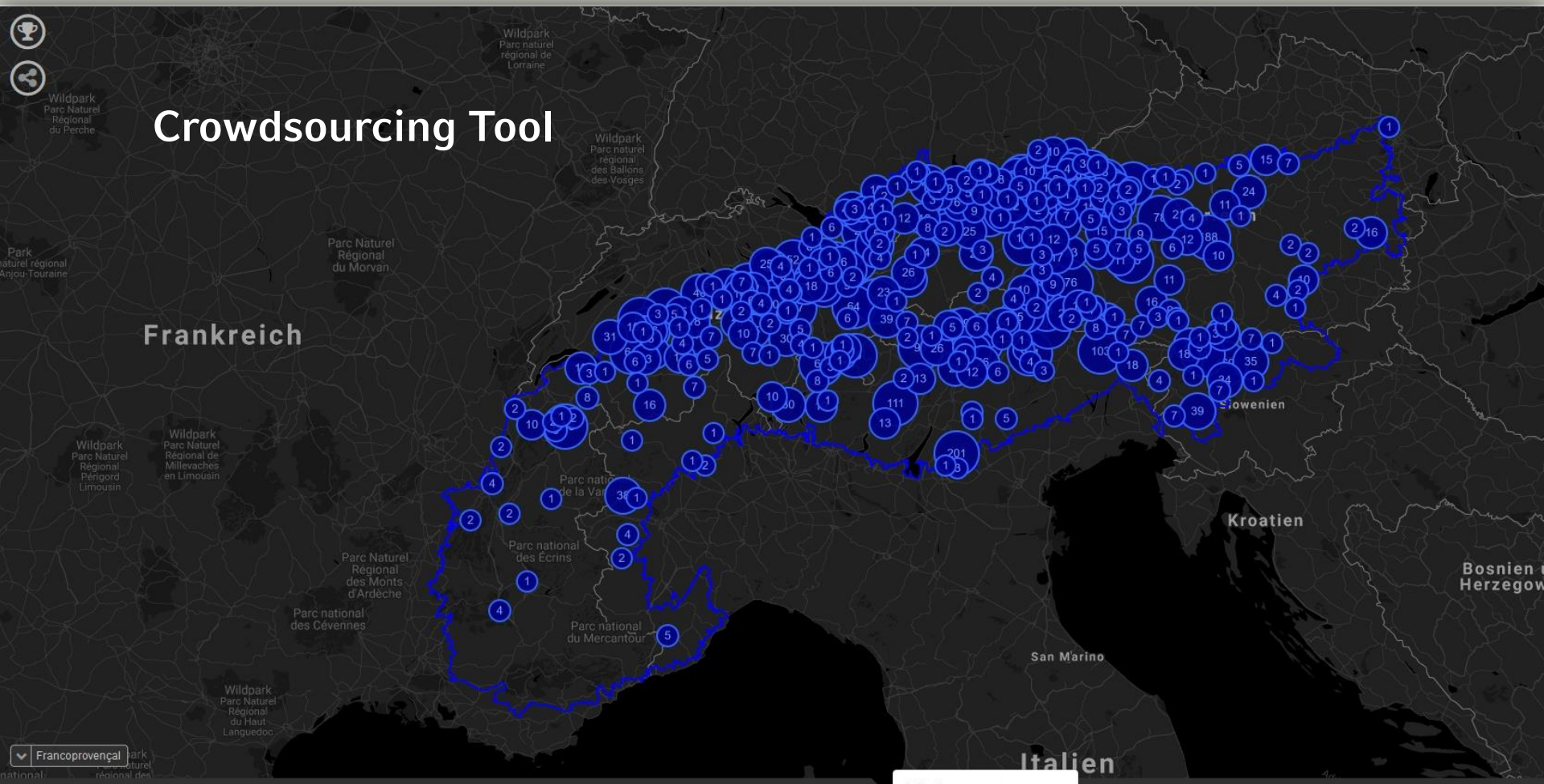
→ currently: IT-Gruppe Geisteswissenschaften of Munich University (LMU Center for Digital Humanities) and archive.org

→ 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# The Virtual Research Environment of VerbaAlpina and its Lexicographic Function






Christina Mutter | Aleksander Wiatr

# Thanks for your attention!